Culture Independent Discovery of New Cyanobacterial Natural Products

Diana Rafaela Reis de Sousa

Master's Degree in Applications in Biotechnology and Synthetic Biology

Chemistry and Biochemistry Department and Biology Department

2022

Supervisor

Pedro Nuno da Costa Leão, Principal Investigator, CIIMAR – Interdisciplinary Centre of Marine and Environmental Research

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Declaração de Honra

Eu, Diana Rafaela Reis de Sousa, inscrita no Mestrado em Aplicações em Biotecnologia e Biologia Sintética da Faculdade de Ciências da Universidade do Porto declaro, nos termos do disposto na alínea a) do artigo 14.º do Código Ético de Conduta Académica da U.Porto, que o conteúdo da presente dissertação reflete as perspetivas, o trabalho de investigação e as minhas interpretações no momento da sua entrega.

Ao entregar esta dissertação, declaro, ainda, que a mesma é resultado do meu próprio trabalho de investigação e contém contributos que não foram utilizados previamente noutros trabalhos apresentados a esta ou outra instituição.

Mais declaro que todas as referências a outros autores respeitam escrupulosamente as regras da atribuição, encontrando-se devidamente citadas no corpo do texto e identificadas na secção de referências bibliográficas. Não são divulgados na presente dissertação quaisquer conteúdos cuja reprodução esteja vedada por direitos de autor.

Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Diana Rafaela Reis de Sousa

Porto, 07 de janeiro de 2023

Acknowledgments

Firstly, I would like to thank Dr. Pedro Leão for the opportunity to develop my master thesis in his lab. Your guidance and trust led me to grow as a researcher. Thanks to your advice, what started as a project that was leading to no results and was making me very sad and doubtful on my abilities ended as a work that I can be proud of.

Of course none of this work would have been possible without the amazing team that is CNP. Everyone is ready to help and will not hesitate to do so. This group proves that it is possible to do great science in a friendly environment. To all of you thank you for making this lab the perfect place to start my scientific career. A special thanks to Adriana and Raquel that always encouraged me throughout the project. You were very good mentors and I hope I can continue to work with you.

To my friends, thank you for being there whenever I needed and for the fun that made me forget all the failed cloning reactions. To Rui, thank you for all the love and support. You really make me the happiest. To my parents and little sister, thank you for the unconditional love and for always believing in me. With all of you by my side I feel like I can do anything!

I would also like to acknowledge the Blue Young Talent Plus program promoted by CIIMAR – Interdisciplinary Centre of Marine and Environmental Research and the funder Fundação Amadeu Dias for the financial support during the development of this thesis.

This work was supported by national funds through FCT – Fundação para a Ciência e Tecnologia, I.P. under the framework of the projects with reference “PTDC/BIA-BQM/5760/2020” and “PTDC/BIA-BQM/29710/2017” and through grants UIDB/04423/2020 and UIDP/04423/2020. This work also received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 952374 and the project ATLANTIDA (ref. NORTE-01–0145-FEDER-000040), supported by the Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement and through the European Regional Development Fund (ERDF).

Resumo

O filo Cianobactéria é responsável pela produção de metabolitos secundários estruturalmente diversos com bioatividades potencialmente benéficas. A dificuldade em cultivar algumas cianobactérias em laboratório juntamente com as suas baixas taxas de crescimento limitam a sua exploração química, que não tem sido tão extensiva como para outros grupos de bactérias. A redução dos custos de sequenciação de genomas e os recentes avanços em ferramentas de bioinformática e metabolómica disponibilizaram várias estratégias para a descoberta de novos produtos naturais através de genomas. A metagenómica permite o estudo de ADN derivado de amostras ambientais (eDNA), facilitando a expressão heteróloga de clusters de genes biossintéticos (BGCs) de organismos não cultivados. Neste projeto foi utilizada uma abordagem metagenómica para revelar o potencial biossintético de cianobactérias que compõe uma amostra de um biofilme recolhido de um lago no Parque da Cidade, Matosinhos, Portugal. Desta amostra foram recuperados 3 genomas montados a partir de metagenomas (MAGs) de cianobactérias: dois pertencentes aos géneros *Planktothricoides* e *Planktothrix* e um que não pôde ser classificado ao nível do género. A análise bioinformática subsequente revelou 39 BGCs completos ou quase completos. Destes, cinco de classes biossintéticas distintas foram escolhidos para serem submetidos a expressão heteróloga: um sintase de policétidos (PKS)/sintase de péptidos não ribossomais (NRPS), duas microviridinas (péptidos ribossomais pós-traducionalmente modificados – RiPPs), um NRPS e uma cianobactina (RiPP). Foi concedida prioridade a um BGC que codifica uma microviridina que pertence ao MAG de cianobactéria mais abundante na amostra. O método de Direct Pathway Cloning acoplado com Sequence- and Ligation-Independent Cloning (DiPaC-SLIC) foi usado para clonar e expressar heterologamente o BGC de microviridina em *E. coli*. Para isolamento dos compostos associados com o BGC alvo, *E. coli* capazes de expressar o composto foram crescidas em culturas de larga escala e uma Extração em Fase Sólida foi realizada. Outro BGC que codifica uma microviridina está atualmente a ser clonado em *E. coli* TOP10. Adicionalmente, genes do BGC que codifica o NRPS foram amplificados diretamente do eDNA recuperado com sucesso. Este trabalho demonstra o potencial do DiPaC-SLIC para expressar heterologamente em *E. coli* BGCs derivados de MAGs de cianobactérias recuperados de biofilmes ambientais através de análise metagenómica.

Palavras-chave: Cianobactérias, Produtos Naturais, Metagenómica, Expressão heteróloga, Genomas montados a partir de metagenomas, DiPaC-SLIC

Abstract

The phylum Cyanobacteria is responsible for the production of structurally diverse secondary metabolites with potentially beneficial bioactivities. The difficulty in culturing some cyanobacteria in laboratory together with their slow growth rates limit their chemical exploration, which has not been as extensive as for other bacterial groups. The reducing costs of genome sequencing and the recent advances in bioinformatics and metabolomics tools have enabled multiple approaches for genome-based discovery of new natural products. Metagenomics allows for the study of DNA derived from environmental samples (eDNA), facilitating the heterologous expression of biosynthetic gene clusters (BGCs) from non-cultured microbes. In this project a metagenomics approach was used to uncover the biosynthetic potential of cyanobacteria that compose a lake biofilm sample collected at Parque da Cidade, Matosinhos, Portugal. From this sample, three cyanobacterial metagenome-assembled genomes (MAGs) belonging to the Oscillatoriales order were recovered: two from the genera *Planktothricoides* and *Planktothrix* and one which could not be classified at the genus level. Further bioinformatic analysis uncovered 39 complete and near-complete BGCs. From these, five of distinct biosynthetic classes were selected to undergo heterologous expression: a type I polyketide synthase (PKS)/non-ribosomal peptide synthase (NRPS), two microviridins (ribosomally synthesized and post-translationally modified peptides – RiPPs), a NRPS and a cyanobactin (RiPP). Priority was given to a microviridin BGC that belonged to the most abundant cyanobacterial MAG in the sample. Direct Pathway Cloning coupled with Sequence- and Ligation-Independent Cloning (DiPaC-SLIC) was used to clone and heterologously express the microviridin BGC in *E. coli*. For isolation of the compounds associated with the target BGC, the expressing *E. coli* strain was grown in large scale cultures and a Solid Phase Extraction was performed. Another microviridin BGC is currently being cloned into *E. coli* TOP10. Additionally, genes from a NRPS BGC were successfully amplified directly from the recovered eDNA. This work demonstrates the potential of DiPaC-SLIC to heterologously express in *E. coli* cyanobacterial MAG-derived BGCs recovered from environmental biofilms through metagenomic analysis.

Keywords: Cyanobacteria, Natural Products, Metagenomics, Heterologous expression, Metagenome-assembled genomes, DiPaC-SLIC

Table of Contents

Table Index

Figure Index

Abbreviations

A – adenylation domain

ACP – acyl carrier protein

Ada – decanoic acid

Adha – hydroxylated decanoic acid

Aeap – 1-amino-2-(N-amidino-Δ3-pyrrolinyl)-ethyl

Ahoa – hydroxylated octanoic acid

AmT – transaminase domain

Aoa – octanoic acid

AT – acyltransferase domain

BACs – Bacterial Artificial Chromosomes

BGC – Biosynthetic Gene Cluster

C – condensation domain

Choi – 2-carboxy-6-hydroxyoctahydroindole

CLR – continuous long read

DBG – De Bruijn Graph

DH – dehydratase

DiPaC – Direct Pathway Cloning

DiPaC-SLIC – Direct Pathway Cloning coupled with Sequence- and Ligation-Independent Cloning

Duk – Decontamination using kmers

eDNA – environmental DNA

ER – enoyl reductase

FAAL – fatty acyl AMP-ligase

GTDB – Genome Taxonomy Database

KR – ketoreductase

KS – ketosynthase

LC/MS – Liquid Chromatography coupled with Mass Spectrometry

LC-HRESIMS – Liquid chromatography-high resolution electrospray ionization mass spectrometry

MAG – Metagenome-assembled genome

Mgs – 2-O-methyl-3-sulfoglyceric acid

MS/MS – Tandem MS

NMR – Nuclear Magnetic Resonance

NP – Natural Product

NRP – non-ribosomal peptide

NRPS – Non-Ribosomal Peptide Synthetase

OLC – overlap-layout consensus

ONT – Oxford Nanopore Technologies

ORF – open reading frame

PacBio – Pacific Biosciences

PCP – peptidyl carrier protein

PCR – polymerase chain reaction

PKS – Polyketide Synthase

QC – Quality Control

RiPPs – ribosomally synthesized and post-translationally modified peptides

RRE – RiPP precursor peptide recognition element

RS – recognition sequence

SBS – sequencing by synthesis

SLIC – Sequence- and Ligation- Independent Cloning

SMRT – single-molecule real-time

SPE – Solid Phase Extraction

T – thiolation domain

TE – thioesterase domain

TNFs – tetranucleotide frequencies

1. Introduction

1.1. Cyanobacteria

Cyanobacteria are ancient photosynthetic prokaryotes which morphological characteristics include unicellular and filamentous forms capable of creating specialized cells for nitrogen fixation (heterocysts).[1,2] Some filamentous cyanobacterial strains are also able to form spore-like cells (akinetes) to survive in unfavourable conditions for long periods of time.[3] Fig. 1 demonstrates some of the morphological diversity presented by Cyanobacteria. By using mainly the pentose phosphate pathway, they fix CO_2 which can be used as sole source of carbon enabling growth of the cells. Most members of this phylum synthesize phycobiliproteins as major light-harvesting pigments.[4] Namely, allophycocyanin and phycocyanin characterized by its bluish green and blue colours, respectively, are present in all cyanobacteria. Some cyanobacteria may also produce the red colour phycobiliprotein, phycoerythrin.[5] The main photosynthetic pigment produced by blue-green bacteria is chlorophyll *a*.[4] Their ability to quickly adapt to a variety of ecosystems allowed them to colonize marine and fresh waters, and even soils and extreme environments such as hot spring, glaciers, hypersaline environments and deserts.[4,6] Cyanobacteria are pointed as responsible for Earth oxygenation and as the symbionts that gave rise to the plastids present in higher plants and algae. These Gram-negative bacteria were considered algae for a long time due to their metabolism, ecological role and occasionally large size.[4] As a consequence, Cyanobacteria nomenclature has been attributed according to both Bacteriological and Botanical Codes causing considerable confusion.[7] Even nowadays, the correct nomenclature system for Cyanobacteria is still undefined.[8]

Early cyanobacterial taxonomic classification was based on phenotypic features.[7,9] Although, with several strains maintaining a similar morphology to fossil representatives, and the lost and arise of morphological features used for higher taxa classification during evolution, this system proved to be inadequate.[8,9] The latest great revision on Cyanobacteria taxonomy used a polyphasic approach to establish monophyletic taxa that accounted for the huge genetic diversity of members of this phylum.[9] The polyphasic approach complements the classic morphological data with electron microscopy analysis and genetic and molecular information to provide a more accurate representation of phylogeny. Using this approach, Komárek et al divided Cyanobacteria through eight orders: *Oscillatoriales*, *Chroococcidiopsidales*, *Nostocales*, *Spirulinales*, *Chroococcales*, *Pleurocapsales*, *Synechococcales*, and *Gloeobacterales*.[9]

Nonetheless, the authors admit the need for further revisions on taxonomy of this phylum with the increase of the available genomic data.[9]

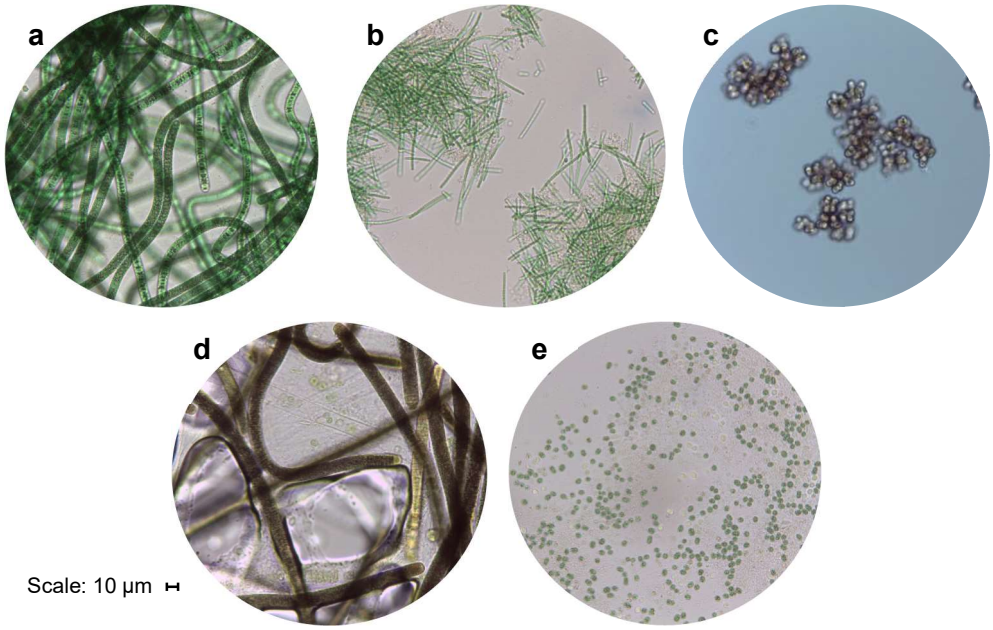


Figure 1 – Light microphotographs illustrating cyanobacterial morphological diversity. Photos taken from LEGEcc[10] cyanobacterial cultures. **a:** *Planktothrix mougeotii* LEGE 07231. **b:** *Alkalinema aff. pantanalense* LEGE 15481. **c:** unidentified *Pleurocapsales* LEGE 10410. **d:** *Limnoraphis robusta* LEGE XX358. **e:** *Microcystis aeruginosa* LEGE 91341.

1.2. Cyanobacterial natural products

Natural products (NPs) are small molecules produced by living organisms that are not strictly required for their survival.[11] The bioactivity of these compounds and their unique chemical structures often serve as inspiration for drug development.[12] Plant-derived NPs have been used in traditional medicine for centuries. Despite that, only with the isolation of morphine from the seed juice of *Papaver somniferum* in 1817[13], plant-derived NPs started to be seen as possible candidates for drug development.[14] Only a decade later, in 1928, when Alexander Fleming discovered penicillin from *Penicillium notatum*[15], microorganisms were recognized as potential sources of bioactive molecules.[11] The unique metabolic activities developed by microbes to survive in a wide range of environments brings structural richness to its secondary metabolites which are associated with distinct bioactivities.[16]

From 1981 to 2019, 22.7% from a total of 1881 newly approved drugs were NPs or NP-derivatives. If we account only for the 1394 approved small-molecule drugs, the percentage of NPs and NP-derivatives rises to 32.6. From these compounds it is possible to count 3 antifungal drugs, 4 multiple sclerosis drugs, 6 antiviral drugs, 6 antiglaucoma drugs, 9 antiparasitic drugs, 9 antidiabetic drugs, 61 anticancer drugs, and 89 antibacterial drugs. Moreover, several synthetically produced drugs are inspired by NPs

functional groups further demonstrating the impact of these compounds on the development of therapeutic approaches. As such, NPs present a reliable source of new leading drugs and templates for drug development with potential to fight the rising problem of drug resistance.[12]

Among bacterial phyla, the antibiotic producing Actinobacteria have been extensively studied. The chemical exploitation of these microbes has led to a significant discovery of bioactive metabolites such as the antibiotics streptomycin (isolated from *Streptomyces griseus*) and chloramphenicol (produced by *Streptomyces venezuelae*)[11], the anticancer anthracyclines (firstly isolated from *Streptomyces peucetius*)[17] and several other bioactive molecules already approved for clinical application or undergoing evaluation.[18]

Unlike Actinobacteria, Cyanobacteria is a less explored phylum responsible for the production of promising compounds for biotechnological applications.[6] The nutritional value of Spirulina (*Arthrospira platensis*) has been recognized by thousands of years, with the Aztecs including this cyanobacterium in their routine diet.[19] Currently, this cyanobacterium has been highlighted by its potential anti-obesity activity.[20,21] With the increasing research on cyanobacterial NPs, several bioactive secondary metabolites with potential applications in cosmetology, agriculture and pharmacology have been described.[6] Dolastatin 10, originally isolated from the sea hare *Dolabella auricularia*, was later recognized as a cyanobacterial metabolite produced by *Symploca* sp. VP642.[22] This NP has demonstrated antitumorigenic activity and has been used in the development of antibody-drug conjugates with some derivatives already approved for therapeutic applications.[23] Although only Dolastatin 10 and derivatives have been approved for clinical use, several cyanobacterial secondary metabolites have demonstrated promising bioactivities for biotechnological applications.[6] Until 2019, approximately 1630 distinct compounds were described from these organisms. These compounds were divided by Demay et al.[6] into 260 families from 10 different chemical classes, namely, alkaloids, terpenes, depsipeptides, polysaccharides, lipids, polyketides, lipopeptides, peptides, macrolides/lactones, and others (Fig. 2). These metabolites present several potentially beneficial activities, such as, antibacterial, anti-inflammatory, antiviral, antioxidant, enzyme/protease inhibition, among others (Fig. 3). Members of the Oscillatoriales and Nostocales orders are responsible for the production of 75.8% of the analyzed metabolite families (Fig. 4). In comparison, the remaining cyanobacterial orders are underexplored.[6] Moreover, most of the explored cyanobacteria belong to marine environments, leaving a huge diversity from distinct habitats unexplored.[6]

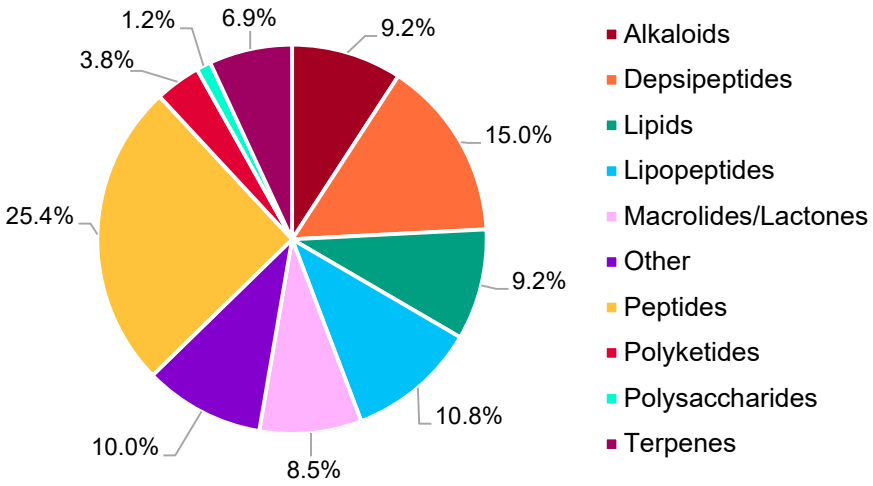


Figure 2 – Classification of the 260 families of cyanobacterial metabolites according to their chemical class. Adapted from Demay et al. 2019[6].

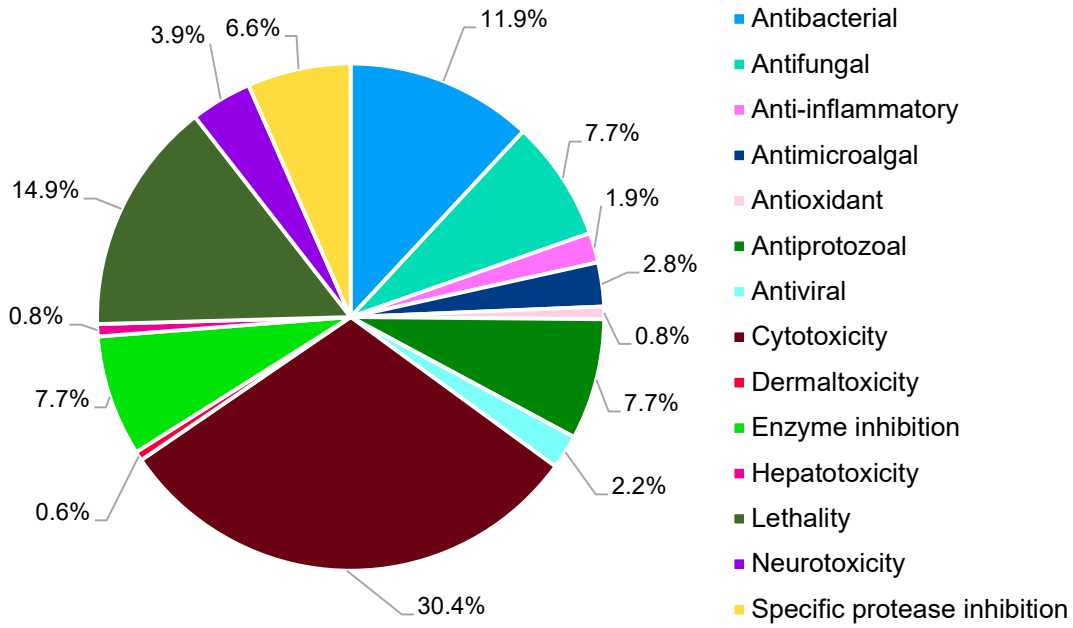


Figure 3 – Classification of the bioactivities displayed by the 260 families of cyanobacterial metabolites. Adapted from Demay et al. 2019[6].

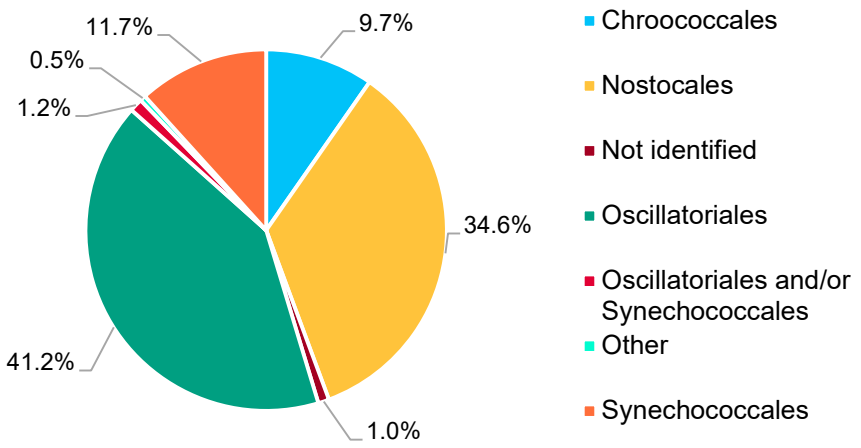


Figure 4 – Taxonomic distribution of the 260 families of cyanobacterial metabolites. Adapted from Demay et al. 2019[6].

The fruitful early research on cyanobacterial NPs relied on a bioactivity-guided isolation of the compounds. Notwithstanding, this approach has been leading to the rediscovery of secondary metabolites.[24] The recent advances in algorithms for bioinformatic analysis and the reducing costs of genome sequencing have been paving the way to the development of genome-derived discovery of new NPs. This is a strategy in which new NPs are discovered using a bioinformatic prediction of the respective biosynthetic gene clusters (BGCs).[24,25] Usually, the biosynthetic genes required for the production of NPs are clustered together on the microbial genome forming the BGCs.[26] Since the genome sequencing of *Synechocystis* sp. PCC 6803 in 1996[27] until 2017, only 400 cyanobacterial genomes became available in public databases representing only 0.6% of the total Bacterial and Archaea genomes. The lengthy isolation procedures and difficulty in obtaining axenic cyanobacterial cultures ideal for genome sequencing can explain the low number of sequenced genomes, compared to other phyla.[28] By September 2021, with the development of next-generation sequencing and metagenomic approaches, GeneBank entries accounted for 3,265 assemblies of cyanobacterial genomes.[29] The genome sequencing efforts together with the study of underexplored taxa will help elucidate the biosynthetic pathways of cyanobacterial metabolites and lead to the discovery of new chemistry.

1.3. Cyanobacterial biosynthetic gene clusters

The enzymes required for the biosynthesis of cyanobacterial secondary metabolites are arranged in BGCs. Cyanobacteria genomes are enriched in polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS), hybrid PKS/NRPS and ribosomally synthesized and post-translationally modified peptides (RiPPs).[30]

NRPSs are multifunctional enzymes that catalyze the assembly of non-proteinogenic and proteinogenic amino acids into a final non-ribosomal peptide (NRP). Firstly, the substrate must pass through the adenylation domain (A) which activates the amino acid to its adenylate form and, afterwards, transfers the adenylated amino acid to the thiol group of the phosphopantetheine cofactor of the peptidyl carrier protein (PCP, also known as thiolation domain – T).[31,32] The PCP is then responsible to deliver the intermediates to the catalytic domains for chain elongation, modification and release. PCPs contain a conserved serine residue that covalently binds to the phosphopantetheine cofactor. This modification occurs after translation and is catalyzed by a phosphopantetheinyl transferase normally encoded within the NRPS cluster. The amino acid and peptide intermediates form a thioester linkage between their carboxyl group and the thiol group of the phosphopantetheine cofactor. The condensation domain (C) performs chain elongation by transferring the aminoacyl or peptidyl group attached to the upstream PCP to the primary amine of the amino acid that was previously linked onto the downstream PCP. The NRPS terminal module has an extra catalytic domain, the thioesterase domain (TE). This domain catalyzes the cyclization, through lactam or lactone formation, or hydrolysis to release the peptide from the final PCP. Additionally, NRPS modules can contain epimerization domains, N-methylation domains and/or reductase domains.[32] A schematic representation of a possible NRPS BGC is represented in Fig. 5. Examples of cyanobacterial NRPs include the hassallidins[33], the spumigins[34] and the aeruginosins[35].

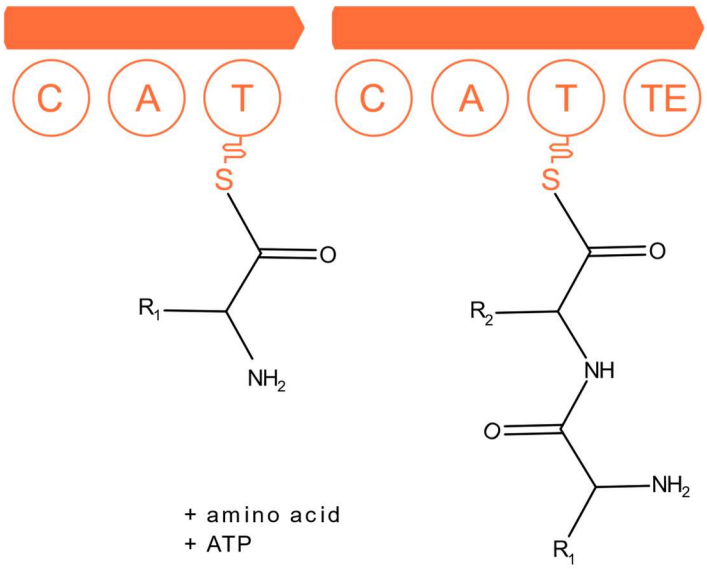


Figure 5 – Schematic representation of a NRPS. A: adenylation. T: thiolation. C: condensation. TE: thioesterase. Adapted from Kehr et al. 2011.[36]

PKS modules have several domains. The first domain to act on polyketide biosynthesis is the acyltransferase domain (AT) that selects the substrate and transfers it to the phosphopantetheinyl cofactor of the acyl carrier protein (ACP also referred to as thiolation domain – T). Chain elongation is then performed by a ketosynthase (KS) domain together with a ketoreductase (KR) domain, a dehydratase (DH) domain and a enoyl reductase (ER) domain. The KS domain loads a ketoacyl group into the growing chain attached to the ACP domain, afterwards the ketoacyl group is converted to a CH_2-CH_2 bond by three cascade reactions promoted by the KR, DH and ER domains. The last PKS module contains a terminal TE domain that catalyzes the release of the polyketide through hydrolysis or macrocyclization. PKS can be divided in three types: 1) Type I PKS that are large proteins composed by all the domains arranged in a sequential manner; 2) in Type II PKS the domains required for chain initiation, elongation and termination are encoded as single proteins and interact transiently; 3) Type III PKS are encoded as single proteins and use soluble malonyl-CoA extender units instead of ACP loaded molecules.[37] Type I and Type II PKSs are commonly encoded in cyanobacterial genomes whereas Type III PKSs are hardly found.[38] A schematic representation of possible Type I and Type II PKS biosynthetic pathways is represented in Fig. 6.

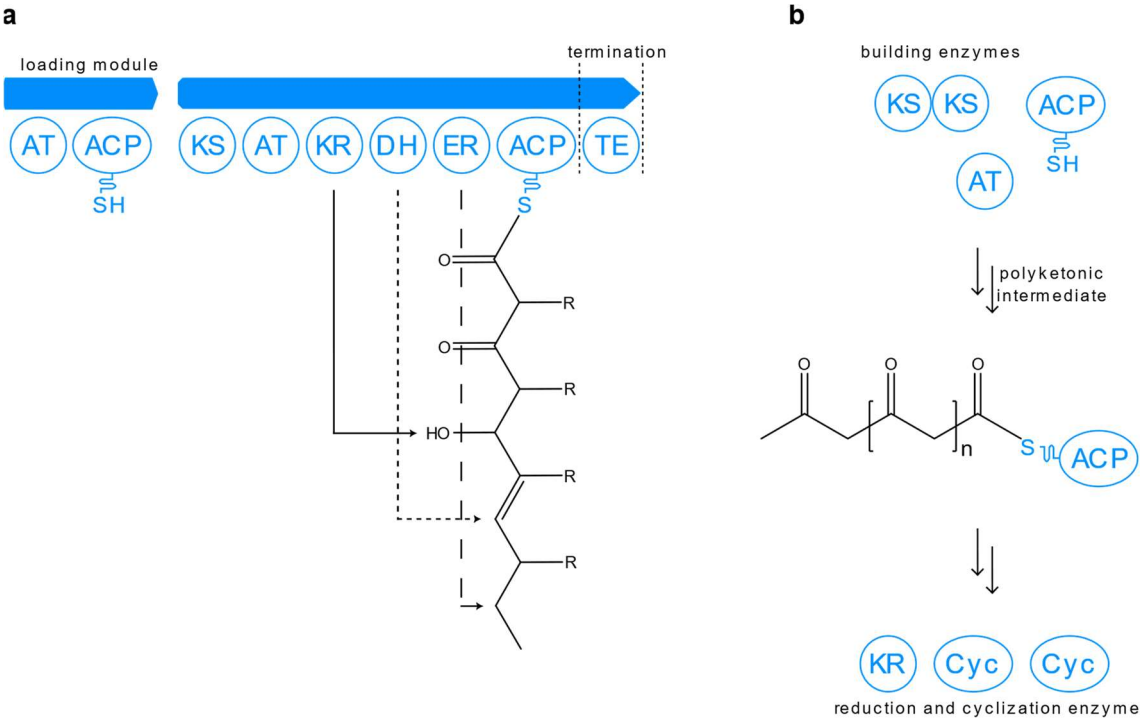


Figure 6 – Schematic representation of **a:** a putative Type I PKS; **b:** a putative Type II PKS. AT: acyltransferase domain. KR: ketoreductase. ACP: acyl carrier protein. KS: ketosynthase. DH: dehydratase. ER: enoyl reductase. TE: thioesterase. Cyc: cyclization domain. Adapted from Kehr et al. 2011[36] and Walsh and Tang 2017[37].

PKS and NRPS modules can also be encountered in hybrid PKS/NRPS BGCs. These hybrid enzymes are responsible for the biosynthesis of a vast array of structurally

distinct secondary metabolites in cyanobacteria such as the desmamides[39], the nodularins[40] and the microcystins[41]. Usually, in cyanobacteria the PKS module precedes the NRPS module. The modules may be contained in different enzymes or on the same enzyme separated by a transaminase domain (AmT) that generates the 3-amino of the polyketide synthesized by the PKS module.[31] The general BGC organization of hybrid PKS/NRPS is represented in Fig. 7.

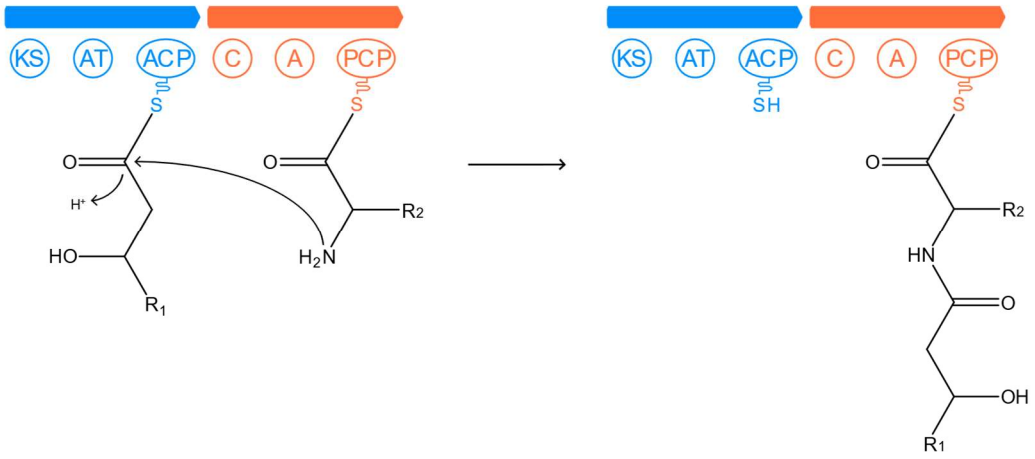


Figure 7 – Schematic representation of the interaction between PKS and NRPS modules. KS: ketosynthase. AT: acyltransferase domain. ACP: acyl carrier protein. PCP: peptidyl carrier protein. C: condensation domain. A: adenylation domain. Adapted from Walsh and Tang 2017[31].

RiPPs are ribosomal synthesized peptides that undergo a certain level of post-translational modifications. The precursor peptide that contains the final peptide sequence is directly encoded in the genome.[42] The process of maturation of the precursor peptide varies for each RiPP family. A schematic representation of a putative RiPP BGC is represented in Fig. 8. Microviridins[43], cyanobactins[44] and lanthipeptides[45] are examples of cyanobacterial RiPPs.

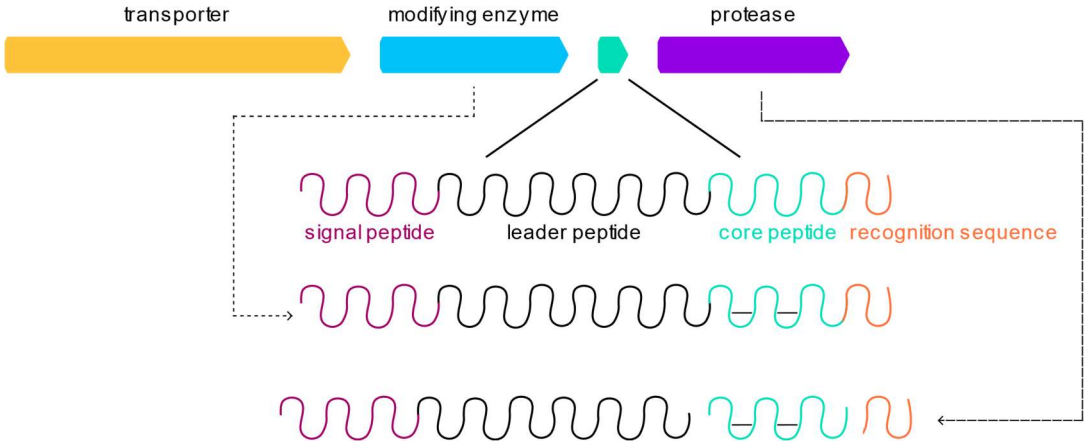


Figure 8 – Schematic representation of a putative RiPP BGC. Adapted from Kehr et al.[36].

Additionally to the core biosynthetic enzymes, Cyanobacteria biosynthetic pathways contain tailoring enzymes responsible for the immense chemical and structural diversity of their secondary metabolites. Halogenases, monooxygenases, desaturases and acyltransferases are examples of tailoring enzymes responsible for the modification of cyanobacterial secondary metabolites.[46]

The rich structural diversity of cyanobacterial secondary metabolites and the wide range of bioactivities they display have been boosting the research on cyanobacterial NPs. Despite that, the discovery of new compounds has been hampered by the ability to culture these microorganisms in laboratory. As such, metagenomics rises as a promising approach to uncover the biosynthetic potential of this phylum.[28]

1.4. Metagenomics

Traditionally, microorganisms have been characterized through genome sequencing of isolates in a culture-dependent manner.[47] Nevertheless, some characteristics of microbial habitats are unreproducible in laboratory making it impossible to isolate and cultivate the microorganisms. Thus, the culture-dependent methodology leaves a great diversity of microbes unexplored.[48]

Metagenomics is a field of research that focuses on the study of nucleotide sequences derived from environmental samples – environmental DNA (eDNA).[49,50] Hence, this culture-independent approach poses as an alternative to traditional methodologies that can enable the sequencing and identification of uncultured microorganisms.[47] The latest improvements in computational tools and the reducing costs of genome sequencing have been pushing the analysis of eDNA towards the *de novo* assembly of metagenome-assembled genomes (MAGs).[47,51] By taking advantage of next-generation sequencing techniques, it is possible to recover sequencing reads and assemble them into contigs.[51] The resulting contigs are then grouped together to represent single organisms (candidate MAGs) in a process named binning.[47,51] This process relies on the analysis of complimentary marker genes, abundances, tetranucleotide frequencies (TNFs)[52], codon usage[53] and taxonomic alignments[54]. The resulting MAGs can be evaluated as high (>90% completeness and <5% contamination), medium (90%>completion>50% and 5%<contamination<10%) or low (completion <50% or contamination >10%) quality MAGs.[55] Usually, it is only possible to obtain high quality MAGs for the most abundant species in the sample. In general, the quality obtained for MAGs is inferior to that achieved for genomes sequenced from bacterial isolates, mostly due to the complex composition of metagenomic samples, the variation on the species abundance across the sample and the difficulty in attributing conserved genomic features

to a specific MAG.[51] The strategies to recover MAGs incorporate sequence quality control (QC), assembly of metagenomes and QC of the assembly, and metagenome binning.[47] The standard workflow for MAG recovery is represented in Fig. 9.

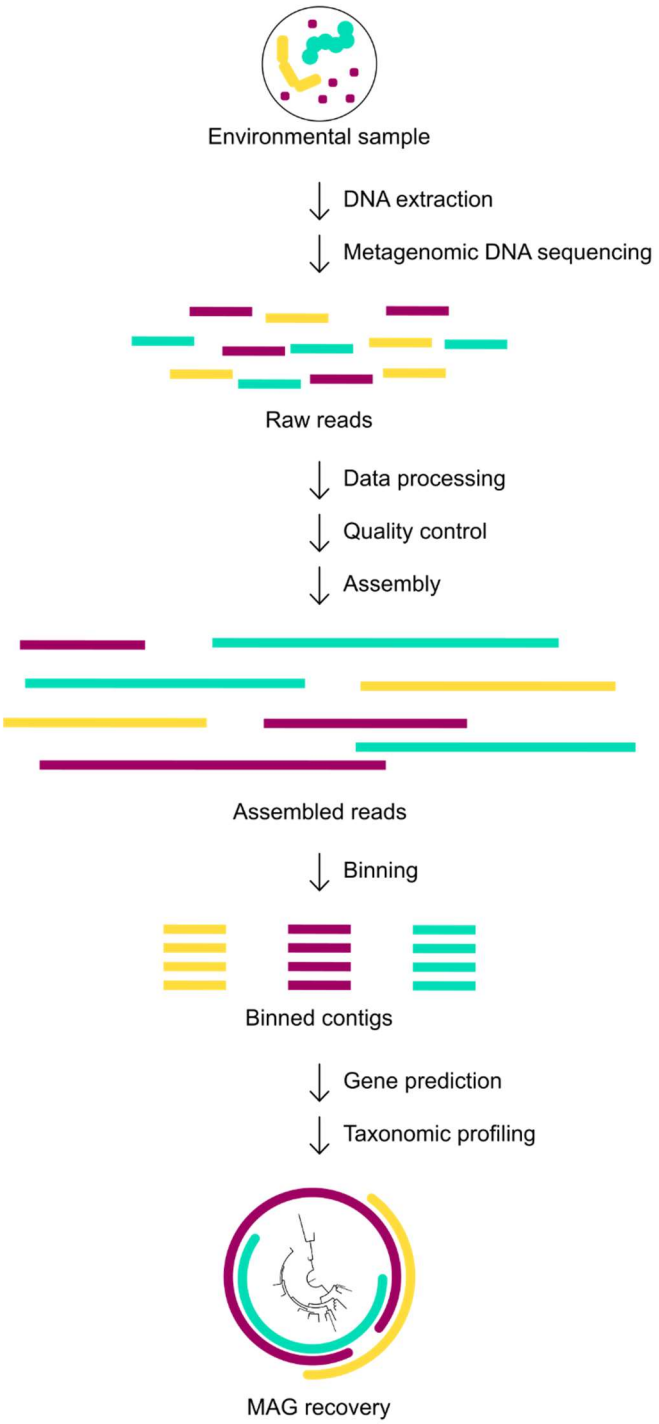


Figure 9 – Schematic representation of the workflow for MAG recovery from environmental samples. Adapted from Yang et al. 2021[47].

1.4.1. Quality Control

MAG construction starts with the QC of the sequencing data derived from metagenomics in order to remove contaminating nucleotide sequences.[47] Firstly, it is important to distinguish between long- and short-read sequencing.

Short-read sequencing produces nucleotide sequences up to 600 base pairs (bps) long.[56] Short-read technologies include NextSeq, MiSeq, HiSeq and NovaSeq (Illumina)[57-60]; Ion Torrent sequencers (Thermo Fisher)[61]; or BGISEQ, MGISEQ and DNBSEQ models (BGI).[62-64] The market leader Illumina systems work with a sequencing by synthesis technology (SBS).[57-60] In this technique, the sequence of linear DNA fragments is determined base by base using fluorescence.[65] BGISEQ, MGISEQ and DNBSEQ models from BGI are based on DNA nanoballs technology.[62-64] In this technology, special enzymes and a circular DNA template are used to amplify short DNA fragments. The resulting fragments form a continuous DNA molecule composed by tandem repeats of the same DNA sequence that is complementary to the circular template.[66,67] Unlike the previous short-read sequencing techniques, the Ion Torrent technology by Thermo Fisher uses a semiconductor chip to translate the pH change caused by a nucleotide incorporation into digital information. The voltage generated is proportional to the amount of incorporated nucleotides.[61]

Long-read sequencing technologies can generate reads with more than 10 kb.[56] These technologies include the Oxford Nanopore Technologies (ONT) developed nanopore sequencing[68] and the Pacific Biosciences (PacBio) single-molecule real-time (SMRT) sequencing[69]. Nanopore sequencing developed by ONT is a real-time sequencing technique in which the bases that pass the nanopore cause a disruption of the current.[68] In SMRT sequencing developed by PacBio adapters are added to the double stranded DNA. The DNA is placed in a SMRT cell containing several wells (named zero-mode waveguides). With the addition of a nucleotide to the DNA strand, fluorescence is emitted and measured to determine the DNA sequence.[69]

During the QC for short-read sequencing it is necessary to remove adapter sequences, generate quality assessment metrics, filter low-quality reads and process enrichment bias.[47] QC software can be distinguished between toolkit-like tools which consist of multiple executables arranged in a sequential manner and workflow-like software in which the functions are arranged in a predefined sequence as an integral workflow. Toolkit-type tools consume a long time to generate intermediate files and to read them, thus hindering acceleration of the QC process. Workflow-type software must be arranged in an optimal sequence once it cannot be altered by the user. When using

this type of QC tools it is only possible to detect if the preprocessing was not adequate for a given dataset in the end of the analysis.[70] SOAPnuke is a 2 step QC workflow-type tool capable of performing trimming, filtering and accessing base composition distribution for each position, read-quality, Q20 and Q30, and distribution of quality score.[70] Q value corresponds to the PHRED score, which follows equation 1, that represents the probability (p) of a base call (identification of the incorporated nucleotide) during sequencing being incorrect. The error rate diminishes as the Q value increases.[71]

$$p = 10^{(-Q/10)} \text{ (equation 1)}^{71}$$

Trimommatic is a preprocessing tool optimized for Illumina data which allows individual functions that can be applied in a user defined order. It can work on paired reads or in individual reads. This bioinformatic tool is capable of performing trimming, quality filtering, removal of adapter sequences and other common QC actions.[72] FastQC is a toolkit-type tool used for the QC of high throughput sequencing data.[73,74] Quality is checked at nucleotide level, GC content, nucleotide bias and sequence length in a modular set of analysis.[73]

Once the sequencing principles are different for long-read sequencing, the bioinformatic algorithms for QC also change. Proovread is a hybrid correction tool developed for PacBio technology. This QC tool has the advantage of being suitable for use in normal computers. This tool uses high-quality short-reads (derived from Illumina sequencers) to correct errors in long reads. This algorithm incorporates a mapping tool used for sequence alignment and assessment of alignment quality. The QC analysis includes trimming, N50 determination analysis (minimum contig length needed to cover 50% of the genome[75]), correction and read length.[76] LongQC is a platform independent QC algorithm composed by distinct computational modules that can process ONT and PacBio sequencing reads. This tool skips the alignment process and uses *k*-mer (DNA substrings of length *k* in a longer DNA fragment)[77] based internal overlaps thus functioning properly without a reference genome. The statistical output includes GC content, quality statistics, estimation of per-read base error and adapter statistics.[78] SequelTools contains three functions: read filtering, QC and read subsampling. Read filtering allows the exclusion of minimum continuous long read (CLR – continuous sequence produced by the polymerase during sequencing) length and/or low-quality reads. By using the read subsampling tool it is possible to subsample reads by random CLR selection or longest subreads per CLR.[79] The QC function produces multiple statistics and quality plots that include N50, count statistics and read length.[79] It is a free tool developed for PacBio sequencing data.[79]

Moreover, certain tools, such as BBMap, can be used for both long- and short-read sequencing QC. BBMap is a fast and extremely accurate splice-aware global aligner tool that can be used for DNA sequencing reads. This bioinformatic program can be used even with genomes that are highly mutated or reads with long indels (insertion or deletion of nucleotides). BBMap does not have an upper limit of number of contigs or genome size and, when compared to other alignment tools, it has a very fast indexing phase. The output statistic files include genome coverage, empirical read quality histograms and insert-size distribution.[80] The decontamination of the reads can then be performed using BBDuk analysis (Duk – Decontamination using *k*-mers). BBDuk is a single high-performance bioinformatic algorithm capable of performing adapter-trimming, GC-filtering, quality-score recalibration, quality-trimming and filtering, sequence masking, length filtering, contaminant-filtering via *k*-mer matching, histogram generation and several other operations.[81]

The adequate tool for each project should be selected by the user.[78] After performing the QC step it is time to assemble the metagenomes.[47]

1.4.2. Metagenome assembly

Assembly is the bioinformatic step in which overlapping reads derived from sequencing are organized in contigs that, subsequently, can be grouped in scaffolds to reconstruct the original DNA sequence.[82] The strategies to assemble metagenomes from long-reads differ from those used for short-reads.[47] Bioinformatic tools for short-read assembly include Omega,[83] MEGAHIT[84] and metaSPAdes.[85] Omega uses an overlap-layout consensus (OLC) approach, in which overlaps between reads are identified, to generate metagenome assembled genomes. Contigs and scaffolds are generated using mate-pair information.[83] MEGAHIT v1.0 uses a compressed version of De Bruijn Graph (DBG) coupled to the *k*-mer selecting process to reduce the memory usage during read assembly. MEGAHIT v1.0 uses a multiple *k*-mer size approach in an attempt to overcome the constraints caused by *k*-mer length choosing.[84] MetaSPAdes starts by using SPAdes to construct a DBG of all the reads. The DBG is then simplified to an assembly graph. Afterwards, paths that correspond to large genomic fragments from a metagenome are reconstructed in the assembly graph.[85]

Long-read sequencing from ONT and PacBio present a higher rate of base errors than short-reads.[47] As such, the bioinformatic tools for assembly of long-read sequencing data are different from those used for short-reads and include Canu[86] and metaFlye[87].[47] These long read assembly systems possess modules for base error correction.[47] This tools can follow either a “assembly then correction” or a “correction then assembly”

strategy. In the “assembly then correction” the correction is done using the assembled genome. This strategy might increase assembly errors. In the “correction then assembly” strategy the algorithm firstly corrects the reads and afterwards does the genome assembly using the corrected reads. This strategy is usually slower but can result in accurate genome assemblies.[88] Canu is a “correction then assembly” bioinformatic tool that reduces the run time required, improves assembly and decreases the depth requirements.[86,88] The output assures that complex genomes can be automatically and completely assembled through a combination of long-range scaffolding information and highly resolved assembly graphs.[86] MetaFlye is also an “assembly then correction” tool.[87,88] MetaFlye analyses global *k*-mer distribution, performs the global *k*-mer counting and detects repeat edges in the output graphs of metagenome assembly. Moreover, MetaFlye is capable of distinguishing between species that share very similar genomes.[87]

To facilitate the assembly and genotyping processes, mapping tools, such as SAMtools[89] and Bowtie[90] can be employed to align the sequencing reads to reference genomes.[89]

Nonetheless, single scaffolds only rarely represent complete genomes. Thus, the scaffolds resultant from assembly are submitted to a binning process[91].

1.4.3. Metagenome Binning

Binning is the process of grouping scaffolds into clusters according to their organism of origin[47,91]. GroopM[92], CONCOCT[93] and MaxBin2[94] are examples of binning tools. GroopM compares the read depths and TNFs of different scaffolds to bin them into metagenomes. This bioinformatic tool generates bins automatically but allows manual refinement of the generated bin cores.[92] CONCOCT performs binning according to coverage across samples and sequence composition, once the characteristic *k*-mer signature varies among different organisms.[93] MaxBin 2.0 recovers bins from the co-assembled reads from distinct metagenomic samples by measuring the contigs TNFs and coverage.[94]

The bioinformatic tool anvi’o is an interactive platform for the analysis and visualization of ‘omics data. This platform combines several algorithms to create an easy to follow assembly-based metagenomic workflow. Automated and human-managed binning and pos-binning manual refinement are examples of the many functions executed by this program.[95]

After the binning process, the bins are evaluated according to the contamination with single-copy genes and the marker gene completeness. Usually, the recovered

MAGs with high and medium quality are submitted to gene annotation. Commonly, this analysis is carried out using CheckM.[47]

1.4.4. Gene prediction and annotation

MAG annotation comprises gene prediction, gene functional annotation, taxonomic classification and profiling.[47] The bioinformatic tools developed for gene prediction include Prodigal[96], MetaGeneMark[97] and Meta-MFDL[98]. Prodigal scores every gene with above 90 bps by examining on the open reading frame (ORF) the GC bias on all three codon positions. Afterwards, the program chooses between highly overlapping ORFs. The output comprehensive list of gene coordinates may also include protein translations.[96] MetaGeneMark predicts nucleotide frequencies to reconstruct codon frequencies and recreates *k*-mer frequencies to uncover genes in short sequences.[97] Meta-MFDL combines various features, such as, mono-codon and mono-amino acid usage, to identify ORFs in genomic sequences. Moreover, this program can distinguish between coding and non-coding ORFs according to ORF length.[98]

The identified genes are then submitted to annotation. BLAST derived programs and Hidden Markov Models based tools, such as PANZZER2, are used to compare the retrieved ORFs with know genes.[47] The identification of new genes has to be performed using gene context-based tools once these genes present no homology to those previously described. GeConT and FlaGs are examples of such bioinformatic tools.[47]

1.4.5. Taxonomic classification of MAGs

Taxonomic classification methodologies that rely on 16S rRNA small subunit genes are not always suitable for MAGs due to poor representation of 16S rRNA genes and limited resolution of the method. Instead, tools that use single-copy marker genes, such as GTDB-Tk and PhyloPhlAn 3.0, have been employed in MAG taxonomic classification. HMMER[99] is used by GTDB-Tk[100] to find, in reference genomes present at the Genome Taxonomy Database (GTDB), marker genes that are used to construct a reference tree. The taxonomic classification of a MAG is attributed according to average nucleotide identity, position in the domain-specific reference tree and relative evolutionary divergence to the reference genome.[100] PhyloPhlAn 3.0 uses more than 80,000 reference genomes and 150,000 MAGs as reference to identify species specific marker genes. Maximum-likelihood phylogeny is used for taxonomic classification.[101]

Finally, MAG abundance can be estimated by using tools based on *k*-mers (Kraken), single nucleotide polymorphism (StrainFinder), marker genes (MetaPhlAn3) or protein (Kaiju).[47]

Metagenomic approaches have been employed to identify different taxa that compose environmental samples. New taxa have been described through metagenomic analysis of eDNA.[102] Moreover, the retrieved MAGs can be bioinformatically analyzed to access their biosynthetic potential, namely by pinpointing genome regions corresponding to BGCs. The use of this strategy has allowed the exploration of uncultured microorganisms from both symbiotic relationships and free-living forms.[102-106] Furthermore, it is possible to capture the biosynthetic pathways from eDNA through cloning, PCR amplification or gene synthesis into plasmids and reveal their products through heterologous expression in suitable hosts.[3,107-109]

1.5. Heterologous expression of MAG-derived BGCs

The study of uncultured microorganisms poses as a reliable path towards the discovery of new chemistry. Heterologous expression into model hosts rises as a promising approach to uncover the metabolites encoded in BGCs directly recovered from eDNA. Fig. 10 schematically represents the cloning strategies used for the heterologous expression of MAG-encoded BGCs. Most of the research dedicated to the study of compounds derived from uncultured microorganisms takes advantage of clone libraries, especially those constructed from fosmid vectors and Bacterial Artificial Chromosomes (BACs).[110] Cosmids have also been used to heterologously express MAG-derived BGCs.[109] Cosmids can be defined as plasmid vectors equipped with *cos* sites which are required for DNA packaging into bacteriophage lambda particles. These vectors do not possess the genes necessary to the biosynthesis of infection particles, as such, cosmids can hold DNA inserts up to 40 kb. Moreover, cosmid vectors contain an origin of replication so that they can replicate in *E. coli* cells and an antibiotic resistance cassette for selection of positive transformants. Cosmid vectors are used to construct libraries of DNA. Nonetheless, these are high copy number plasmids, meaning that they recruit a large portion of the cell machinery to replicate which can translate into an unstable propagation on the host cell.[111] Similar to cosmid vectors, fosmids have the *cos* site necessary for packing DNA into phage lambda particles. Fosmids are low copy number plasmids due to the F-factor origin of replication thus allowing a more stable transformation. The selection of positive transformants can be done based on antibiotic resistance.[112] Fosmid plasmids can hold DNA fragments with up to 45 kbs.[113] BACs are plasmids modified with the *E. coli* F-factor derived origin of replication that were constructed to uphold DNA fragments larger than 300 kb. Commonly, these plasmids are able maintain and replicate inserts with 100 to 200 kb in a stable way.[114] All these strategies can be employed to construct eDNA libraries that can posteriorly be screened to uncover new NPs. Once all the proteins required for NP biosynthesis in

microorganisms are encoded together in the genome, eDNA clones with large inserts have a high chance of containing fully functional BGCs. Nonetheless, as the cloning libraries are generated in a random way, promising BGCs for heterologous expression can be split into different clones, thus hampering the expression of the encoded compound.[108,109,115] Moreover, the screen of metagenomic libraries requires the sequencing of the transformants and the subsequent heterologous expression of the compounds of interest may require new cloning into a new vector.[107,108,115]

De novo DNA synthesis is a promising approach to the heterologous expression of eDNA-derived BGCs as it circumvents the constraints associated with working with the organisms or their eDNA. This strategy allows for codon optimization and manipulation of transcriptional regulatory elements (ribosome binding sites, promoters and terminators).[116,117] Nonetheless, de novo DNA synthesis is still unfeasible for BGCs rich in GC and with repetitive sequences. Moreover, due to its high cost, this technique has been only applied to the heterologous expression of BGCs smaller than 10 kb.[116] This technique has been employed to study a NRPS family produced by the gut microbiota[118] and to express a metagenome-derived novel chitinase[117].

Direct Pathway Cloning coupled with Sequence- and Ligation-Independent Cloning (DiPaC-SLIC) is a recently developed synthetic biology technique that allows the heterologous expression of microbial BGCs in *E. coli*.[119] In this technique long-amplicon polymerase chain reaction (PCR) is used to amplify the target BGC and to introduce the homologous overhangs to enable cloning into the selected vector. The primers designed for PCR amplification introduce the nucleotide sequence complementary to the vector to enable cloning. The commonly used vector contains a tetracycline inducible promoter, PtetO, and a *gfp* gene to track BGC expression.[119] As such, this technique allows the amplification of target BGCs directly from the eDNA and is a promising alternative to the currently recognized synthetic biology techniques for the heterologous expression of BGC derived from environmental samples. DiPaC-SLIC has already been used to express cyanobacterial BGCs in *E. coli*.[119,120] However, to the best of our knowledge, this is the first report on the use of this synthetic biology tool to heterologously express MAG-derived gene clusters.

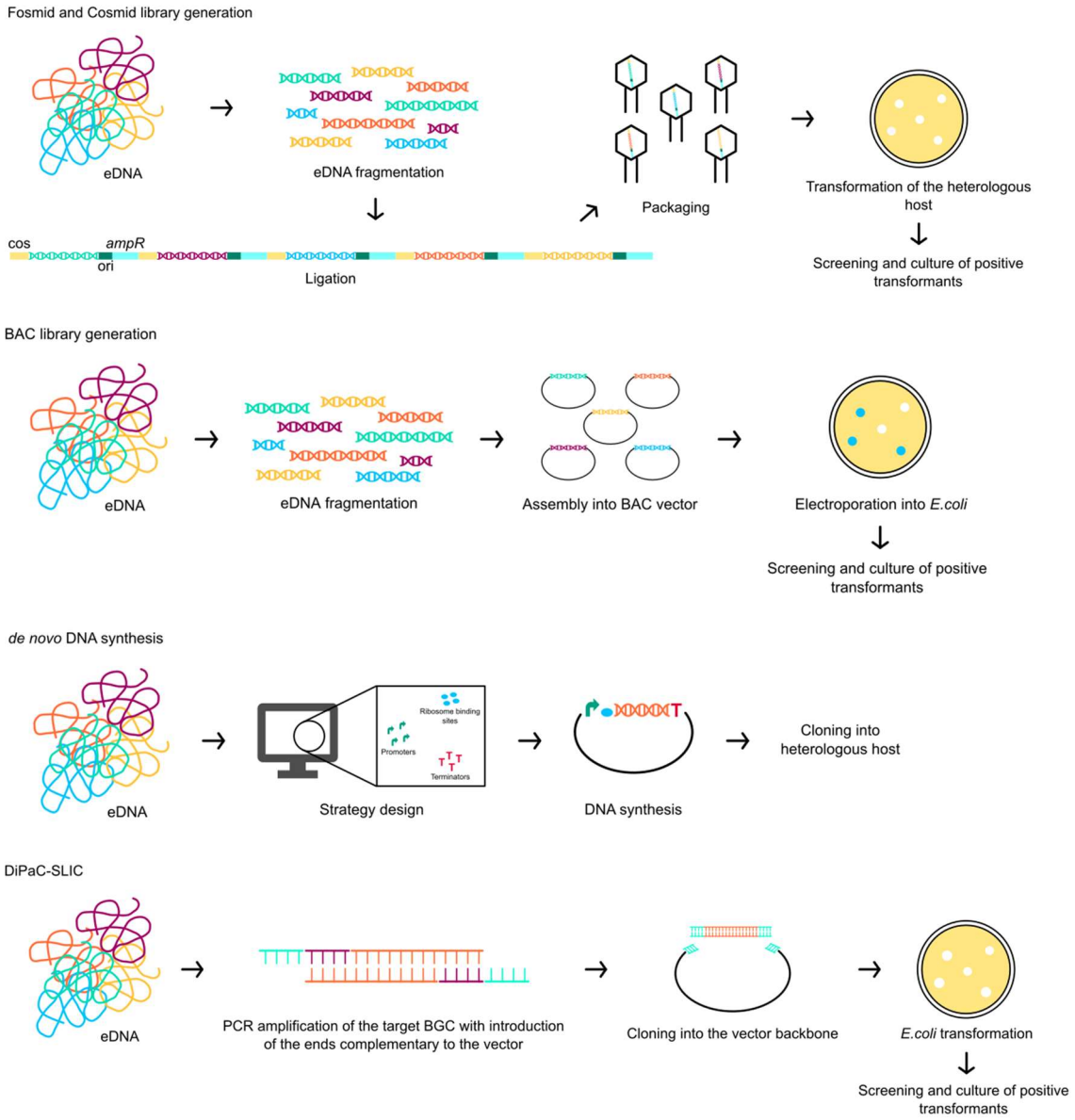


Figure 10 – Schematic representation of the cloning strategies used to access the biosynthetic potential of uncultured microbes.

1.6. Aim of this work

The aim of this work is to find a proof of concept that DiPaC-SLIC is suitable for the heterologous expression of cyanobacterial MAG-derived BGCs. For that, a dark green lake biofilm sample recovered from Parque da Cidade, Matosinhos, Portugal was analysed. Shotgun metagenomic data revealed three cyanobacterial MAGs from which we recovered 39 complete/near-complete BGCs. From those, five BGCs from different biosynthetic classes were selected to undergo heterologous expression in *E. coli* through DiPaC-SLIC. A microviridin BGC was successfully expressed in *E. coli* BL21 (DE3). Another microviridin BGC is currently being cloned into *E. coli* TOP10. Moreover, genes from a NRPS BGC were successfully amplified from the recovered eDNA.

2. Material and Methods

2.1. Sampling, eDNA extraction and sequencing

The sample used in this study consisted of a dark green floating cyanobacterial biofilm collected at a lake at Parque da Cidade do Porto, Matosinhos, Portugal (41.167485, -8.677425). eDNA from the sample was extracted using PowerSoil® DNA Isolation Kit (MO BIO Laboratories, Inc.). DNBseq sequencing technology (PE 150bp) at BGI genomics was used for shotgun metagenomic sequencing resulting in a 10 GB clean data file.

2.2. Metagenome pre-processing and quality verification

Paired-end reads were quality trimmed at Q20 and BBDuk tool from BBMap (https://sourceforge.net/projects/bbmap/) was used to remove sequences shorter than 45 bps. These parameters resulted in 0 trimmed reads because the sequencing company had already quality clipped the reads.

2.3. Metagenome-assembled genomes

MEGAHIT v.1.2.9[84] was used to assemble paired-end reads to recover MAGs. Bowtie2 v.2.3.5.1[90,121] and samtools v.1.10[89] were used to perform mapping and the remaining steps were performed using Anvi'o v.6.1[95] according to the protocol "Metagenomic workflow" (http://merenlab.org/2016/06/22/anvio-tutorial-v2/). In short, "anvi-gen-contigs-database" was used to create a contigs database. Prodigal[96] was used to identify ORFs and genes from sample contigs that matched bacterial single-copy core genes were identified using HMMER[99]. Functional annotation of genes from the contigs database was performed using "anvi-run-ncbi-cogs". Profile was done using "anvi-profile". The standalone tool CONCOCT v.1.1.0[93] was used to perform binning. The resulting bins were imported to Anvi'o using "anvi-import-collection". Afterwards, "anvi-estimate-genome-taxonomy" was used to estimate taxonomy. Bins were visualized using "anvi-interactive" and "anvi-refine" was used for manual refinement. GTDB-Tk v0.3.3b[100] was used to attribute the final taxonomy to the refined bins. To determine the quality of the recovered MAGs "anvi-estimate-genome-completeness" and CheckM[122] were used. Refined bins with >90% completeness and <5% contamination were considered high-quality MAGs whereas those with 90%>completion>50% and 5%<contamination<10% were considered medium quality MAGs. Low quality MAGs (completion <50% or contamination >10%) were excluded.

2.4. Bioinformatic analysis of cyanobacterial BGCs

Putative cyanobacterial biosynthetic gene clusters were accessed using antiSMASH version 3[123].

BlastP (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins) was used to uncover the biosynthetic genes from the antiSMASH hits. Putative complete BGCs were selected for heterologous expression. RiPPMiner-Genome (http://202.54.226.242/~priyesh/rippminer2/new_predictions/index.php)[124] was used to analyse the precursor peptides of a cyanobactin BGC.

2.5. Genomic DNA extraction

Environmental samples were incubated with an EDTA solution at 26 ºC for 45 min prior to genomic DNA extraction. Liquid nitrogen was used for biomass maceration. Genomic DNA was extracted following the "NZY Plant/Fungi gDNA Isolation Kit" (NZYTech). The elution step was carried out using molecular biology grade water at 65 ºC. DNA concentrations were evaluated using Nanodrop DeNovix DS-11 FX.

2.6. Cloning strategy design for BGCs 52.1, 418.1 and 91.1

Geneious Prime 2020.1.2 and Geneious Prime 2022.0.2 (https://www.geneious.com) were used for analysing ORFs, primer design and development of the cloning strategies. The presence of transcriptional terminators on the putative biosynthetic gene clusters was bioinformatically analysed using ARNold (http://rssf.i2bc.paris-saclay.fr/toolbox/arnold/). Primer secondary structures and dimer formation were evaluated using OligoCalc (http://biotools.nubic.northwestern.edu/OligoCalc.html) and OligoEvaluator (http://www.oligoevaluator.com/LoginServlet).

BGCs were divided in fragments for the cloning strategy to facilitate PCR amplification and/or eliminate intergenic terminators. Two PCR amplifications were performed. The first PCR step was used to amplify the BGC fragments from the environmental DNA. The second PCR step was a nested PCR to introduce the complementary tails required for cloning. In Table 1, the primers designed for the first amplification are identified as BGC_Fx/Rx and primers designed for the second amplification are identified as BGC_T_Fx/Rx. The restriction site for PmlI (highlighted in **bold** on the sequence of the primers BGC_T_Fx/Rx) was added to the 3' end of each fragment to allow linearization of the vector containing the fragments.

Table 1 – Sequence, amplicon size and secondary structures of the primers designed for the cloning and heterologous expression of the selected BGCs.

Primer name	Sequence	Fragment size / bps	Secondary structures
52.1_F1	GCAAAAAGCCTTACAGATAAATC	2733	No
52.1_R1	CTCTAGGATGAAATATTACAACC		No
52.1_F2	CTTAAACTTCCAGATAATAAGTCTC	1813	No
52.1_R2	TTACGCACTCAAGTGTCATG		No
52.1_F3	AATTCTGTCATTCTGAGCTCC	1865	No
52.1_R3	GCTACTCCCTTAATTTTCGATC		No
52.1_F4	GAAAATTAAGGGAGTAGC	896	No
52.1_R4	GAAATTGTAACCTGGATG		No
52.1_T_F1	tcagtgatagagaagaggatcgaccAAAATGGATAGAAACGAATGG	2685	Weak
52.1_T_R1	cagttcttcacctttgctaaccatg**CACGTG**TCTGGAAGTTTAAGATTTTTTAAG		Moderate
52.1_T_F2	cttaaaaaatcttaaacttccaga**CAC**ATGGGTGATCTTGTTACC	1584	Weak
52.1_T_R2	cagttcttcacctttgctaaccatg**CACGTG**ATTTTAGGGAATTAGGGTATC		Moderate
52.1_T_F3	gccgataccctaattccctaaaat**CAC**ATGACGGTTTTAATTTTCACTTTC	1724	Moderate
52.1_T_R3	cagttcttcacctttgctaaccatg**CACGTG**GTCTACTTAACTATCTTCCCAATC		Moderate
52.1_T_F4	gattgggaagatagttaagtagac**CAC**CAGGAGAAAATAATGTTTGAGTCG	750	Weak
52.1_T_R4	cagttcttcacctttgctaaccatg**CACGTG**TTAGATCAAGGTGACTCCACC		Strong
418.1_F1	TACTTATTATCGCGGTGATTAG	2049	No
418.1_R1	GCATTCCTATTAGATTAGTTCAG		No
418.1_F2	TTAGCGATTATTCACGTTTTTGC	1784	No
418.1_R2	TCACTAAGAGAATTCCTGATTTG		No

Table 1 – Continuation.

Primer name	Sequence	Fragment size / bps	Secondary structures
418.1_F3	AGAATACCCCCTAAATCCTC	1447	No
418.1_R3	TTGACATGACCTAATTTCCTTAC		No
418.1_T_F1	tcagtgatagagaagaggatcgaccATGCAAAC TCAAATTACACCTGAC	2105	Moderate
418.1_T_R1	cagttcttcacctttgctaaccatg**CACGTG**AA ACGTGAATAATCGCTAATCTCG		Strong
418.1_T_F2	cgagattagcgattattcacgttt**CAC**TCTGAA CTAATCTAATAGGAATGC	1842	Weak
418.1_T_R2	cagttcttcacctttgctaaccatg**CACGTG**GT CCCCCTTAATCAGGAG		Moderate
418.1_T_F3	atcctcctgattaaggggac**CAC**CCAAGGAAG TTGCTCATG	1502	Moderate
418.1_T_R3	cagttcttcacctttgctaaccatg**CACGTG**AG AGCAATATCGGGACAG		Moderate
91.1_F1	GCAATCTTTTTCTTTCTTTCT	4462	No
91.1_R1	CCTGACCAGAAATATCAAATA		No
91.1_F2	CCAGAAAATAAACCCAAATCC	7194	No
91.1_R2	CAGTCGCCATAAGCTTAA		No
91.1_F3	GATTAAGCTTATGGCGAC	8840	No
91.1_R3	GAAAAATTGACACAAAAATCC		No
91.1_T_F1	tcagtgatagagaagaggatcgaccGCTGTTCC CTTTTATATGGTA	4430	Weak
91.1_T_R1	cagttcttcacctttgctaaccatg**CACGTG**CT CTAAAAAGCCACAATCAAG		Weak
91.1_T_F2	tatcttgattgtggctttttagag**CAC**CGAATG AAAGTAGTAGAA	7162	Weak
91.1_T_R2	cagttcttcacctttgctaaccatg**CACGTG**CA AGTTTAATTTGTTCCCAAATTT		Moderate
91.1_T_F3	tttgggaacaaattaaacttg**CAC**GACTGCTAT ACTTAAAGAAAC	8873	Weak
91.1_T_R3	cagttcttcacctttgctaaccatg**CACGTG**GT ATTAAACTTTCTCAAAGAACTT		Moderate

2.7. DiPaC-SLIC protocol, bacterial strains, and plasmids

All PCR amplifications were performed using a Veriti® 96-Well Thermal Cycler (ThermoFisher Scientific). Bacterial strains and plasmids used in this study are listed in Table 2. The vector backbone, pET28-ptetO::gfpv2, was linearized by PCR amplification. PCR mixture was prepared to a final volume of 50 µL containing a final concentration of 0.5 µM of forward and reverse primer, 1× Q5 High-Fidelity 2× Master Mix (NEB) and 70-100 ng of template DNA. Thermal cycling program was as follows: initial denaturation step at 98 ºC for 30 sec, followed by 35 cycles of a denaturation step at 98 ºC for 10 sec, annealing at 62 ºC for 30 sec, and extension at 72 ºC for 6 min, followed by a final extension step at 72 ºC for 10 min. The resulting PCR product was treated with DpnI (NEB) at 37 ºC for 1 h followed by a deactivation step at 65 ºC for 20 min, to remove vector plasmids that could re-circularize.

Fragments amplification was carried out as 12.5 µL PCR mixtures. For amplification of genes *mvdB-mvdC* and *mvdD-mvdF* from BGC 52.1 a final concentration of 0.5 µM of forward and reverse primer, 1× Q5 High-Fidelity 2× Master Mix (NEB) and 8-60 ng of template DNA was used. Thermal cycling program was as follows: initial denaturation step at 98 ºC for 30 sec, followed by 35 cycles of a denaturation step at 98 ºC for 10 sec, annealing at 52/55 ºC for 30 sec, and extension at 72 ºC for 2 min, followed by a final extension step at 72 ºC for 10 min. For the amplification of genes orf1-*mvdA* from BGC 52.1 Bin 108 and genes *mvdA*, *mvdB*-orf1-*mvdC* and *mvdED* from BGC 418.1 Bin 90.1 a final concentration of 0.4 µM of each primer, 0.2 mM dNTPs, 1 mM of $MgCl_2$, 8-60 ng of template DNA and 1 U of KOD DNA polymerase. Thermal cycling program was as follows: initial denaturation step at 95 ºC for 30 sec, followed by 35 cycles of a denaturation step at 95 ºC for 15 sec, annealing at 55 ºC for 30 sec for genes *mvdA* from BGC 418.1 Bin 90.1, at 57 ºC for 30 sec for genes orf1-*mvdA* from BGC 52.1 Bin 108 and *mvdB*-orf1-*mvdC* from BGC 418.1 Bin 90.1 and at 62 ºC for 30 sec *mvdED* from BGC 418.1 Bin 90.1, and extension at 72 ºC for 2 min, followed by a final extension step at 72 ºC for 10 min. PCR to introduce complementary tails was prepared at a final volume of 12.5 µL with a final concentration of 0.4 µM of each primer, 0.2 mM dNTPs, 1 mM of $MgCl_2$, 8-60 ng of template DNA and 1 U of KOD DNA polymerase. Thermal cycling programs were as follows: initial denaturation step at 95 ºC for 30 sec, followed by 35 cycles of a denaturation step at 95 ºC for 15 sec, annealing at 55/57 ºC for 30 sec, and extension at 72 ºC for 2 min, followed by a final extension step at 72 ºC for 10 min.

For the amplification of genes *aerA*, *aerB-aerDEF* and *aerG*-orf1-*aerM* from BGC 91.1 Bin 108 a PCR mixture with a final volume of 12.5 µL and a final concentration of 0.4 µM of each primer, 0.2 mM dNTPs, 1 mM of $MgCl_2$, 8-60 ng of template DNA and 1

U of KOD XL DNA polymerase was prepared. Thermal cycling program was as follows: initial denaturation step at 94 ºC for 30 sec, followed by 35 cycles of a denaturation step at 94 ºC for 30 sec, annealing at 52/55/57 ºC for 30 sec for gene *aerA*, at 52/55/62 ºC for genes *aerB-aerDEF* and 52/55 ºC for genes *aerG*-orf1-*aerM* and extension at 72 ºC for 7 min, followed by a final extension step at 72 ºC for 10 min. PCR to introduce complementary tails was prepared at a final volume of 12.5 µL with a final concentration of final concentration of 0.4 µM of each primer, 0.2 mM dNTPs, 1 mM of $MgCl_2$, 8-60 ng of template DNA and 1 U of KOD XL DNA polymerase. Thermal cycling program was as follows: initial denaturation step at 94 ºC for 30 sec, followed by 35 cycles of a denaturation step at 94 ºC for 30 sec, annealing at 52/57 ºC for 30 sec and extension at 72 ºC for 7 min, followed by a final extension step at 72 ºC for 10 min.

All fragments, as well as the vector backbone, were purified by gel band excision using NZYGelpure kit (NZYTech) and eluted in 30 µL of molecular biology grade water. SLIC reaction was prepared to a 10 µL final volume containing: 0.5 µL of T4 DNA polymerase (NEB), 1 X Buffer 2.1 (NEB), variable volumes of vector and fragments (concentrations variating between 0.02 and 0.5 pmol, according to Greunke et al.[125]) and molecular biology grade water. The SLIC reaction mixture was incubated for 35 sec at 50 ºC followed by 15 min on ice. 50 µL of *E. coli* TOP10 cells were transformed by heat shock with 5 µL of reaction mixture. Colony PCR was used to screen positive clones using the primer pairs listed in Table 3. Colony PCRs were performed using GoTaq® Flexi (Promega). The PCR mixture was prepared at a final volume of 20.0 µL and contained: 1x Green GoTaq® Flexi Buffer, 1.0 mM of $MgCl_2$ solution, 0.20 mM of PCR Nucleotide Mix, 0.20 µM of each primer, molecular biology grade water and template DNA picked directly from the colony. The PCR conditions were: initial denaturation step at 95 ºC for 5 min, followed by 35 cycles of a denaturation step at 95ºC for 45 sec, annealing at 48 ºC for 30 sec and extension at 72 ºC for 90 sec, followed by a final extension step at 72 ºC for 5 min. Clones that led to amplicons with the expected size were grown overnight in liquid LB medium supplemented with 50 µg mL^{-1} of kanamycin, with 180 rpm shaking, at 37 ºC. NZYMini-prep kit was used to isolate plasmids from overnight cultures (NZYTech). The elution step was carried out using molecular biology grade water at 65 ºC. The isolated plasmids were sent to sequencing to confirm the integrity of the nucleic acids sequence at ligation sites. Positive transformants were cryopreserved. For the insertion of a new fragment, the vector, containing already a set of genes, 1 µg of vector plasmid DNA was digested with PmlI at 37 ºC for 90 min followed by a deactivation step at 65 ºC for 20 min. Purification by gel excision and SLIC reaction were carried out as detailed above. Positive transformants were screen trough colony

PCR as described above. The integrity of the nucleic acids sequence at ligation sites was confirmed through direct sequencing. Clones with the correct DNA sequence were cryopreserved. For BGC 52.1 Bin 108, *E. coli* TOP10 transformed with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 were cryopreserved and the plasmid DNA was used for heterologous expression.

Table 2 – Strains and plasmids used in this study.

Strain	Description	Reference or source
E. coli TOP10	Host strain for cloning	ThermoFisher Scientific
E. coli BL21 (DE3)	Heterologous expression strain	NZYTech
Plasmid	**Description**	**Reference or source**
pET28b-ptetO::gfpv2 (6,552 bps)	Expression plasmid containg a tetracycline inducible promoter (PtetO) with a gfp reporter gene downstream of promoter, kanamycin resistance cassette, ColE1	Duell et. al[126]
pET28b-ptetO::orf1-*mvdA*-gfpv2 (9,205 bps)	pET28b-ptetO::gfpv2 cloned with orf1-*mvdA* genes from BGC 52.1 between PtetO and gfp gene	This study
pET28b-ptetO::orf1-*mvdABC*-gfpv2 (10,792 bps)	pET28b-ptetO:: orf1-*mvdA*-gfpv2 cloned with *mvdBC* genes from BGC 52.1 between the previous ligation site and gfp gene	This study
pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 (12,519 bps)	pET28b-ptetO:: orf1-*mvdABC*-gfpv2 cloned with *mvdDEF* genes from BGC 52.1 between the previous ligation site and gfp gene	This study
pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2 (13,272 bps)	pET28b-ptetO:: orf1-*mvdABCDEF*-gfpv2 cloned with orf2 genes from BGC 52.1 between the previous ligation site and gfp gene	This study
pET28b-ptetO::*mvdA*-gfpv2 (8,607 bps)	pET28b-ptetO::gfpv2 cloned with *mvdA* genes from BGC 418.1 between PtetO and gfp gene	This study
pET28b-ptetO::*mvdAB-orf1-mvdC*-gfpv2 (10,394 bps)	pET28b-ptetO::*mvdA*-gfpv2 cloned with *mvdB*-orf1-*mvdC* genes from BGC 418.1 between PtetO and gfp gene	This study

Table 3 – Sequence, amplicon size and secondary structures of the primers designed for the colony PCR used to confirm the correct insertion of the amplicons into the vector backbone.

Colony PCR primer	Primer sequence	Size / bp	Secondary structure
screen_ptetF2	TCCGACCTCATTAAGCAGC	413	No
52.1_colony_R1	CCACAATTGCTGCTGAGG		No
52.1_colony_F2	CCATACCACGGATCAAATTGC	553	No
screen_GFP_R	TTACCGTTGGTCGCATCACC		No
52.1_colony_F2	CCATACCACGGATCAAATTGC	670	No
52.1_colony_R2	ACCAATTAGGGTATTTTCTGC		No
52.1_colony_F3	TGCGGAGTCATTACGTTATTGTCC	517	No
screen_GFP_R	TTACCGTTGGTCGCATCACC		No
52.1_colony_F3	TGCGGAGTCATTACGTTATTGTCC	765	No
52.1_colony_R3	CGTTGATCGGGCTAACAATAAGCA		No
52.1_colony_F4	CCGGTTTCTGCGTAAGTCC	413	No
screen_GFP_R	TTACCGTTGGTCGCATCACC		No
52.1_colony_F4	CCGGTTTCTGCGTAAGTCC	559	No
52.1_colony_R4	TCTCCGATCGCCGTTACG		No
52.1_colony_F5	GAGGATACTAGCATCTGGAGC	324	No
screen_GFP_R	TTACCGTTGGTCGCATCACC		No
screen_ptetF2	TCCGACCTCATTAAGCAGC	386	No
418.1_colony_R1	TATCAAGCACATAGCGATTCC		No
418.1_colony_F2	GACAAAAAGCAATCGCGATCG	428	No
screen_GFP_R	TTACCGTTGGTCGCATCACC		No
418.1_colony_F2	GACAAAAAGCAATCGCGATCG	580	No
418.1_colony_R2	ACGTAGCCAATGAGGAGG		No

Table 3 – Continuation.

Colony PCR primer	Primer sequence	Size / bp	Secondary structure
418.1_colony_F3	TCCAAGAACTAATCCCCAAGC	521	No
screen_GFP_R	TTACCGTTGGTCGCATCACC		No
418.1_colony_F3	TCCAAGAACTAATCCCCAAGC	653	No
418.1_colony_R3	CACTCACCTAGAGGCTTAACC		No
418.1_colony_F4	TACCTATGATTGGCGAAAGGAAGG	424	No
screen_GFP_R	TTACCGTTGGTCGCATCACC		No

2.8. Heterologous expression of BGC 52.1 from Bin 108 (putative microviridin)

For heterologous expression experiments, *E. coli* BL21 (DE3) cells were transformed by heat shock with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2. The transformation with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 was confirmed through direct sequencing. *E. coli* BL21 (DE3) cells previously transformed with pET28b-ptetO::gfpv2 (empty vector) were used as control. 1 mL of a starter culture grown overnight in LB media supplemented with 50 µg mL^{-1} of kanamycin at 37 ºC was inoculated in 50 mL of M9 minimal medium in 100 mL erlenmeyer flasks. M9 minimum medium [3.0 g L^{-1} KH_2PO_4, 0.5 g L^{-1} NaCl, 17.1 g L^{-1} $Na_2HPO_4 \cdot 12\ H_2O$, 0.2 ml L^{-1} $CaCl_2$ (0.5 M), 1.0 ml L^{-1} $MgSO_4$ (2 M), 1.0 g L^{-1} NH_4Cl, pH 7.0] used for heterologous expression was supplemented with 10.0 mL L^{-1} of filtered glucose solution (40% (w/v)), 2 mL L^{-1} of filtered Vitamin mix (500×) and 50 µg mL^{-1} of kanamycin after autoclaving. Cultures were grown to an OD600 of 0.25-0.3 at 37 ºC with 180 rpm shaking. Afterwards the cultures were placed at 4 ºC for 30 min. Expression was tested with induction using 50 µL of 0.5 mg mL^{-1} tetracycline and without tetracycline induction, at 20 ºC and 37ºC with shaking at 180 rpm. 25 mL of culture medium were harvested by centrifugation (5000×*g*, 10 min at 4 ºC) at day 3. Extraction from cell pellets was carried out using MeOH (15 mL) during 2 h with shaking. The supernatant was recovered by centrifugation (5000 *xg*, 10 min at 4 ºC) and transferred to a previously weighted vial (8 mL). Extraction from culture supernatants was carried out using 0.4 mg of resin during 2 h with shaking. Resin was recovered by centrifugation (5000 *xg*, 10 min at 4 ºC), washed two times with distilled water and recovered by centrifugation (5000 *xg*, 10 min at 4 ºC). Finally, resin

was incubated with MeOH (15 mL) during 2 h with shaking. The supernatant was recovered by centrifugation (5000 *xg*, 10 min at 4 ºC) and transferred to a previously weighted vial (8 mL). Extracts were weighted and dissolved in MeOH to a concentration of 5 mg mL^{-1}. The resulting solutions were filtered (0.2 µm) and submitted to Liquid chromatography-high resolution electrospray ionization mass spectrometry (LC-HRESIMS) analysis. LC-HRESIMS data was generated in an UltiMate 3000 UHPLC (Thermo Fisher Scientific) system composed of a WPS-3000SL autosampler and a LPG-3400SD pump coupled to a Q Exactive Focus Hybrid Quadrupole-Orbitrap Mass Spectrometer controlled by Xcalibur 4.1 and Q Exactive Focus Tune 2.9 (Thermo Fisher Scientific). Full Scan mode with a capillary voltage set to -3.8kV, resolution of 70,000 FWHM, sheath gas flow rate to 35 units and capillary temperature to 300°C were used for LC-HRESIMS analysis. The system for LC-HRESIMS included an ACE 3 C8-300 (50 mm x 2.1 mm) column. The column oven was set to 40 °C. The flow rate for sample elution was 0.4 mL min^{-1}. Elution started at 98.0% solution A (99.9% H_2O, 0.1% HCOOH v/v) and 2.0 % solution B (99.9% acetonitrile, 0.1% HCOOH, v/v). After the first minute, a linear gradient from the initial conditions was used to reach 99% of solution B after 9 min. These conditions were maintained for 1.5 min. Afterwards, a linear gradient was established to returned to initial conditions after 1.5 min. Initial conditions were maintained for 2 min. Differences between *E. coli* carrying the empty vector pET28b-ptetO::gfpv2 and *E.coli* carrying the expression vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 were analyzed by manual inspection of the resulting data.

2.9. Large Scale heterologous expression of BGC 52.1

50 mL of a starter culture grown overnight in LB media supplemented with 50 µg mL^{-1} of kanamycin at 37 ºC was inoculated in 2.5 L of M9 minimal medium in 5 L erlenmeyer flasks. M9 minimum medium was prepared as described above. 10 L of culture were grown for large scale experiments. Cultures were grown to an OD600 of 0.25-0.3 at 37 ºC with 180 rpm shaking. Afterwards the cultures were placed at 4 ºC for 30 min. Expression was induced using 2.5 mL of 0.5 mg mL^{-1} tetracycline at 20 ºC with shaking at 180 rpm. The culture medium was harvested by centrifugation (5000 *xg*, 10 min at 4 ºC) at day 3. Pellets were divided by 50 mL flasks. Extraction from cell pellets was carried out two times using MeOH (35 mL) during 3 h with shaking. The supernatant was recovered by centrifugation (5000 *xg*, 10 min at 4 ºC) and transferred to a previously weighted vial (40 mL). Extraction from culture supernatants was carried out using 20 g of resin during 3 h on the 5 L erlenmeyers with shaking at 180 rpm. Resin was recovered by centrifugation (5000 *xg*, 10 min at 4 ºC), washed two times with distilled water and recovered by centrifugation (5000 *xg*, 10 min at 4 ºC). The resulting resin was divided by

50 mL flasks. Finally, the resin was incubated twice with MeOH (35 mL) during 3 h with shaking. The supernatant was recovered by centrifugation (5000 *xg*, 10 min at 4 ºC) and transferred to a previously weighted vial (40 mL). The resultant pellet and supernatant extracts were combined and submitted to a Solid Phase Extraction (SPE) in a Strata® C18-E (55 µm, 77 Å) column (Phenomenex®). The *E. coli* culture extract was dissolved with MeOH and C18 silica was added in a 1:1 (m/m) ratio. The MeOH was dried and the remaining powder was macerated prior to the SPE procedure. Five elutions were performed: 5% MeOH and 95% H_2O; 25% MeOH and 75% H_2O; 50% MeOH and 50% H_2O; 75% MeOH and 25 % H_2O; 100% MeOH. The resulting extracts were recovered separately and transferred to previously weighted vials (16 mL). The column was cleaned three times using MeOH and the resulting extract was recovered to a previously weighted vial (16 mL). Extracts were weighted and dissolved in MeOH to a concentration of 5 mg mL^{-1}. Samples were prepared for LC-HRESIMS has described above. LC-HRESIMS analysis was performed as described above.

3. Results and discussion

3.1. Metagenomic analysis of the lake biofilm sample allowed the recovery of three cyanobacterial MAGs

146 Bins were recovered from the dark green lake biofilm sample. The completeness and contamination (also known as redundancy) of each Bin were evaluated using both Anvi'o and CheckM. Once anvi'o estimated completeness and contamination was going to guide manual refinement, in the first cut-off, bins with a completeness inferior to 50%, according to anvi'o were eliminated. This resulted in the maintenance of bins 103 and 95, which would have been discarded if the cut-off was performed according to CheckM results. Bin 25 had more than 49% completeness according to both bioinformatic tools, and a contamination below 20%. Thus, the bin was selected for manual refinement as it could reach the standards required to be qualified as a medium-quality genome. In total, this preliminary analysis resulted in 26 recovered bins (Table 4). The taxonomy assignment for the recovered bins was performed using GTDB-Tk v0.3.3b (Fig. 11). The taxonomy analysis revealed that three of the recovered MAGs belonged to the phylum Cyanobacteria (Fig. 11), namely, bin 90, bin 108, bin 112. Bin 108 represented a high-quality MAG whilst bins 90 and 112 represented medium-quality MAGs (Table 4).

Manual refinement using the interactive script "anvi-refine" was performed for bins in which the decontamination could lead to an increase in overall MAG quality (Table 4). "anvi-refine" tracks the impact that the exclusion of contaminant sequences has on the completeness and contamination of a bin during manual refinement. Hence, it is possible to reduce bin contamination without lowering its completion in a way that affects the overall MAG quality. After this process, Anvi'o and CheckM were used to evaluate the completeness and contamination of each retrieved MAG (Table 5). According to Anvi'o, manual refinement allowed the recovery of 25 MAGs, from which, 11 represent high-quality MAGs and 14 represent medium-quality MAGs. The results from CheckM differ from those obtained using Anvi'o. According to CheckM, 14 MAGs were recovered after manual refinement, from which 5 are high-quality MAGs and 9 are medium-quality MAGs (Table 5). Manual refinement was performed according to Anvi'o completeness and contamination estimation. Therefore, it was expected that the results would be optimized in Anvi'o analysis. Nonetheless, the difference encountered was higher than expected. For several bins, manual refinement did not cause a significant decrease in contamination, thus not having an impact on the overall quality of the bin. Moreover, in

some cases, manual refinement caused a decrease on MAG completeness thus lowering bin quality.

However, the quality of the recovered cyanobacterial bins matches in both bioinformatic tools. Bin 90.1 and Bin 108 represent high-quality MAGs while Bin 112 represents a medium-quality MAG (Tables 4 and 5). The three cyanobacterial MAGs belong to the filamentous cyanobacterial order Oscillatoriales. Bin 112 represents a cyanobacterium from the *Planktothricoides* genus and Bin 108 belongs to the *Planktothrix* genus. Bin 90.1 could not be classified at the genus level. Information on the recovered MAGs is presented in Table 6.

Table 4 – Length and percentage of completion and contamination of the recovered Bins. The cyanobacterial bins are highlighted in **orange**.*submitted to manual refinement.

Bin	Anvio			CheckM		
	Total length / bps	Completion / %	Redundancy / %	Total length / bps	Completion / %	Redundancy / %
40*	5001296	98.59	49.30	5001296	95.01	48.32
89*	9560381	98.59	305.63	9560381	94.83	227.51
90*	**6776506**	**98.59**	**9.86**	**6776506**	**97.38**	**2.34**
102	3334117	97.18	0.00	3334117	94.56	1.00
140	4583015	97.18	1.41	4583015	96.97	1.76
32*	6381040	95.77	12.68	6381040	96.45	2.33
30*	7305705	94.37	88.73	7305705	96.21	77.08
108	**5495089**	**94.37**	**4.23**	**5495089**	**98.97**	**1.52**
134*	2866613	93.42	73.68	2866613	87.85	57.40
68*	4655816	91.55	35.21	4655816	87.62	27.22
122*	2871362	91.55	9.86	2871362	75.38	3.14
56*	4659144	80.28	61.97	4659144	70.60	35.57
112	**7147993**	**80.28**	**9.86**	**7147993**	**79.43**	**4.01**
31	2359340	78.87	1.41	2359340	72.74	0.05
107*	10183916	77.46	33.80	10183916	86.33	59.54
26*	3442711	73.24	60.56	3442711	61.89	39.50
88*	3358063	73.24	22.54	3358063	78.79	26.49
105	2076919	69.01	4.23	2076919	54.04	3.14
13*	2800733	64.79	19.72	2800733	65.62	25.51
22	2687535	64.79	9.86	2687535	75.29	6.09
103	1608769	54.93	4.23	1608769	39.14	2.95

Table 4 – Continuation.

Bin	Anvio			CheckM		
	Total length / bps	Completion / %	Redundancy / %	Total length / bps	Completion / %	Redundancy / %
95*	1677452	53.52	26.76	1677452	35.19	12.02
96*	2219840	53.52	5.63	2219840	57.27	14.28
98	1479001	52.63	0.00	1479001	57.82	0.65
69	675536	50.70	1.41	675536	51.00	1.23
25*	5874802	49.30	15.49	5874802	49.25	17.01

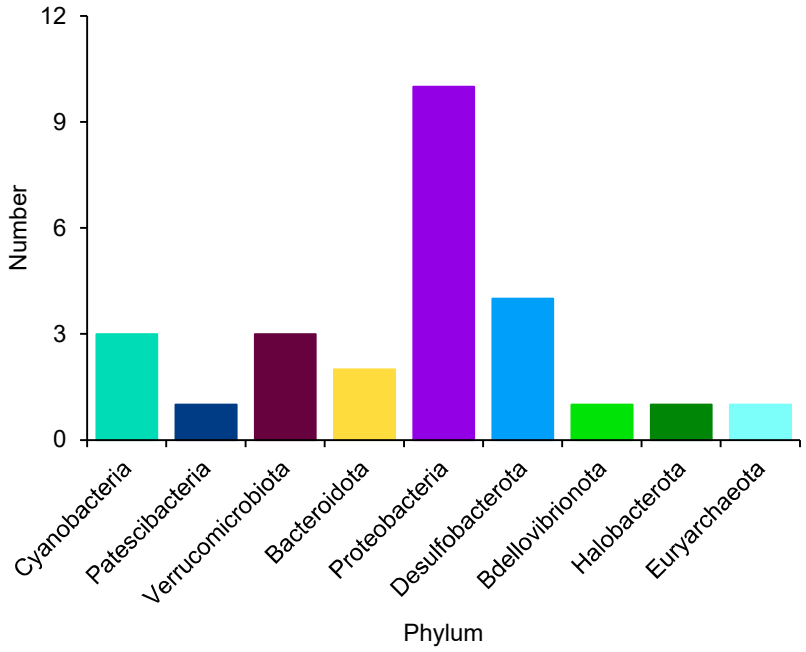


Figure 11 – Graphical representation of the MAG taxonomy assignment performed using GTDB-Tk v0.3.3b.

Table 5 – Length and percentage of completion and contamination of the refined Bins. The cyanobacterial bin is highlighted in **orange**.

Bin	Anvio			CheckM		
	Total length / bps	Completion / %	Redundancy / %	Total length / bps	Completion / %	Redundancy / %
90.1	**6727695**	**97.18**	**4.23**	**6727695**	**95.63**	**1.90**
32.1	6346806	95.77	2.82	6346806	96.45	1.61
30.1	6762233	94.37	1.41	6762233	96.21	37.07
89.1	9327595	92.96	2.82	9327595	94.32	161.03

Table 5 – Continuation.

Bin	Anvio			CheckM		
	Total length / bps	Completion / %	Redundancy / %	Total length / bps	Completion / %	Redundancy / %
40.1	4899380	92.96	2.82	4899380	90.87	38.99
134.1	2689760	92.11	1.32	2689760	87.47	4.29
122.1	2864847	91.55	2.82	2864847	75.36	2.48
68.1	4611400	90.14	0.00	4611400	82.24	9.88
26.1	3402256	73.24	4.23	3402256	54.75	17.84
13.1	2789275	64.79	5.63	2789275	47.65	8.34
107.1	10077034	77.46	0.00	10077034	85.80	53.72
56.1	4615869	74.65	8.45	4615869	60.17	23.36
88.1	3333238	73.24	2.82	3333238	77.82	21.93
96.1	2214602	53.52	1.41	2214602	57.10	13.36
95.1	1664938	52.11	4.23	1664938	17.76	1.89
25.1	5862970	49.30	2.82	5862970	48.88	15.90

Table 6 – Taxonomy classification of the cyanobacterial MAGs recovered from the lake biofilm sample.

MAG	Order	Family	Genus	Abundance (in the sample)*
Bin 108	Oscillatoriales	Phormidiaceae	Planktothrix	12.67
Bin 112	Oscillatoriales	Oscillatoriaceae	Planktothricoides	0.42
Bin 90.1	Oscillatoriales	Phormidiaceae	Non-classified	3.86

*mean coverage of a contig divided by overall sample coverage

3.2. Bioinformatic analysis of the recovered cyanobacterial MAGs revealed thirty-nine complete/near-complete BGCs

The recovered cyanobacterial MAGs were analysed in antiSMASH to uncover its BGC content. The analysis identified a total of 39 BGCs in the three MAGs, namely, 13 NRPS, 2 hybrid PKS/NRPS, 10 RiPPs, 5 terpenes and 9 from other biosynthetic classes. The BGC distribution across the three MAGs is represented in Fig. 12. More detailed information on the BGCs identified in each MAG is listed in Tables 7, 8 and 9.

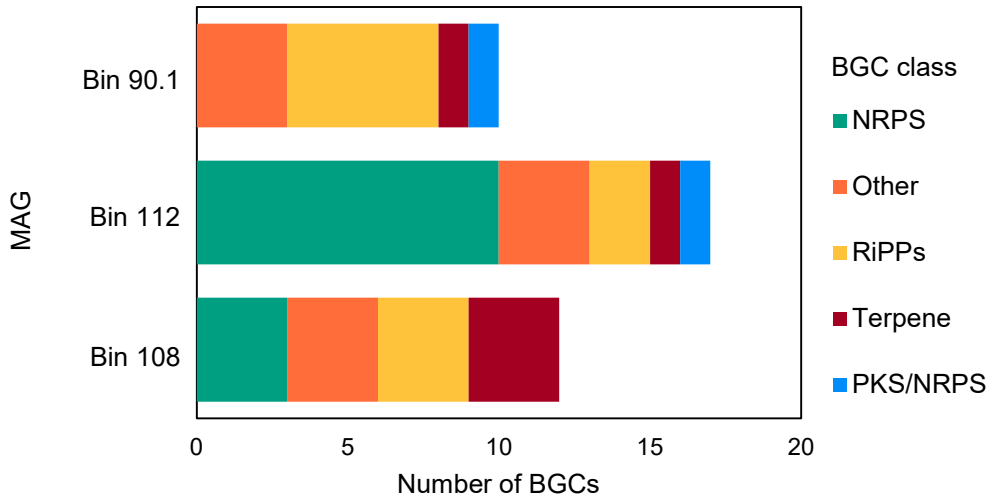


Figure 12 – Distribution of the different BGC classes across the cyanobacterial MAGs.

Table 7 – Putative BGCs identified by antiSMASH analysis in Bin 108.

BGC	Type	Most similar known cluster		Similarity / %	Complete cluster / Size
9.1	RRE-containing	-	-	-	Yes / 20,462 nt
52.1	microviridin	microviridin K	RiPP	100	Yes / 17,169 nt
66.1	cyanobactin	prenylagaramide B / prenylagaramide C	RiPP:Lanthipep-tide	47	Yes / 15,447 nt
74.1	cyanobactin	prenylagaramide B / prenylagaramide C	RiPP:Lanthipep-tide	17	No / 11,858 nt
91.1	NRPS	aeruginoside 126B / aeruginoside 126A	NRP:Glycopepti-de + Polyketide:Other + Saccharide:Hy-brid/ tailoring	35	No / 23,097 nt
96.1	NRPS, betalactone	anabaenopeptin	NRP	100	Yes / 46,613 nt
106.1	terpene				No / 11,134 nt
107.1	ladderane	-	-	-	Yes / 27,400 nt
161.1	terpene				Yes / 19,381 nt
191.1	NRPS-like	anabaenopeptin NZ857 / nostamide A	NRP	100	No / 21,912 nt
206.1	terpene	-	-	-	Yes / 21,950 nt
218.1	NRPS	cyanopeptin	NRP	75	No / 8,993 nt

Table 8 – Putative BGCs identified by antiSMASH analysis in Bin 112.

BGC	Type	Most similar known cluster		Similarity / %	Complete cluster / Size
74.1	NRPS	cyanopeptolin	NRP	85	No / 8,242 nt
125.1	NRPS-like				No / 16,770 nt
144.1	RRE-containing				Yes / 6,894 nt
266.1	NRPS	microcystin	NRP + Polyketide	33	No / 12,958 nt
271.1	NRPS				No / 10,727 nt
318.1	cyanobactin	microcyclamide	RiPP:Cyanobactin	22	Yes / 5,913 nt
523.1	NRPS,T1PKS	puwainaphycin A / puwainaphycin B / puwainaphycin C / puwainaphycin D	NRP + Polyketide	70	Yes / 28,029 nt
675.1	NRPS				No / 9,481 nt
806.1	NRPS	hectochlorin	NRP:Lipopeptide + Polyketide: Modulear Type I	25	Yes / 14,224 nt
851.1	cyanobactin				No / 3,797 nt
853.1	RRE-containing,cyanobactin	tenuecyclamide A / tenuecyclamide C	RiPP:Cyanobactin	57	? / 7,317 nt
873.1	ectoine				Yes / 7,145 nt
911.1	terpene				Yes / 8,367 nt
912.1	NRPS	pseudospumigin A / pseudospumigin B / pseudospumigin C / pseudospumigin D / pseudospumigin E / pseudospumigin F	NRP + Polyketide	66	No / 13,482 nt
947.1	NRPS				No / 5,249 nt
1071.1	NRPS-like				No / 3,412 nt
1163.1	NRPS	hapalosin	NRP:Cyclic depsipeptide + Polyketide : Modular type I	40	No / 3,718 nt

Table 9 – Putative BGCs identified by antiSMASH analysis in Bin 90.1.

BGC	Type	Most similar known cluster		Similarity / %	Complete cluster / Size
78.1	cyanobactin	piricyclamide	RiPP	75	Yes / 12,948 nt
83.1	RRE-containing				No / 10,887 nt
121.1	cyanobactin	piricyclamide	RiPP	25	Yes / 5,308 nt
251.1	terpene				No / 11,937 nt
418.1	microviridin	microviridin K	RiPP	75	Yes / 12,401 nt
439.1	T1PKS,NRPS				? / 18,971 nt
445.1	cyanobactin				No / 9,321 nt
448.1	microviridin				Yes / 17,179 nt
525.1	ladderane				Yes / 13,048 nt
636.1	RiPP-like				Yes / 7,786 nt

Complete BGCs were considered candidates for heterologous expression. Based on size, biosynthetic class and on previous successful heterologous expression attempts conducted in our lab, BGC 52.1 (microviridin – RiPP) and BGC 91.1 (NRPS) from Bin 108, BGC 523.1 (T1PKS/NRPS) from Bin 112 and BGC 418.1 (microviridin – RiPP) from Bin 90.1 were selected for further bioinformatic characterization. Additionally, manual inspection of the BGCs identified by antiSMASH analysis led us to hypothesise that BGC 66.1 and BGC 74.1 from Bin 108 represented a single cyanobactin (RiPP) BGC. As such, we also selected this BGCs for bioinformatic characterization.

3.3. Bioinformatic analysis of the BGCs selected for heterologous expression

Promising BGCs were selected for heterologous expression. BlastP was used to annotate the biosynthetic gene products from the selected BGCs.

3.3.1. BGC 52.1, Bin 108 – putative microviridin

Bioinformatic analysis of BGC 52.1 (Bin 108) indicated similarities with microviridin gene clusters. Microviridins are cyclic depsipeptides produced by cyanobacteria composed of thirteen to fourteen amino acids. This family of RiPPs usually exhibits protease inhibitory activity.[127] Gene annotation of BGC 52.1 is listed in Table 10 and the BGC architecture is represented in Figure 13. MvdA is an ATP-binding cassette transporter which is pointed as a scaffolding protein responsible to maintain the biosynthetic complex on the cytosolic side of the membrane.[128] MvdB is a putative

acetyltransferase that transfers an acetyl group from acetyl-CoA to diverse substrates.[129] In microviridin biosynthesis, MvdB transfers an acetyl group to the N-terminal amino acid.[127,130] MvdC is responsible for the cyclization of the compound by forming a conserved amide bond between the amino acids Lys and Asp/Glu/Gly. MvdD catalyses cyclization trough the formation of ester bonds between Thr and Asp and/or Ser and Glu.[127] These enzymes recognize and interact with a conserved PFFARFL motif on the leader peptide.[128] *mvdE* and *mvdF* encode the precursor peptides.[130] The amino acid sequence of the core peptide of MvdE corresponded to the sequence of microviridin I. Microviridin I is a group I microviridin, meaning that it has two ester bonds: one between Thr2 and Asp1, and another between Ser and Glu.[130,131] The second precursor, MvdF, did not show similarities with previously described microviridins (Table 11).[130] Moreover, the core peptide encoded by *mvdF* has two unique amino acids in conserved positions when compared to the already characterized microviridins.[130] Namely, a tryptophan (W) instead of a tyrosine (Y) on the seventh position and a serine (S) instead of a tyrosine (Y), tryptophan (W) or phenylalanine (F) on the fourteenth position from the N-terminal (Table 11). These changes do not interfere with the amino acids required for ester bond formation.[130]

Microviridin I was firstly isolated from *Planktothrix agardhii* extracts and lacks a corresponding BGC.[131] Nonetheless, MIBiG comparison analysis demonstrated that the structure of BGC 52.1 was similar to that of Microviridin K BGC. Phylmus and co-workers characterized the biosynthetic gene cluster responsible to produce microviridin K, initially isolated from a *Planktothrix agardhii* extract.[127] *mvdA-E* are attributed to the biosynthesis of the isolated oligopeptide. In this BGC there is also a second precursor peptide whose product is not detected on the cyanobacterial extract. A methyltransferase and a short chain dehydrogenase flank the gene cluster, upstream and downstream, respectively. These genes are hypothesized to belong to an ancient anabaenopeptilide pathway and do not play a role on the microviridin biosynthesis.[127] Nonetheless, as the second peptide was not found on the isolate, we cannot discard the possibility of these genes being a part of the biosynthesis of the second depsipeptide. As such, the strategies designed for the heterologous expression of BGC 52.1 include the two orfs (Orf1 and Orf2). The possible chemical structures for the compounds encoded in BGC 52.1 were predicted according to microviridin biosynthesis (Fig. 14 and 15). As there are no evidence of the possible role for Orf1 and Orf2 on the biosynthesis, at this point the possible modifications introduced by the encoded enzymes were not considered.

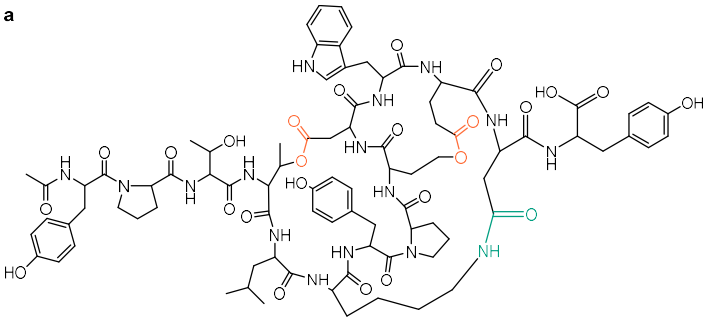
Table 10 – BlastP homologies for BGC 52.1 from Bin 108.

Gene	Proposed function	Organism	Identity / %	Similarity / %	Size / aa
orf1	Putative methyltransferase	*Planktothrix agardhii* No758	95.85	98	265
mvdA	ATP-binding cassette transporter	*Planktothrix agardhii*	98.83	99	597
mvdB	GNAT family N-acetyltransferase	*Planktothrix agardhii*	97.77	98	179
mvdC	ATP-grasp ribosomal peptide maturase	*Planktothrix agardhii*	99.69	100	324
mvdD	ATP-grasp ribosomal peptide maturase	*Planktothrix agardhii* KL2	99.39	99	330
mvdE	Microviridin I prepeptide	*Planktothrix agardhii*	91.67	95	48
mvdF	Microviridin prepeptide	*Planktothrix agardhii*	100.00	100	50
orf2	Putative acyl-carrier protein	*Planktothrix agardhii* NIES-204	98.78	99	245

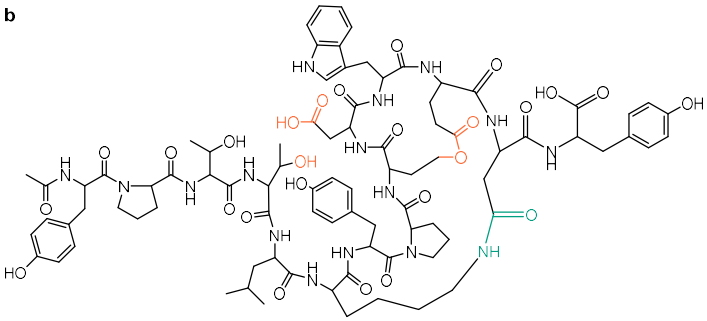
Figure 13 – Architecture of BGC 52.1.

Table 11 – Amino acid sequence of the precursors encoded in BGC 52.1. Highlighted in **green** is the conserved motif of the leader peptide recognized by the ATP-grasp ligases. Highlighted in **bold** are the conserved amino acids for microviridin core peptides. Highlighted in **red** are the amino acids that differ from those that are already described for this RiPPs in that position.

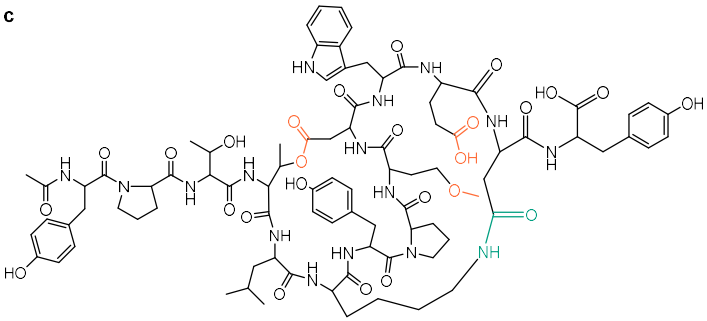
Gene	Precursor peptide sequence	Leader peptide sequence	Core peptide sequence
mvdE	MSKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	MSKNVKVSAPKAV**PFFARF**LAEQAVEANNSNSAP	YPT**T**L**KYPSDW**ED**Y**
mvdF	MSKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	MSKNIKVSTGSAV**PFFARF**LSEQDTETGDSTSTDIP	TIW**T**F**KW**PSDWED**S**



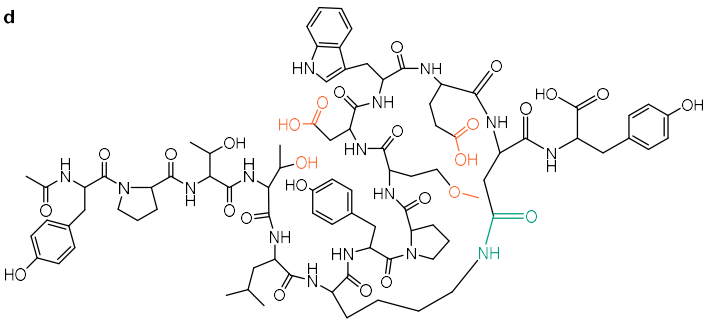
Exact Mass: 1778.7828



Exact Mass: 1796.7934

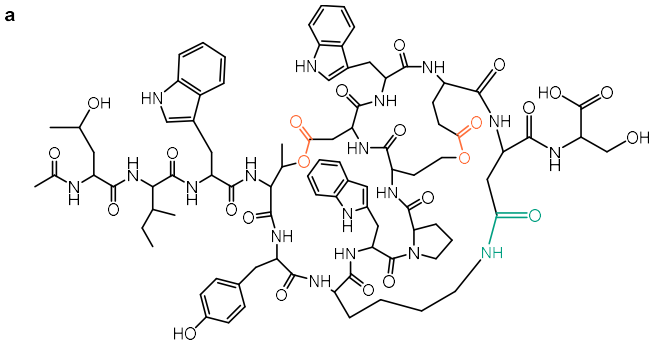


Exact Mass: 1810.8090

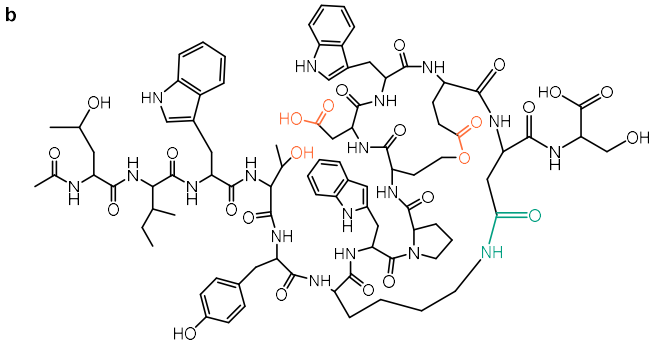


Exact Mass: 1828.8196

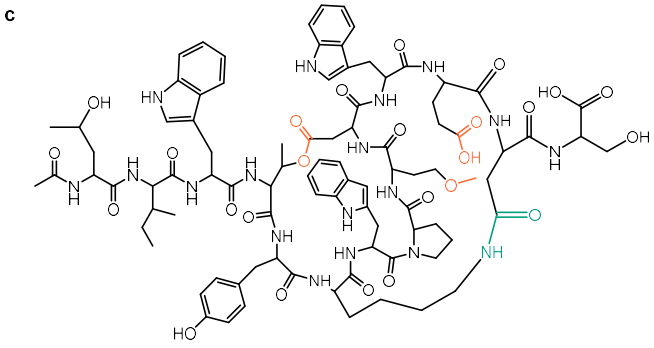
Figure 14 – Predicted structures for the microviridins encoded by *mvdE*. The atoms that participate in the ester bonds are represented in orange. The atoms that participate in the amide bond are represented in green. **a:** two ester bonds; **b:** one ester bond between serine and glutamate; **c:** one ester bond between tyrosine and aspartate; **d:** no ester bonds.



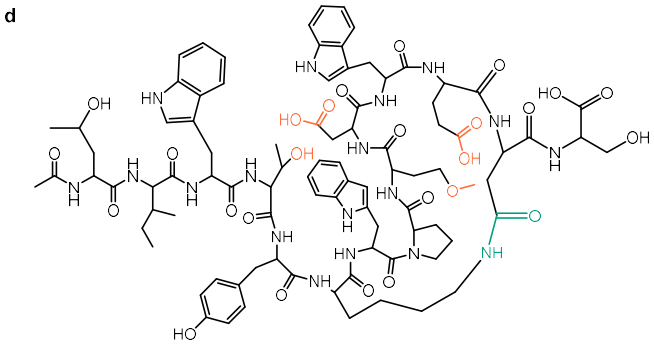
Exact Mass: 1828.8097



Exact Mass: 1846.8203



Exact Mass: 1860.8359



Exact Mass: 1878.8465

Figure 15 – Predicted structures for the microviridins encoded by *mvdF*. The atoms that participate in the ester bonds are represented in orange. The atoms that participate in the amide bond are represented in green. **a:** two ester bonds; **b:** one ester bond between serine and glutamate; **c:** one ester bond between tyrosine and aspartate; **d:** no ester bonds.

3.3.2. BGC 418.1, Bin 90.1 – putative microviridin

The bioinformatic analysis of BGC 418.1 (Bin 90.1) also revealed similarities with microviridin gene clusters. Gene annotation is listed in Table 12 and the BGC architecture is represented in Figure 16. This gene cluster only encodes one precursor peptide, *mvdE*. Although all the amino acids from the conserved motifs are present, the amino acid sequence of the core peptide is different from those described to date (Table 13).[130] The first two amino acids of the core peptide sequence, tryptophan and serine, are not present in any microviridin previously reported. Moreover, the gene cluster organization does not show similarities to known microviridin clusters.[130] The strategy to clone and heterologously express this BGC was devised to include all genes listed in Table 12. Having in mind the microviridin biosynthesis, the possible chemical structures for the compound encoded in BGC 418.1 were predicted (Fig. 17).

Table 12 – Gene annotation for BGC 418.1 from Bin 90.1.

Gene	Proposed function	Organism	Identity / %	Similarity / %	Size / aa
mvdA	ATP-binding cassette transporter	*Planktothrix sp.* PCC 11201	82.74	92	677
mvdB	GNAT family N-acetyltransferase	*Planktothrix*	81.01	87	179
orf1	No significant similarity found				37
mvdC	ATP-grasp ribosomal peptide maturase	*Planktothrix sp.* UBA10369	83.59	91	326
mvdE	Microviridin prepeptide	*Planktothrix agardhii*	70.83	85	47
mvdD	ATP-grasp ribosomal peptide maturase	*Planktothrix agardhii* KL2	82.52	92	326

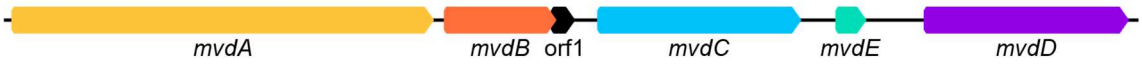


Figure 16 – Architecture of BGC 418.1.

Table 13 – Amino acid sequence of the precursor encoded in BGC 418.1. Highlighted in **green** is the conserved motif of the leader peptide recognized by the ATP-grasp ligases. Highlighted in **bold** are the conserved amino acids for microviridin core peptides.

Gene	Precursor peptide sequence	Leader peptide sequence	Core peptide sequence
mvdE	MSKNVKATAPQTVPFFARFLEE QAAKSNSSNAPWSQTLKYPSDW EEY	MSKNVKATAPQTV**PFFARFL** EEQAAKSNSSNAP	WSQ**T**L**KYPSDW**EE**Y**

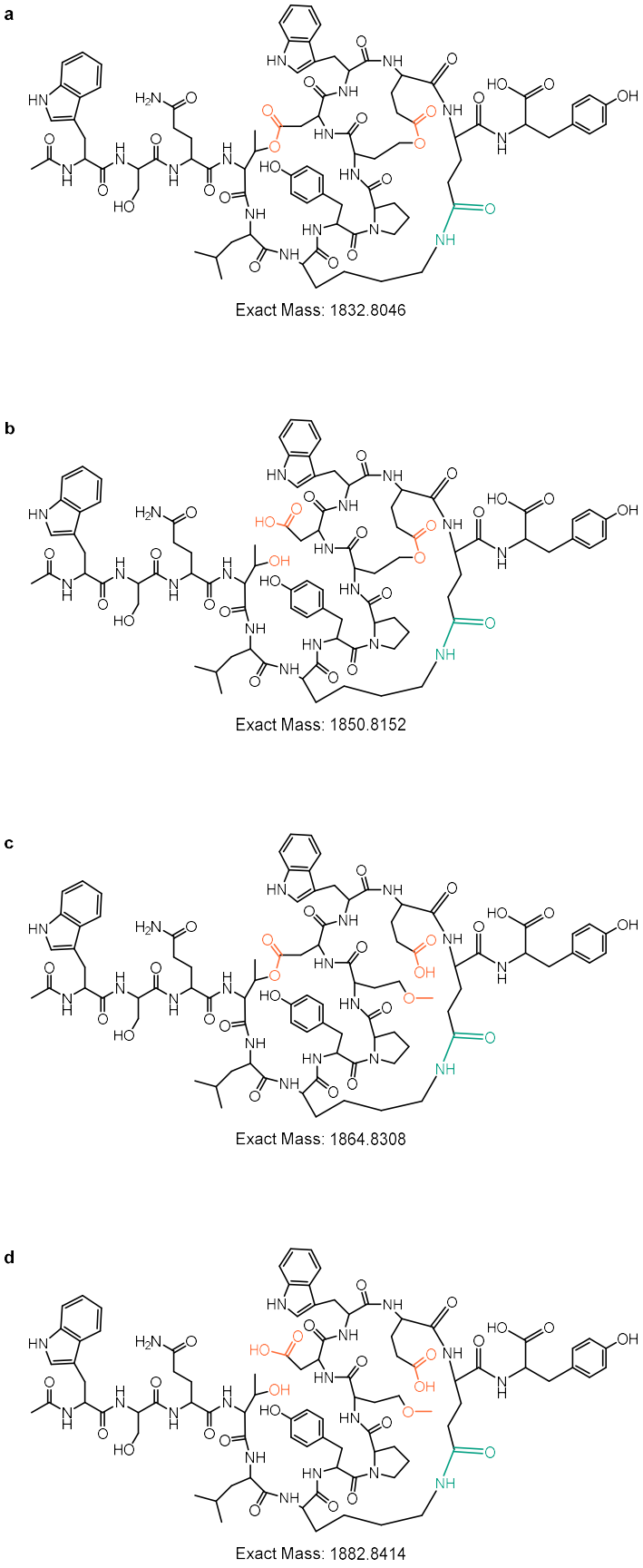


Figure 17 – Predicted structures for the microviridins encoded by *mvdF*. The atoms that participate in the ester bonds are represented in orange. The atoms that participate in the amide bond are represented in green. **a:** two ester bonds; **b:** one ester bond between serine and glutamate; **c:** one ester bond between tyrosine and aspartate; **d:** no ester bonds.

3.3.3. BGC 66.1 and BGC 74.1, Bin 108 – putative cyanobactin

The antiSMASH analysis of Bin 108 revealed two putative cyanobactin BGCs. Cyanobactins are a family of RiPPs in which the precursor peptide (PatE) is composed of the leader peptide and the final hypervariable peptide sequence (core peptide) flanked by the highly conserved recognition sequences (RS) that direct the activities of the biosynthetic enzymes. The core peptide together with RSII and RSIII forms the cassette.[132] Apart from Cyanobacteria, these ribosomal peptides have been identified in several marine animals and could not be linked to a symbiont cyanobacterium. Despite that, cyanobactin BGCs were only encountered in members of the phylum Cyanobacteria. Cyanobactin BGCs length ranges from 8 to 19 kbs. The final peptide structure consists of 6 to 20 amino acids.[133] A single *patE* homolog might contain one or more cassettes encoding distinct natural products. The first cyanobactin BGC described was the patellamide cluster composed of genes *patA-G*. The subtilisin-like serine proteases PatA/G are regulated by RSII and RSIII, respectively. PatA and PatG are subtilisin-like serine proteases implicated in the maturation of cyclic cyanobactins. PatA cleaves the peptide after the RSII releasing the N-terminal end of the peptide.[132] The RSII recognized by PatA and homologs include the amino acid sequences: AVLAS, GVDAS, GLTPH, GLEAS, and GVEPS.[134] PatA is composed by a C-terminal domain of unknown function (DUF) and a protease domain.[132] PatG binds to the RSIII to free the C-terminal end of the core peptide and further catalyses the C-N macrocyclization of the final peptide sequence.[132] PatG and its homologs recognize the following RSIII: AYD, AYDGE, FAGDDAE, SYD, SYDD, and SYEGDEAE.[134] To occur macrocyclization, the last residue of the core peptide sequence must be an azoline or the amino acid proline preceded by an L-amino acid. Moreover, for proper cyclization, the N-terminal amino acid of the core peptide must be a nucleophilic amine. PatG has C-terminal and N-terminal DUFs, a flavin-dependent oxidase domain and a protease domain.[132] The heterocyclase PatD is directed by RSI. This last recognition sequence is only present in the cassette when the BGC has a PatD homolog.[132] This enzyme catalyses the conversion of one or more serine and/or threonine residues to the corresponding azoline. PatD and its homologs are composed of three domains: RiPP precursor peptide recognition element (RRE) that recognizes and binds to the leader peptide; E1-like domain which is a docking scaffold; and YaCO protein that catalyses the activation of the amide backbone in an ATP dependent manner to enable heterocyclization.[132] PatF is an ABBA-type prenyltransferase that is present in all cyanobactin BGCs except the one that encodes the non-prenylated trichamide.[132,133] Despite being non-prenylated, PatF is essential for the synthesis of patellamides raising the hypothesis of these enzyme having a distinct activity in biosynthetic pathways that lack prenylation.[133] Nonetheless,

several PatF homologs are responsible for the addition of a prenyl or genranyl group, in reverse or forward orientation, to tyrosine (Y), serine (S), tryptophan (W) and threonine (T) residues from the core peptide. The biosynthetic step catalysed by these enzymes was shown to be the final step through heterologous expression in *E. coli*. Most of the F-family enzymes act on the cyclic peptide. Still, one PatF homolog is able to prenylate the α-amine at the N-terminus from the linear peptide. These prenyltransferases act on a variety of peptide sequences whereas the isoprene donor is always dimethylalyl pyrophosphate or isopentenyl diphosphate.[132] PatB and PatC are two widely distributed proteins throughout cyanobactin BGCs without a known role on the biosynthesis of these RiPPs. These enzymes are not required for *in vitro* synthesis nor for *in vivo* production of cyanobactins when the enzymes are expressed under the influence of different promoters. Nonetheless, if the BGC is expressed as a sole construct in *E. coli* these enzymes are required. Additionally, cyanobactin biosynthetic pathways may include oxidases and/or methyltransferases, which can be attached to the PatG, as well as several proteins of unknown function.[132]

Manual inspection of the two hits demonstrated that both were incomplete cyanobactin BGCs. BGC 66.1 had in its composition two prenylated precursor peptides from the anacyclamide family (Table 14). Anacyclamides are a family of low molecular weight cyanobactins firstly described in *Anabaena* strains. As such, the characterization of this cyanobactin BGC followed the nomenclature for the anacyclamide BGC in which: AcyA corresponds to PatA; AcyB to PatB; AcyE to PatE; AcyF to PatF; and AcyG corresponds to PatG. The anacyclamide BGC lacks a PatD homolog,[135] thus, this cyanobactin BGC should also lack this gene. Furthermore, BGC 66.1 possesses a AcyG homolog (Table 14). According to the general composition of cyanobactin BGCs, these BGC is missing another AcyA homolog, as well as AcyB, AcyC and AcyF homologs. Analysis of the amino acid sequence from the precursor peptides demonstrated that these lack the RSI (Table 15), further demonstrating that this BGC may not include a PatD homolog. Moreover, both precursors have a conserved proline residue on the C-terminus. As such, a peptide length ranging from 7 to 20 amino acids is expected.[133] BGC 74.1 was composed by AcyA, AcyB and AcyC homologs (Table 16). As such, these putative BGC lacks, at least, the precursor peptide. The two putative cyanobactin BGCs appeared to complement each other, so the subtilisin-like serine proteases were used as a query against complete *Planktothrix* genomes from the NCBI database to verify if the two proteins clustered together. The results showed that homologs for the two proteins belong to the same cluster in *Planktothrix* genomes. The architecture of those BGCs is similar to that of BGC 66.1 together with BGC 74.1 (Fig. 18 a and b).

Furthermore, the AcyF homolog, which is not present in neither BGC 66.1 nor BGC 74.1, is in between the two subtilisin-like serine proteases in the other *Planktothrix* BGCs (Fig. 18 b). This indicated that there was a missing sequence in the reported cyanobactin BGCs. Primers were designed to recover the nucleotide sequence located between the AcyA and AcyG homologs. However, this analysis was not concluded during the development of this thesis. Moreover, according to the composition of some of the homolog BGCs, here an additional precursor peptide might be present. The core peptide sequence as well as the RSII and RSIII for AcyE2 were predicted according to the precursor peptide sequences described in literature (Table 15). It was not possible to perform this prediction for the precursor peptide AcyE1 based on literature. Hence, the bioinformatic tool RiPPMiner-Genome was used to analyse the amino acid sequence of this precursor. The output core peptide sequence is listed in Table 15. The RSII sequence annotated in Table 14 consists of the 5 amino acids that precede the core peptide. RiPPMinner-Genome analysis of AcyE2 confirmed the literature-based prediction. The possible chemical structures for the cyanobactins encoded by *acyE1* and *acyE2* are represented in Fig. 19 and 20. The predicted core peptide sequence of AcyE1 lacks amino acids that are prenylated by AcyF – S, W, Y, T – thus the designed structure does not contain a prenyl group. AcyE2 contains three T residues, as such, it can contain more than one prenyl group. All the predicted structures for this peptide are represented in Fig. 20. No further experimental work was developed with this BGC during the development of this thesis.

Table 14 – Gene annotation for BGC 66.1 from Bin 108.

Gene	Proposed function	Organism	Identity / %	Similarity / %	Size / aa
orf1	No significant similarity found				69
orf2	Hypothetical protein A19Y_3483/Uma2 family endonuclease	*Planktothrix agardhii* NIVA-CYA 126/8/*Plankothrix agardhii*	98.96/98.96	99/99	192
acyE1	Hypothetical protein NIES204_32140/Anacyclamide/piricyclamide family prenylated cyclic peptide	*Planktothrix agardhii* NIES-204/*Planktothrix*	63.93/63.93	75/70	59
orf4	Type II toxin-antitoxin system HicA family toxin	*Planktothrix agardhii*	97.30	100	74

Table 14 – Continuation.

Gene	Proposed function	Organism	Identity / %	Similarity / %	Size / aa
orf5	Type II toxin-antitoxin system HicB family antitoxin	*Planktothrix sp.* PCC 11201	98.53	100	68
orf6	Hypothetical protein	*Planktothrix rubescens*	100.00	100	75
orf7	Hypothetical protein	*Planktothrix agardhii*	100.00	100	91
acyE2	Anacyclamide/piricyclamide family prenylated cyclic peptide	*Planktothrix*	98.04	100	51
acyG	PatA/PatG family cyanobactin maturation protease	*Planktothrix*	95.94	98	690

Table 15 – Amino acid sequences of the precursor peptides. Highlighted in **orange** are the predicted RSII. Highlighted in **blue** are the predicted core peptides. Highlighted in **green** are the predicted RSIII.

Gene	Precursor peptide sequence
acyE1	MTKKNLKPQQTAPVQREINTTSIPHSEDGTT**GLIPQ**LILKDGREKGIFGGFP**FAGDDAE**
acyE2	MIKKNIRPQQSAPVQRQVTTTSCQGEK**ALFAS**TECDILTKQCTP**FAGDNAE**

Table 16 – Gene annotation for BGC 74.1 from Bin 108.

Gene	Proposed function	Organism	Identity / %	Similarity / %	Size / aa
acyC	Cyanobactin biosynthesis PatC/TenC/TruC family protein	*Planktothrix agardhii*	89.23	92	65
acyB	Cyanobactin biosynthesis system PatB/AcyB/McaB family protein	*Planktothrix*	97.06	98	68
acyA	PatA/PatG family cyanobactin maturation protease	*Planktothrix agardhii*	97.90	99	618

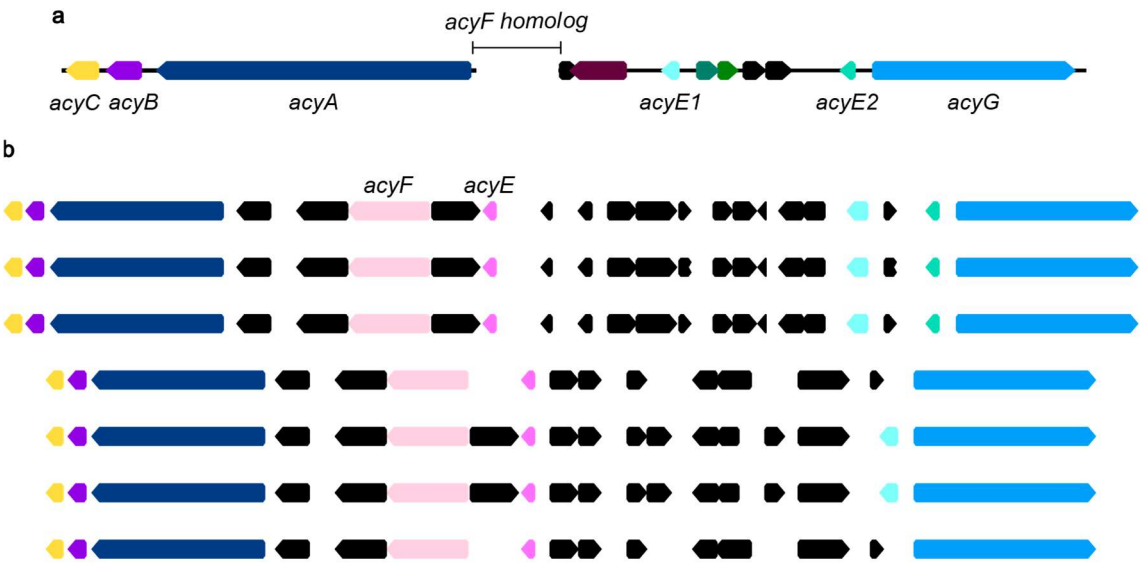


Figure 18 – a: Architecture of BGCs 74.1 and 66.1. **b:** First hits from the alignment. It is possible to verify that the *acyF* homolog is in between the subtilisin-like serine proteases. There is also another precursor peptide on the region that is missing from our cyanobactin BGC, indicating that we might have an additional prenylated precursor peptide.

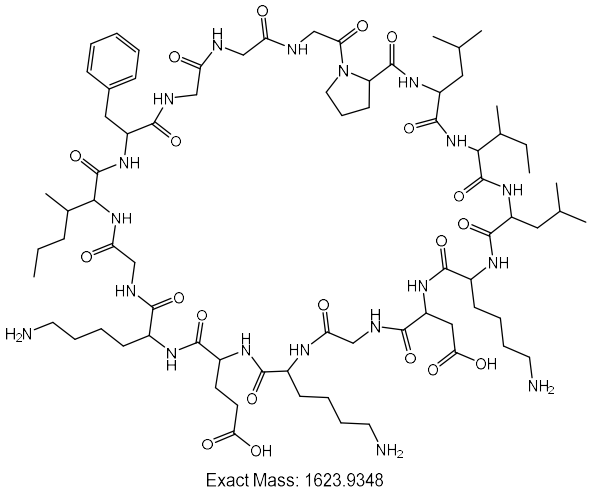


Figure 19 – Predicted structure for the core peptide encoded by *acyE1*.

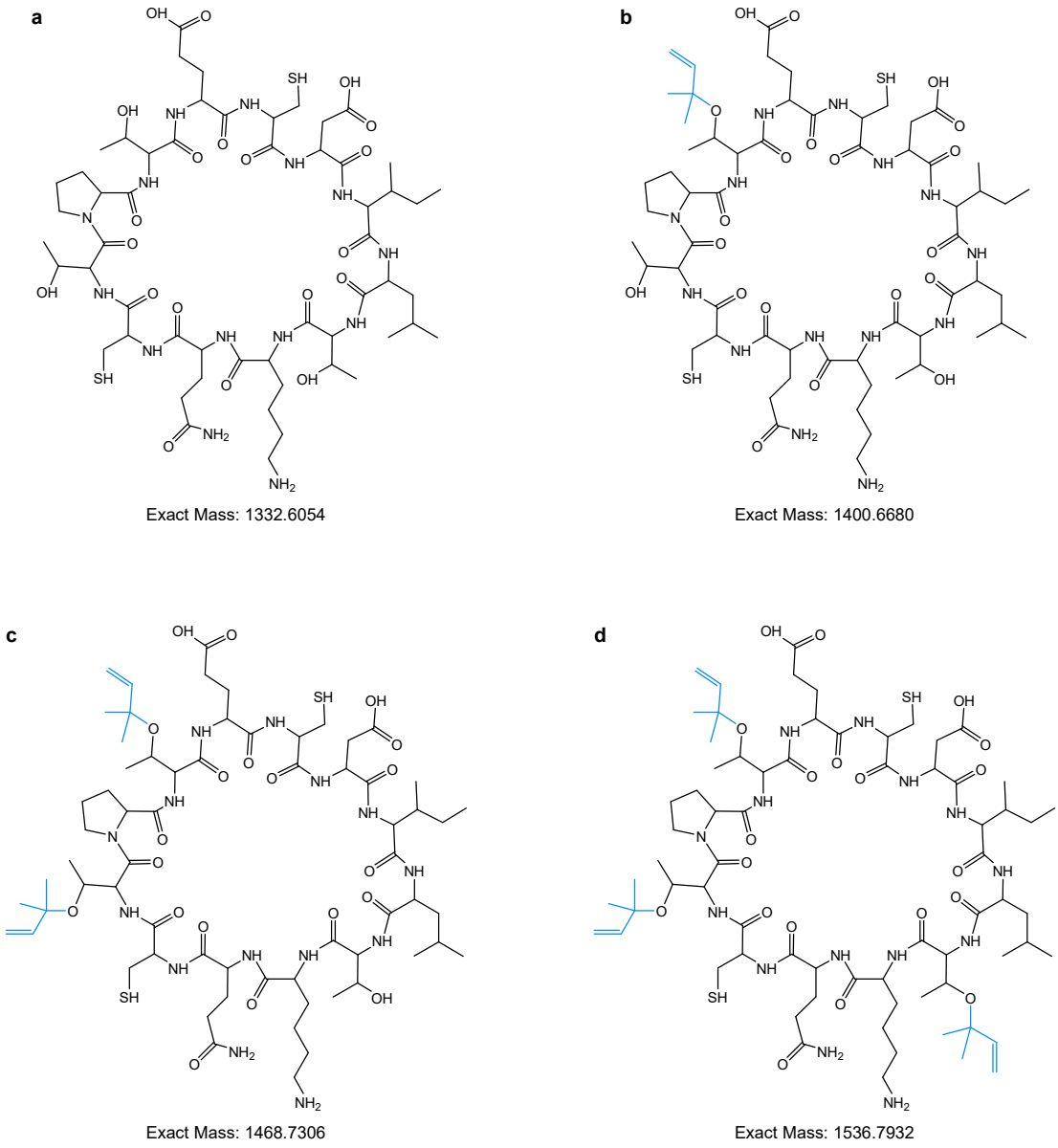


Figure 20 – Predicted structure for the core peptide encoded by *acyE2*. The prenyl groups are highlighted in **blue**. **a:** without prenyl groups; **b:** one prenyl group; **c:** two prenyl groups; **d:** three prenyl groups.

3.3.4. BGC 91.1, Bin 108 – aeruginosin-type

Bioinformatic analysis of BGC 91.1 (Bin 108) indicated similarities with aeruginosin-type gene clusters. Aeruginosins are nonribosomal peptides with a serine protease inhibitory activity. These family of linear peptides has several characteristic structural features, namely: the arginine derivatives at C-terminus, such as, argininol, argininal, agmatine, 1-amidino-2-ethoxy-3-aminopiperidine and rarely 1-amino-2-(N-amidino-Δ3-pyrrolinyl)-ethyl (Aeap) moiety; the 2-carboxy-6-hydroxyoctahydroindole (Choi) moiety; and, at the N-terminal position, it is common to have phenyl lactic acid derivatives or, in some cases, 2-O-methyl-3-sulfoglyceric acid (Mgs) or hexanoic acid.[136] The minimal aeruginosin BGC is composed of genes *aerA-B* and *aerD-G*. AerA is

responsible for the activation of a fatty acid or a monocarboxylic acid. Posteriorly, AerB adds a hydrophobic D-amino acid to the compound. The unusual Choi moiety is then synthesized by AerD-F and incorporated into the compound by AerG which also has the off-loading module. The variation in the loading and off-loading mechanisms of the biosynthetic enzymes as well as the diversity of tailoring enzymes present in these BGCs are responsible for the huge structural variability of aeruginosins. BGC 91.1 has the minimal genes responsible for the production of aeruginosins (Table 17). Additionally, it encodes a second SDR family oxidoreductase and the NRPS AerM (Table 17). Different from AerG, AerM has a thioester-reductase domain that releases the peptide as a aldehyde.[137] The reported cluster does not contain tailoring enzymes. According to the general gene disposition of aeruginosin BGCs, upstream of *aerA* there are no additional biosynthetic genes. Usually in this BGC-type, the tailoring enzymes are encoded in between or after the NRPS modules. The antiSMASH analysis of this cluster revealed that it is on contig edge, meaning that some genes might be absent from the retrieved sequence. Nonetheless, as the last NRPS module is intact, an attempt to heterologously express the compound encoded in BGC 91.1 will be performed in the future. The architecture of BGC 91.1 and the antiSMASH monomers predicted for each module are represented in Fig. 21. The heterologous expression strategy for BGC 91.1 will include genes *aerA-M*. AerA present a high homology towards the AerA identified in *Planktothrix agarghii* (Table 17). The homolog accession number from this enzyme leads to the paper that describes aeruginoside 126A and 126B.[35] The authors refer that in the biosynthesis of aeruginoside 126A and 126B, AerA catalyses the addiction of a phenyl lactic acid moiety. This was attributed to the similarity between the A domain of AerA and the A domain of McyG from the microcystin biosynthesis which is responsible for the activation of phenylpyruvate. In the biosynthesis of aeruginoside 126A and 126B the phenylpyruvate is further reduced to phenyl lactic acid by the KR domain.[35] BlastP analysis of the A and KR domains of the AerA from our BGC revealed a 98.34% and 97.79% similarity, respectively, with the A and KR domains of the AerA from the biosynthetic pathway of aeruginoside 126A and 126B. Furthermore, the domains of the homologs were conserved. As such, for the structure prediction of the aerguinosin encoded in BGC 91.1 was assumed that AerA would incorporate a phenyl lactic acid moiety into the N-terminus of the compound. The second SDR family oxidoreductase is predicted to have no function on the biosynthesis on aeruginosins, as in the biosynthetic pathway of aeruginosin NAL2 this enzyme is present but no function could be attributed to it.[137] AntiSMASH analysis was not able to identify the amino acid incorporated by AerM into the compound. Instead, the bioinformatic tool proposed four possible amino acids for this position – glutamate, glycine, aspartate or asparagine. According to all the

retrieved information, the chemical structures of the aeruginosin encoded in BGC 91.1 from Bin 108 were predicted (Fig. 22).

Table 17 – Gene annotation for BGC 91.1 from Bin 108.

Gene	Proposed function	Organism	Identity / %	Similarity / %	Size / aa
aerM	Amino acid adenylation domain-containing protein	*Microcystis aeruginosa* PMC 728.11	76.23	86	1512
orf1	SDR family oxidoreductase	*Planktothrix* sp. PCC 11201	85.95	91	258
aerG	Amino acid adenylation domain-containing protein	*Microcystis aeruginosa* K13-10	82.97	92	1092
aerF	SDR family oxidoreductase	*Microcystis viridis* Mv_BB_P_1995 1000_S69	88.35	96	266
aerE	Cupin domain-containing protein	*Microcystis aeruginosa* F13-15	75.96	89	208
aerD	Aeruginoside biosynthesis prephenate decarboxylase AerD	*Planktothrix agardhii*	87.75	94	210
aerB	Aeruginoside biosynthesis non-ribosomal peptide synthetase AerB	*Planktothrix agardhii* KL2	78.17	86	1593
aerA	NRPS/PKS hybrid enzyme, involved in aeruginosin biosynthesis aerA	*Planktothrix agardhii*	95.97	98	1428
orf1	Aldo/keto reductase	*Planktothrix agardhii*	92.51	96	374
orf2	Circadian clock protein KaiA	*Planktothrix agardhii*	96.88	97	321

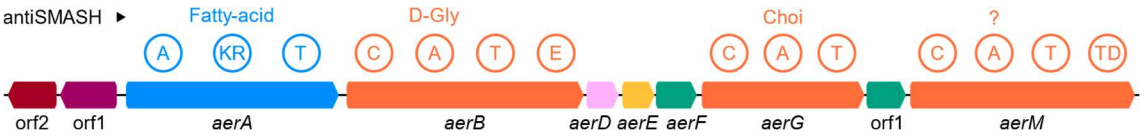


Figure 21 – Aeruginosin BGC 91.1 from Bin 108. In circles are represented the domains of the NRPS and PKS modules. Above the domains are represented the antiSMASH predicted monomers for each module. A: adenylation. KR: ketoreductase. T: thiolation. C: condensation. E: epimerization. TD: thioester-reductase.

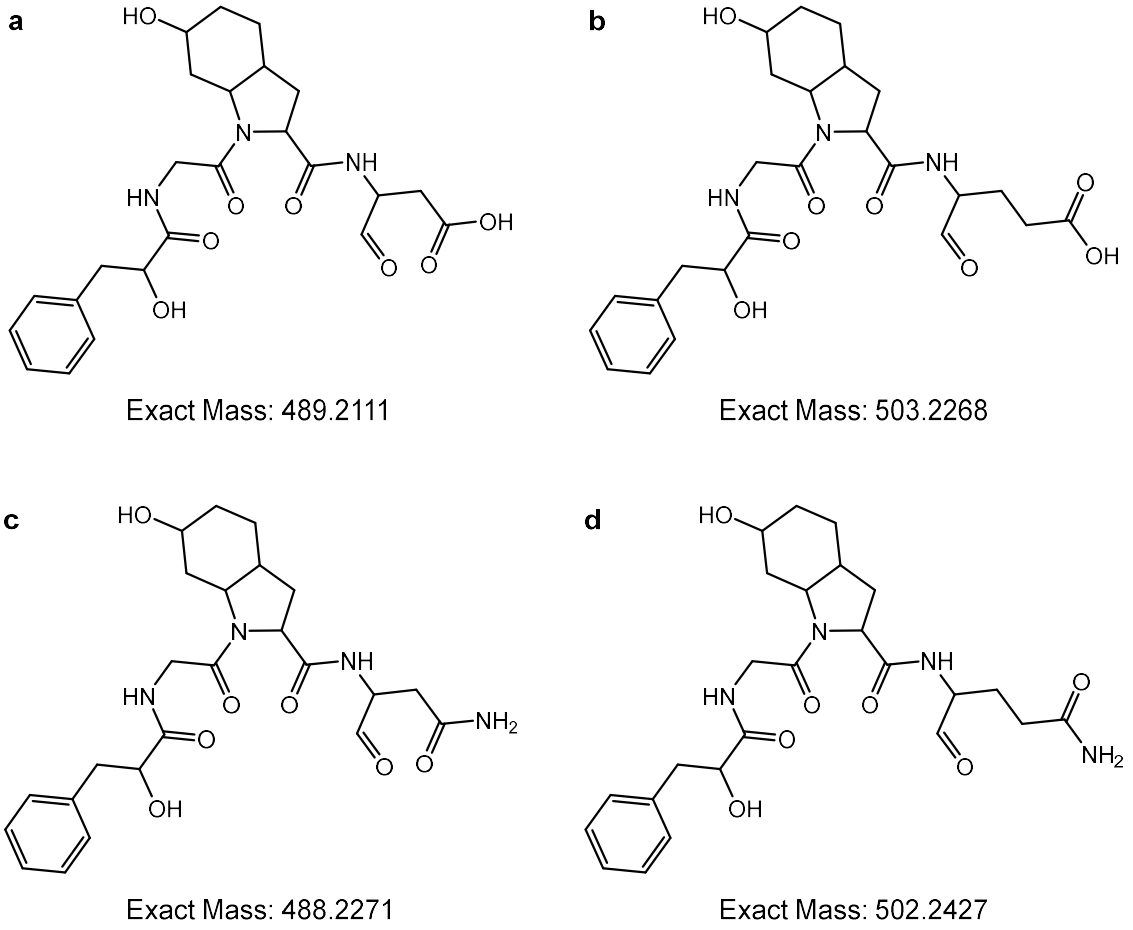


Figure 22 – Predicted structures for the aeruginosin encoded by BGC 91.1 from Bin 108. **a:** aspartate incorporated by AerM; **b:** glutamate incorporated by AerM; **c:** aparagine incorporated by AerM; **d:** glutamine incorporated by AerM.

3.3.5. BGC 523.1, Bin 112 – putative microginin

Bioinformatic analysis of BGC 523.1 from Bin 112 revealed it was a hybrid NRPS/T1PKS. Further analysis of the BGC structure and of the domains of each NRPS and PKS modules led us to hypothesize that this BGC encoded a microginin. Microginins are non-ribosomal linear peptides produced by cyanobacteria. These peptides are characterized by a N-terminal fatty-acid moiety followed by three to six amino acids.[138] The fatty acid moiety is composed by the 3-amino derivatives of decanoic (Ada) acid or octanoic acid (Aoa). Commonly, those derivatives present N-methylation, terminal halogenation or hydroxylation (Ahda/Ahoa).[138,139] According to the literature, for the

biosynthesis of microginins, a fatty acyl AMP-ligase (FAAL), an acyl carrier protein (ACP), a T1PKS/NRPS module and NRPS modules are essential. The FAAL (*micA*) activates and loads octanoic acid onto the ACP which then interacts with the Type I PKS module to generate the final fatty acid moiety.[140] Posteriorly, the NRPS modules will introduce the amino acid residues into the compound.[141]

BGC 523.1 has all the modules required for the biosynthesis of microginins namely, the FAAL, the ACP, the hybrid NRPS/T1PKS, the NRPS (Table 18 and Fig. 23). In the case of MicC, its closest homolog is a hypothetical protein, nonetheless, analysis of its domains demonstrated close similarity to NocO and NocN as well as ColD and ColE. All of these enzymes are dimetal-carboxylate halogenases from the CylC family. Typically, these enzymes catalyze the addition of one or two halogen atoms mid-chain or at the chain termini.[142] Previously, our group discovered twelve new microginins produced by a CylC-homolog harboring BGC from *Microcystis aeruginosa* LEGE 91341.[120] The halogenated microginins presented mono- or di-chlorination at the terminus of the fatty acid chain. The possible compounds encoded in BGC 523.1 were predicted based on BGC architecture and literature data (Fig. 24 and 25). The exact masses of the expected products were compared to the microginin structures present at CyanoMetDB[139] to verify if BGC 523.1 could encode new compounds. CyanoMetDB is a recently created database of cyanobacterial secondary metabolites comprising 2010 distinct structures by the time it was published.[139] Comparison results revealed several close masses (Table 19). Nevertheless, the building block structure of the previously described microginins is different from the predicted secondary metabolites encoded in this BGC (Table 19). Although, the structure of Nostoginin BN741 is very similar to the structure obtained for the compounds here described using 3-amino-2-hydroxy-octanoic acid as the fatty acid moiety without any chlorination, switching only the NMeLeu for NMeIle (Table 19). As BGC 523.1 contains the dimetal-carboxylate halogenase, most likely its compounds will be halogenated, thus adding chemical diversity to the microginin family. The strategy for heterologous expression of BGC 523.1 from Bin 112 will include all genes listed in Table 18. No further experimental work was developed with this BGC during the development of this thesis.

Table 18 – Gene annotation for BGC 523.1 from Bin 112.

Gene	Proposed function	Organism	Identity / %	Similarity / %	Size / aa
micA	Fatty acyl-AMP ligase	*Planktothrix*	99.83	100	588
micB	Acyl carrier protein	*Planktothrix*	100.00	100	88
micC	Hypothetical protein	*Planktothrix*	93.76	97	465
micD	Amino acid adenylation domain-containing protein	*Microcystis aeruginosa*	89.57	92	2704
micE	Non-ribosomal peptide synthetase	*Microcystis aeruginosa*	90.03	94	3069
micF	Amino acid adenylation domain-containing protein	*Microcystis aeruginosa*	86.79	91	1396
micG	ATP-binding cassette domain-containing protein	*Planktothrix*	98.50	99	668

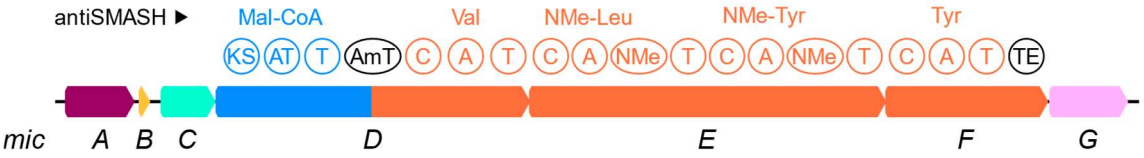


Figure 23 – Microginin BGC 523.1 from Bin 112. In circles are represented the domains of the NRPS and PKS modules. Above the domains are represented the antiSMASH predicted monomers for each module. KS: ketosynthase. AT: acyltransferase. T: thiolation. AmT: Transamination. C: condensation. A: adenylation. NMe: N-methylation. TE: thioesterase.

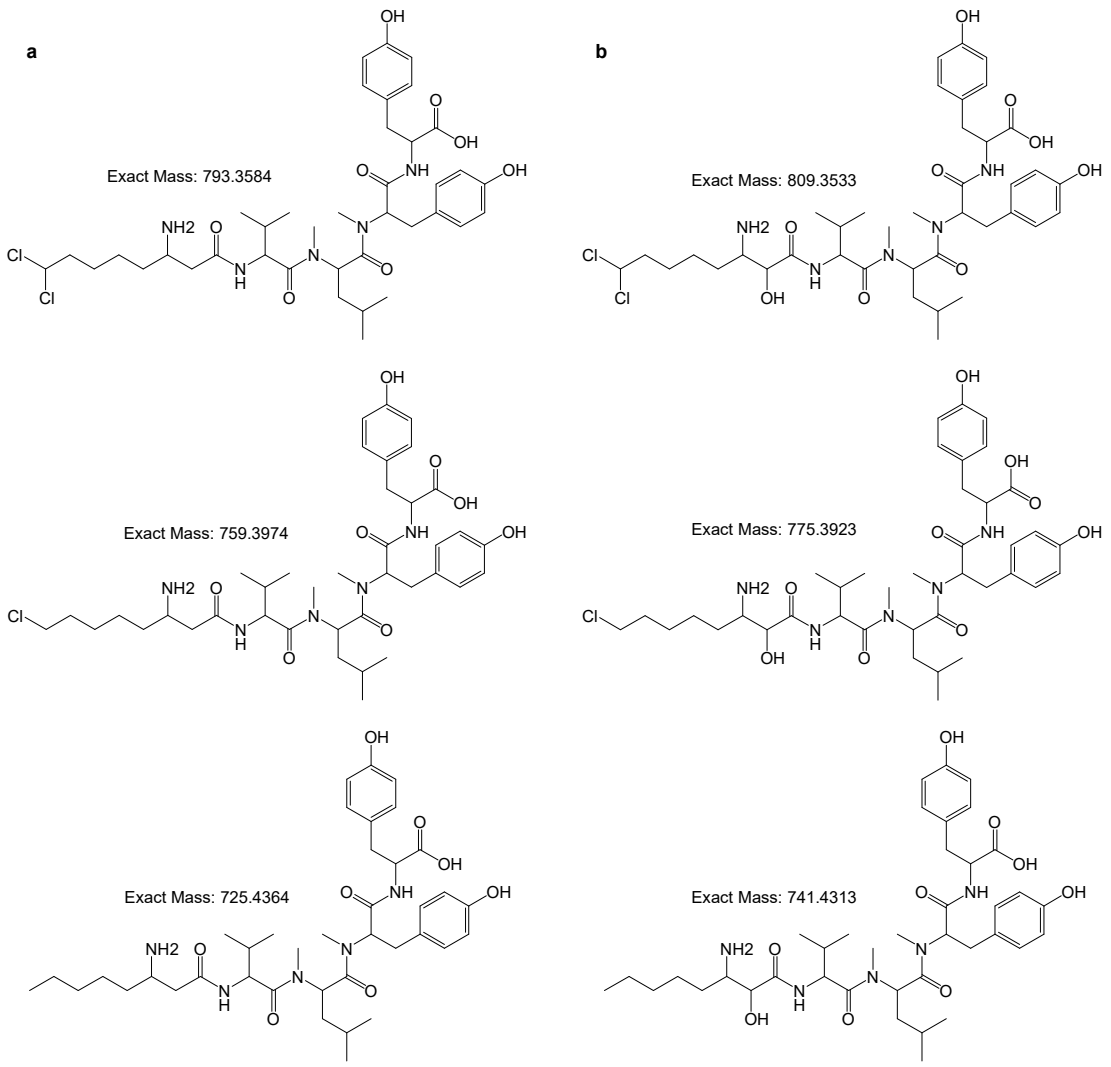


Figure 24 – Predicted structures for the microginins encoded by BGC using **a:** 3-amino-octanoic acid or **b:** 3-amino2-hydroxy-octanoic acid as fatty acid chain.

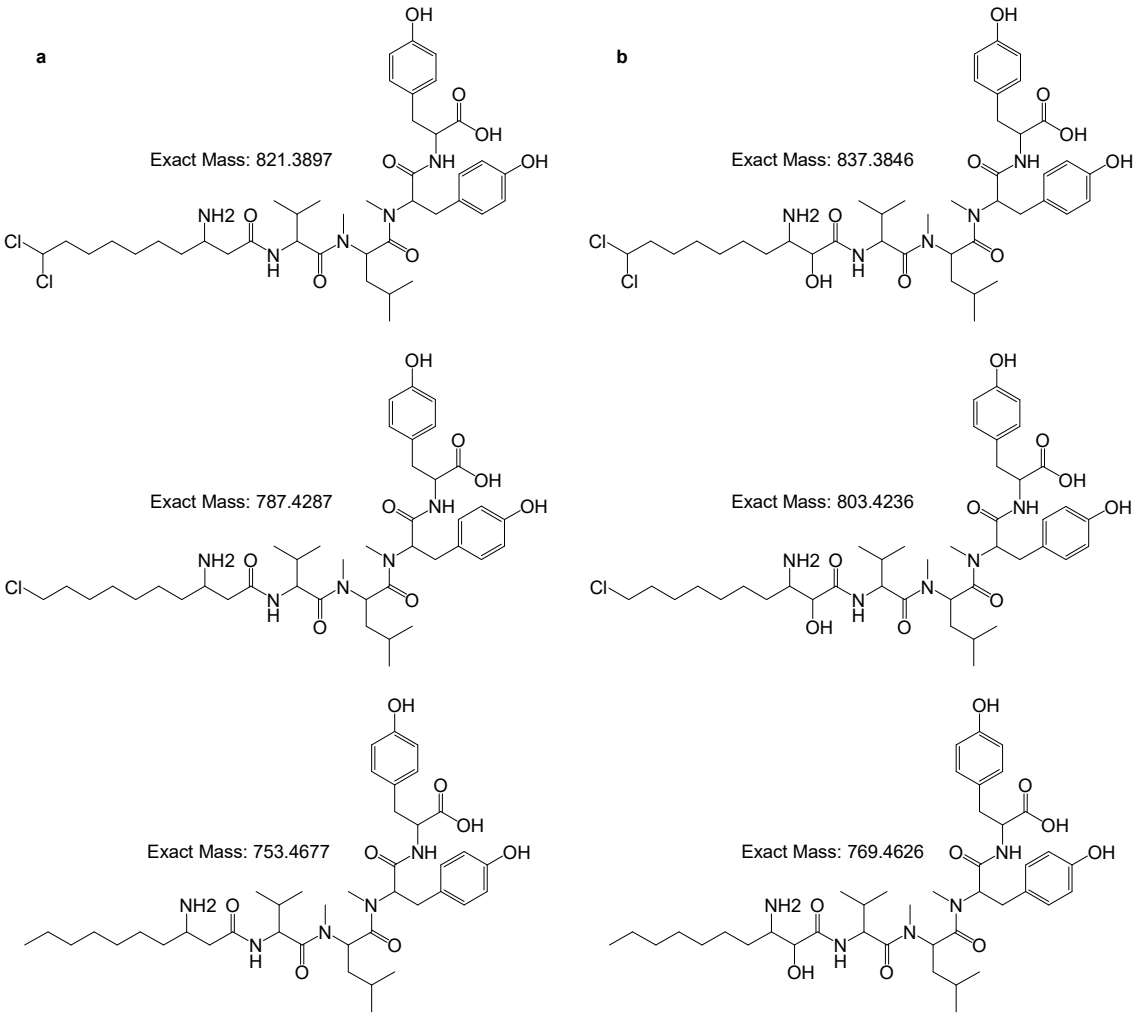


Figure 25 – Predicted structures for the microginins encoded by BGC using **a:** 3-amino-decanoic acid or **b:** 3-amino2-hydroxy-decanoic acid as fatty acid chain.

Table 19 – Description of the microginins present on the CyanoMetDB database with the closest monoisotopic mass to that predicted for our compounds.

Compound name	Monoisotopic mass	Building block string
Microginin FR5	725.39998	Ahda-Val-Pro-Tyr-Tyr
Microginin 725	725.41304	ClAhda-Ala-MeLeu/MeIle-MeLeu/MeIle-Tyr
Microginin 741B	741.39489	MeAhda-Ser-Pro-Tyr-MeTyr/Hty
Microginin FR4	741.39489	MeAhda-Thr-Pro-Tyr-Tyr
Microginin FR2	741.39490	MeAhda-Thr-Pro-Tyr-Tyr
Microginin 741A	741.43128	MeAhda-Ala-MeLeu/MeIle-Tyr-Tyr
Microginin 741C	741.43128	Ahda-Ala-MeLeu/MeIle-MeTyr/Hty-Tyr
Nostoginin BN741	741.43128	(2S,3S)Ahoa-Val-NMeIle-NMeTyr-Tyr

Table 19 – Continuation.

Compound name	Monoisotopic mass	Building block string
Cyanostatin B	753.43128	Ahda-Tyr-MeIle-Pro-Tyr
Microginin 478	769.46258	MeAhda-Val-MeVal-MeTyr-Tyr
Microginin 770	769.46258	MeAhda-Val-Ile-Tyr-Tyr
Microginin GH787	787.39231	ClAhda-Tyr-MeIle-Pro-Tyr
Microginin KR787	787.39231	ClAhda-Tyr-MeLeu-Pro-Tyr

3.4. Heterologous expression of BGC 52.1 from Bin 108 (version without orf2)

BGC 52.1 was assembled into a modified pET-28 through DiPaC-SLIC. The vector backbone includes a tetracycline inducible promoter (PtetO) and a *gfp* gene (Fig. 26). Three amplicons were generated through PCR for the cloning strategy (Fig. 26). The PCR conditions were optimized and the amplified genes were confirmed through direct sequencing. Both the amplicon and vector backbone in each step were purified through gel band excision. Colony PCR was used to verify if the SLIC reactions were successful. The colonies selected for colony PCR were left to grow in plates containing LB agar medium supplemented with 50 μg mL^{-1} of kanamycin (Fig. 27-29). A positive colony was chosen in each step and sent to sequencing to verify the integrity of the ligation zones (Fig. 30). Colonies with no mutations were selected to proceed with the DiPaC-SLIC strategy. In the case of the SLIC reaction between the vector backbone pET28b-ptetO::orf1-*mvdABC*-gfpv2 and genes *mvdDEF*, the colonies that developed with the overnight incubation period (colonies 1 to 3) did not have the complete BGC. Positive transformants (colonies 4 and 5) appeared later on the same day. The precursor peptides *mvdE* and *mvdF* are included in this set of genes. Therefore, the delayed growth might indicate that *E. coli* cells were redirecting cellular resources towards the production of the compounds encoded by the BGC.

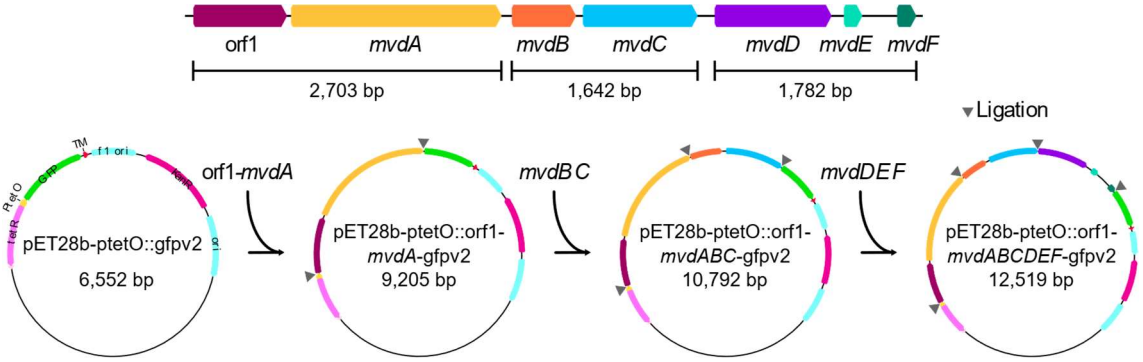


Figure 26 – Cloning Strategy for BGC 52.1. The gene cluster was divided in three fragments and cloned into the vector backbone pET28b-ptetO::gfpv2.

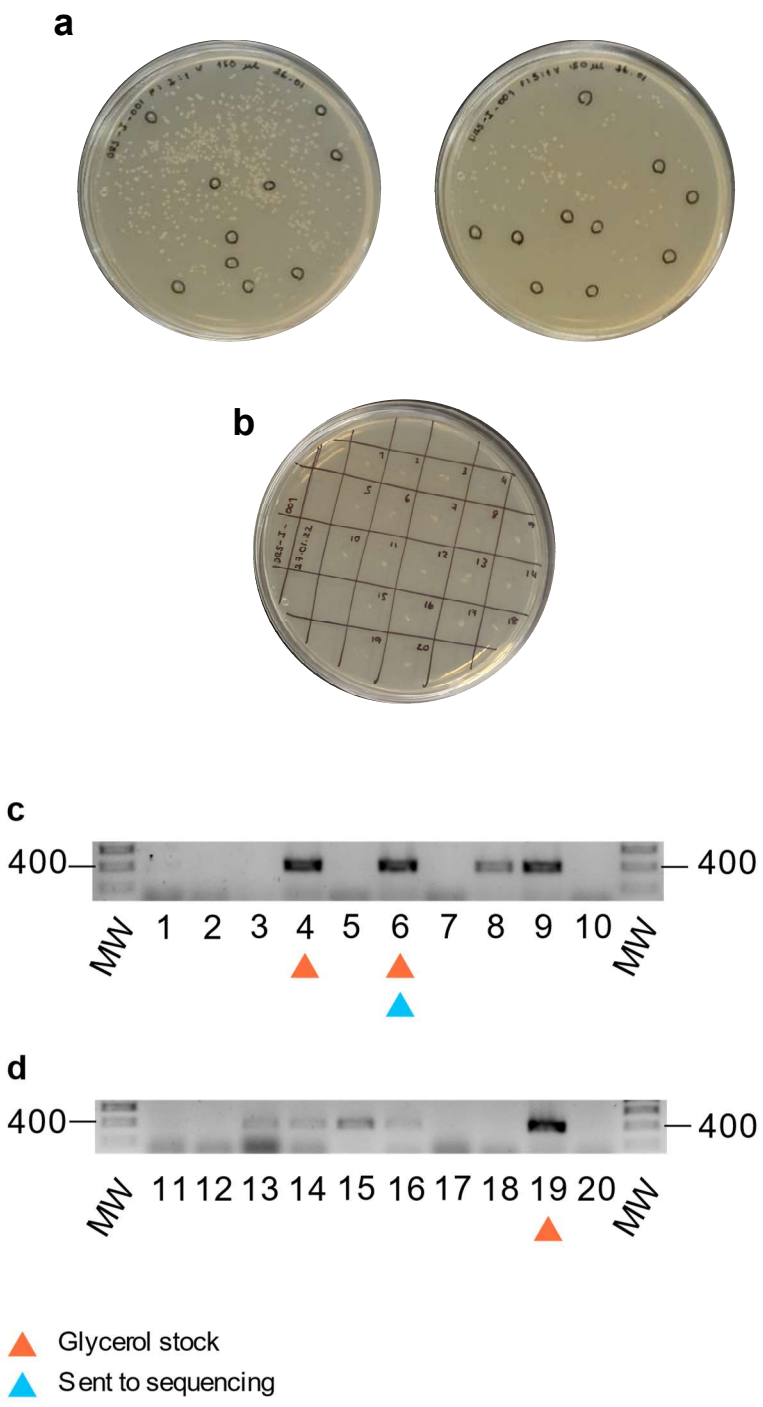


Figure 27 – Colony PCR screening for the SLIC reaction between pET28b-ptetO::gfpv2 and genes orf1-*mvdA*. **a:** Selected colonies to undergo colony PCR. On the right is the LB agar medium plate supplemented with 50 µg mL^{-1} of kanamycin for ratio pET28b-ptetO::gfpv2 1:2 orf1-*mvdA*. On the left is the LB agar medium plate supplemented with 50 µg mL^{-1} of kanamycin for ratio pET28b-ptetO::gfpv2 1:5 orf1-*mvdA*. **b:** LB agar medium plate supplemented with 50 µg mL^{-1} of kanamycin used to grow the colonies selected for colony PCR. **c:** Resulting electrophoresis gel from the colony PCR for ratio pET28b-ptetO::gfpv2 1:2 orf1-*mvdA* using primers Screen_ptetF2 and 52.1_colony_R1 (expected size – 413 bps). **d:** Resulting electrophoresis gel from the colony PCR for ratio pET28b-ptetO::gfpv2 1:5 orf1-*mvdA* using primers Screen_ptetF2 and 52.1_colony_R1 (expected size – 413 bps). MW: NZYDNA Ladder III (NZYTech).

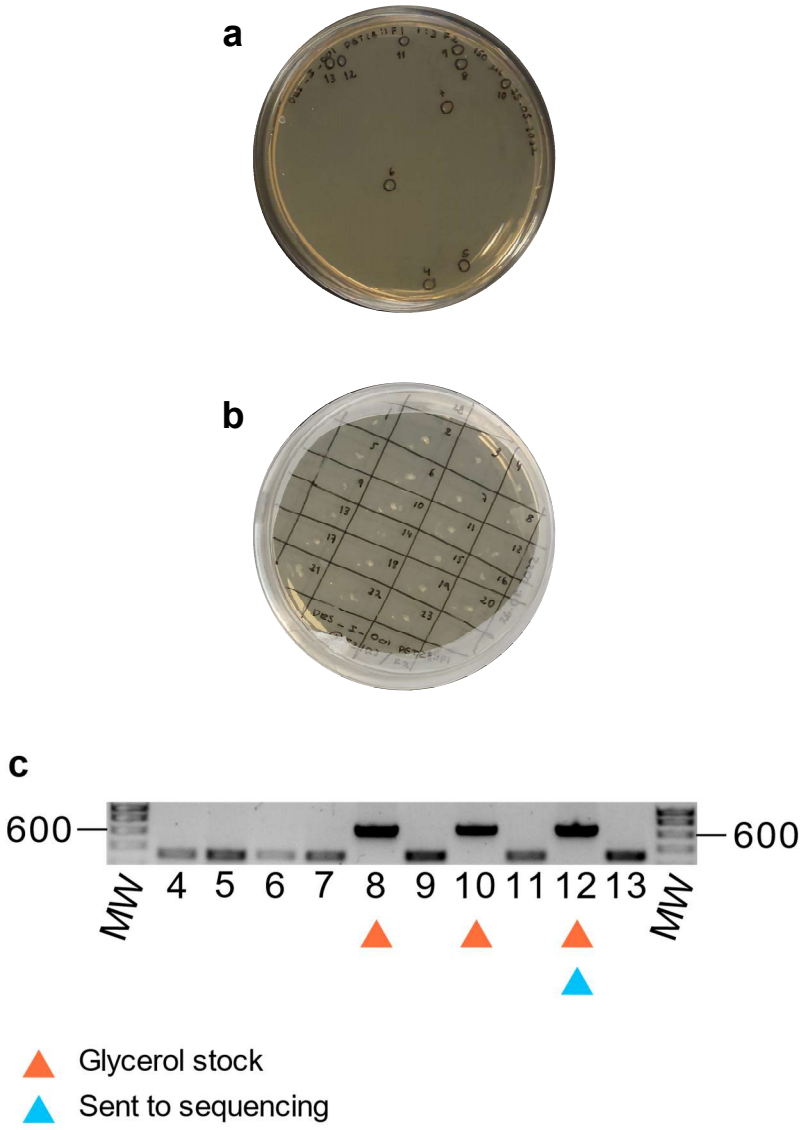


Figure 28 – Colony PCR screening for the SLIC reaction between pET28b-ptetO::orf1-*mvdA*-gfpv2 and genes *mvdBC*. **a:** Selected colonies to undergo colony PCR for the ratio pET28b-ptetO::orf1-*mvdA*-gfpv2 1:3 *mvdBC*. **b:** LB agar medium plate supplemented with 50 μg mL^{-1} of kanamycin used to grow the colonies selected for colony PCR. **c:** Resulting electrophoresis gel from the colony PCR using primers 52.1_colony_F2 and 52.1_colony_R2 (expected size – 670 bps). MW: NZYDNA Ladder III (NZYTech).

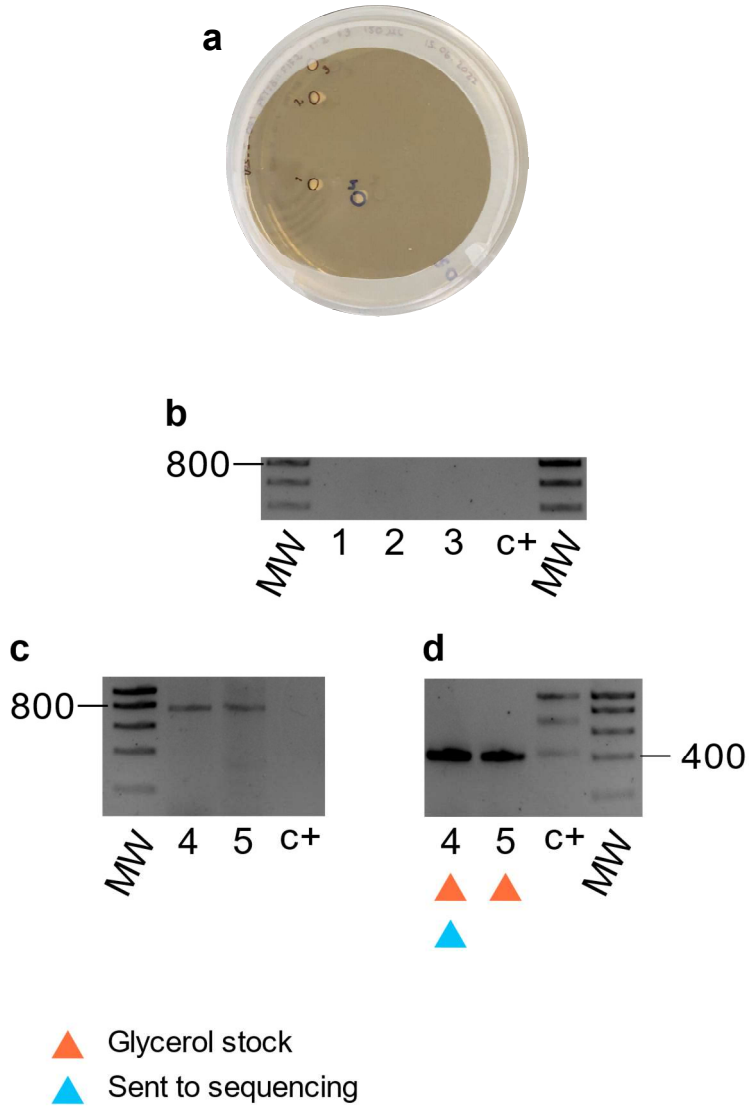


Figure 29 – Colony PCR screening for the SLIC reaction between pET28b-ptetO::orf1-*mvdABC*-gfpv2 and genes *mvdDEF*. **a:** Selected colonies to undergo colony PCR for the ratio pET28b-ptetO::orf1-*mvdABC*-gfpv2 1:2 *mvdDEF*. **b:** Resulting electrophoresis gel from the colony PCR of colonies 1 to 3 using primers 52.1_colony_F3 and 52.1_colony_R3 (expected size – 765 bps). **c:** Resulting electrophoresis gel from the colony PCR of colonies 4 and 5 using primers 52.1_colony_F3 and 52.1_colony_R3 (expected size – 765 bps). **d:** Resulting electrophoresis gel from the colony PCR of colonies 4 and 5 using primers 52.1_colony_F4 and screen_GFP_R (expected size – 413 bps). MW: NZYDNA Ladder III (NZYTech). c+: positive control using the vector backbone.

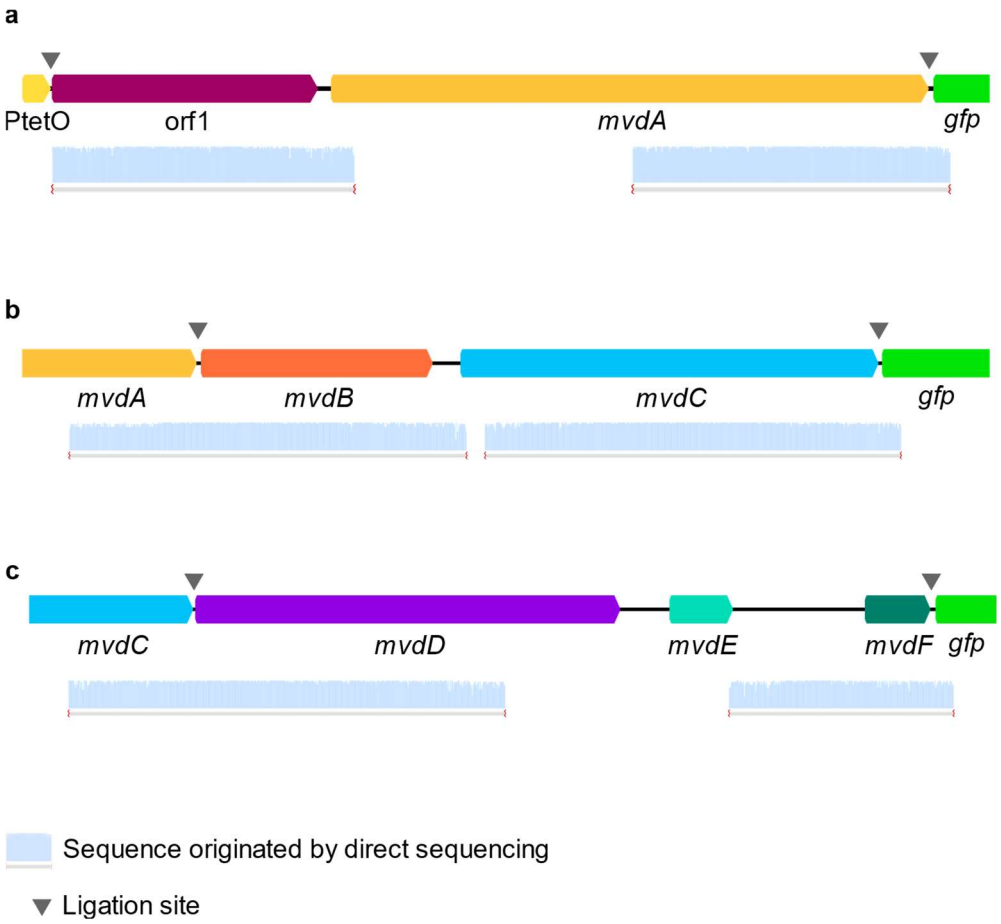


Figure 30 – Direct sequencing results. **a**: Sequencing results at the ligation zones between vector backbone pET28b-ptetO::gfpv2 and genes orf1-*mvdA* from BGC 52.1. **b**: Sequencing results at the ligation zones between vector backbone pET28b-ptetO::orf1-*mvdA*-gfpv2 and genes *mvdBC* from BGC 52.1. **c**: Sequencing results at the ligation zones between vector backbone pET28b-ptetO:: orf1-*mvdABC*-gfpv2 and genes *mvdDEF* from BGC 52.1.

The final 12,519 bp construct was cloned into *E. coli* BL21 (DE3). Two transformants – colony 4a (large) and colony 4b (small) – were selected for heterologous expression (Fig. 31). The correct transformation with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 was confirmed through direct sequencing (Fig. 31). *E. coli* BL21 (DE3) cells previously transformed with pET28b-ptetO::gfpv2 (empty vector) were used as control. The presence of the vector pET28b-ptetO::gfpv2 was confirmed through direct sequencing (Fig. 31). Cultures were incubated at 20 ºC and 37 ºC for 3 days. Expression was tested with and without tetracycline supplementation. Fig. 32 shows the aspect of the cultures after the three-day incubation.

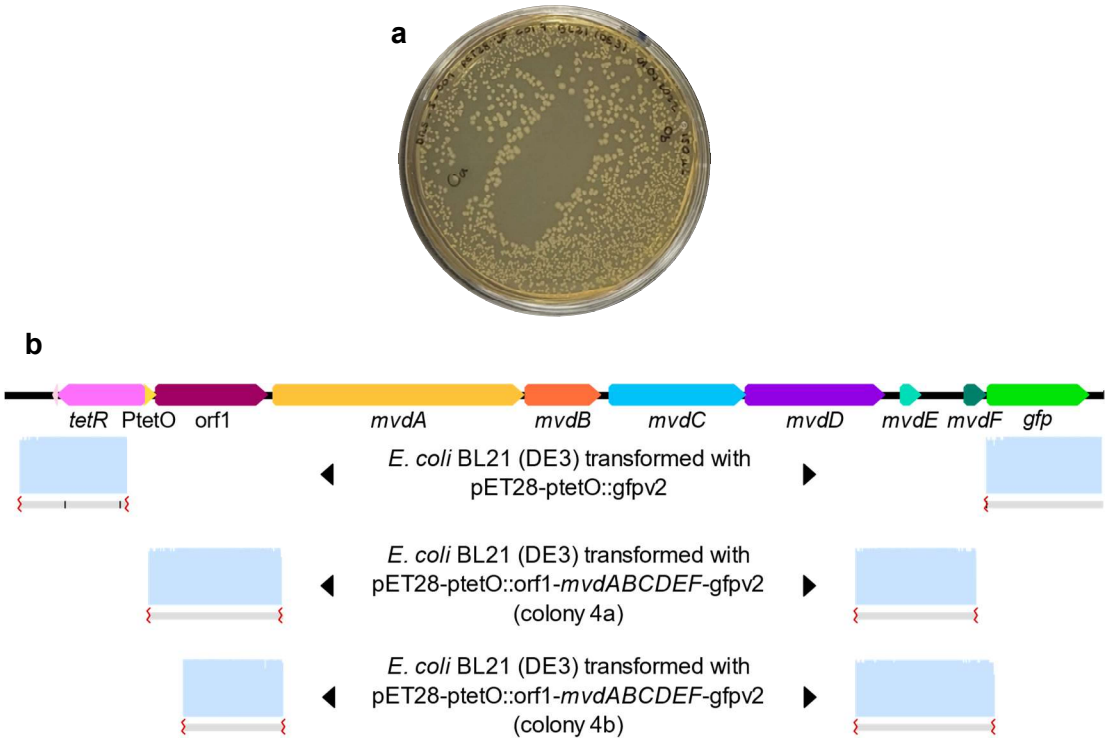


Figure 31 – Cloning of vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 into *E.coli* BL21 (DE3). **a:** Resulting plate from the chemical transformation of *E. coli* BL21 (DE3) with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2. **b:** Direct sequencing results.

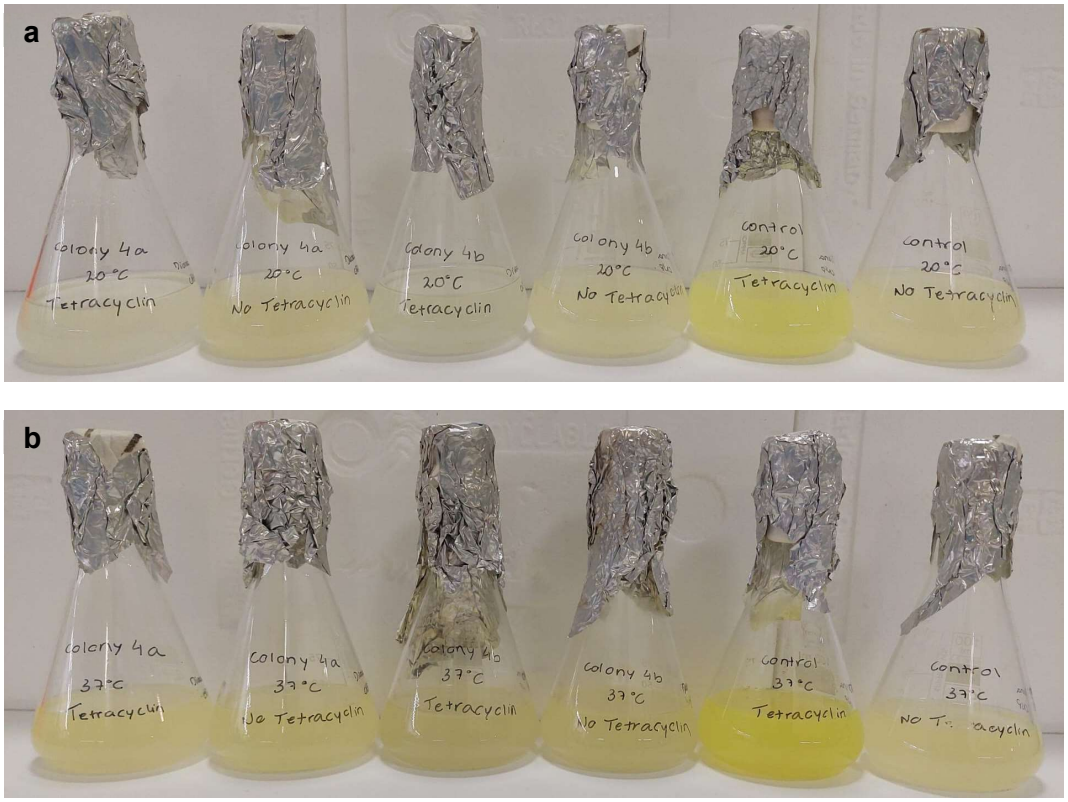


Figure 32 – Cultures after a period of three day incubation. **a:** Cultures incubated at 20 ºC. **b:** Cultures incubated at 37 ºC.

Pellets and supernatants from 25 mL of culture medium were harvested at day 3. LC-HRESIMS analysis of the resulting pellet extracts revealed a peak at approximately 6.12 min that was present only in *E. coli* transformed with the microviridin BGC thus confirming the expression of pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 (Fig. 33 a). This peak was more abundant when the cells were cultured at 20 ºC with tetracycline supplementation. Nonetheless, we could detect this peak when the cells were cultured at 20 ºC without tetracycline supplementation and 37 ºC with tetracycline supplementation (Fig. 33 a). Analysis of the peak identified several *m/z* for *E. coli* carrying the vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 that were not present in *E. coli* transformed with the empty vector (Table 20). Moreover, at approximately min 7 several other masses exclusively present in *E. coli* holding the *mvd*BGC were identified (Table 20). Although the encountered ions were more abundant in colonies supplemented with tetracycline, several were also present on the extracts from colonies without tetracycline induction demonstrating leaky expression of the BGC (Fig. 33 b). On the LC-HRESIMS spectra of the supernatant extracts we were not able to encounter differences between *E. coli* transformed with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 and *E. coli* transformed with the empty vector. The spectra are complex possibly due to the composition of the medium and *E. coli* produced metabolites (Fig. 33 c). However, most of the *m/z* encountered on the pellet extracts, were also present on the supernatant extracts. Additionally, several *m/z* were also detected in supernatant extracts from cultures without tetracycline induction further proving the existence of leaky expression of the gene cluster.

The most abundant ions are represented in Fig. 34. These *m/z* were mainly present on the pellet extracts for the conditions tested with tetracycline induction at 20 ºC and on the supernatant extracts from cultures grown at 37 ºC with tetracycline induction (Fig. 34). In general, the relative abundance of these *m/z* is very similar on the pellet extracts for the conditions tested with tetracycline induction at 20 ºC and on the supernatant extracts from cultures grown at 37 ºC with tetracycline induction. The major difference was between *m/z* 1005.9595 and *m/z* 1278.5578. The ions were abundant on the pellet for the conditions tested with tetracycline induction at 20 ºC but were almost not present on the extracts from cultures grown at 37 ºC. *m/z* 1005.9595 is one of the ions that we could match with a mass for a predicted peptide (Table 20 and Supplementary Tables 1 to 8). As such, for our attempts to isolate the compounds for structural elucidation, the culture conditions used will be 20 ºC with tetracycline induction.

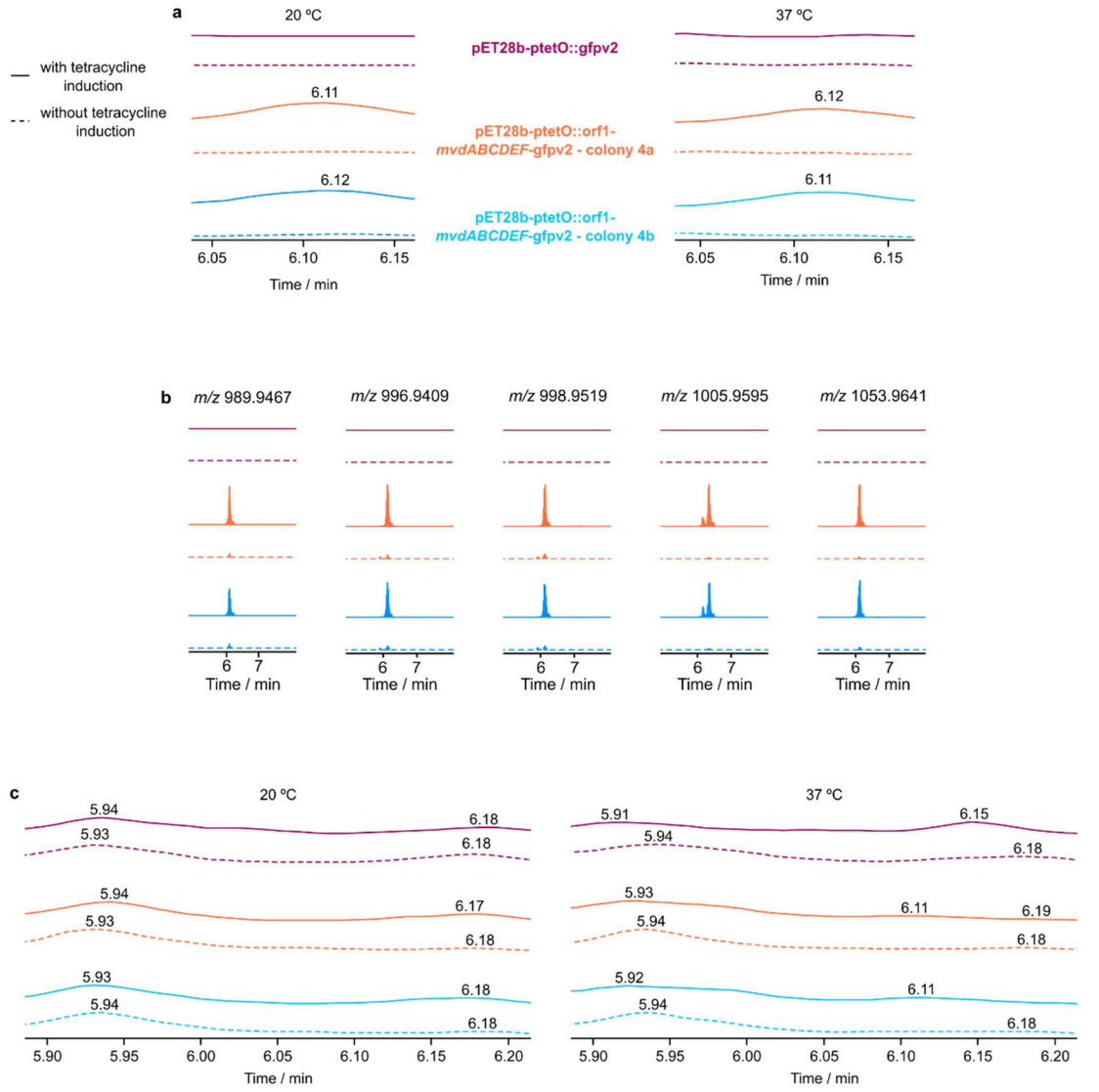


Figure 33 – Results from de MS analysis of *E.coli* extracts from 3-day old cultures. **a:** TIC spectra of the pellet extracts. At approximately min 6.12 there is a peak only visible for colonies carrying the vector ptetO::orf1-*mvdABCDEF*-gfpv2. **b:** Evidence of leaky BGC expression in cultures incubated at 20 ºC. **c:** TIC spectra of the supernatant extracts. The peak at approximately min 6.12 is not visible due to the complexity of the spectra.

Table 20 – *m/z* identified in *E. coli* carrying the *mvd*BGC and possible correspondent peptide.

m/z	Ion type	Calculated mass	Putative peptide
910.9120	M+2H	1819.8094	
917.9061	M+2H	1833.7977	PYPTTLKYPSDWEDY
919.9169	M+2H	1837.8192	
926.9245	M+2H	1851.8345	PYPTTLKYPSDWEDY
989.9467	M+2H	1977.8789	
996.9409	M+2H	1991.8672	SAPYPTTLKYPSDWEDY
998.9519	M+2H	1995.8892	
1005.9595	M+2H	2009.9044	SAPYPTTLKYPSDWEDY
1053.9641	M+2H	2105.9136	NSAPYPTTLKYPSDWEDY
1062.9806	M+2H	2123.9466	NSAPYPTTLKYPSDWEDY
1078.9712	M+2H	2155.9278	NSAPYPTTLKYPSDWEDY

Table 20 – Continuation.

m/z	Ion type	Calculated mass	Putative peptide
1097.4801	M+2H	2192.9456	SNSAPYPTTLKYPSDWEDY
1211.5227	M+2H	2421.0308	NNSNSAPYPTTLKYPSDWEDY
1220.5371	M+2H	2439.0596	NNSNSAPYPTTLKYPSDWEDY
1278.5578	M+H	1277.5505	
1161.8438	M+3H	3482.5096	SEQDTETGDSTSTDIPTIWTFKWPSDWEDS
1742.2635	M+2H	3482.5124	SEQDTETGDSTSTDIPTIWTFKWPSDWEDS

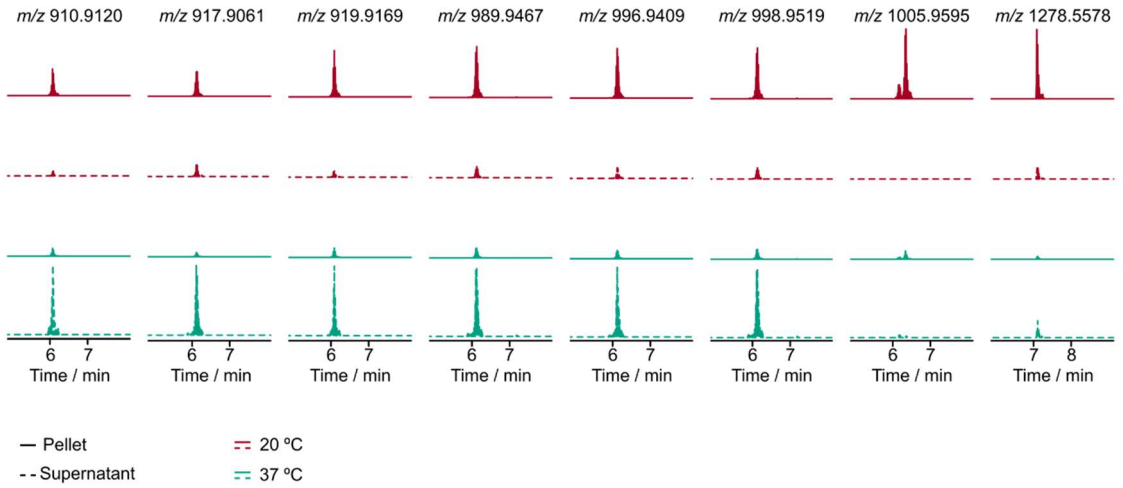


Figure 34 – Relative abundance of the most abundant ions in extracts from *E. coli* carrying the vector ptetO::orf1-*mvdABCDEF*-gfpv2 (colony 4a) cultured at 20 and 37 ºC with tetracycline supplementation.

The calculated mass for some of the most abundant *m/z* did not match any of the predicted peptide masses. Nonetheless, several of the encountered ions corresponded to a predicted peptide (Table 20 and Supplementary Tables 1 to 8). However, none of the encountered ions corresponded to the mass of the predicted final compound (with or without acetylation). Ziemert et. al previously succeeded in expressing microviridin B in *E. coli* demonstrating the suitability of this host to produce the final compound encoded in microviridin BGCs.[143] Having this in account, our results might indicate that an additional enzyme is participating on the biosynthesis of the peptides corresponding to the most abundant ions, possibly orf1, or that a catalyst is missing on the cloned BGC, possibly orf2. Nevertheless, it is needed to consider the possibility of *E. coli* proteases cleaving the peptide differently from the BGC-endoded cyanobacterial proteases. To analyze the hypothesis of a possible role of orf2 on the biosynthesis of this compounds, during the development of this thesis, orf2 was cloned into the vector that already contained orf1 and all the *mvd* genes. In the future, to better understand which modifications are being introduced by *E. coli* into the peptides, Tandem MS (MS/MS) analysis will be performed for the most abundant ions.

3.5. Heterologous expression of BGC 52.1 from Bin 108 (version with orf2)

orf2 from BGC 52.1 was assembled into the previously cloned vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 (Fig. 35). The amplicon of orf2 was generated through PCR for the cloning strategy. The PCR conditions were optimized and the amplified gene was confirmed through direct sequencing. Both the amplicon and vector backbone were purified through gel band excision. Colony PCR was used to verify if the SLIC reactions were successful (Fig. 36 a and b). The positive colony was sent to sequencing to verify the integrity of the ligation zones (Fig. 36 c).

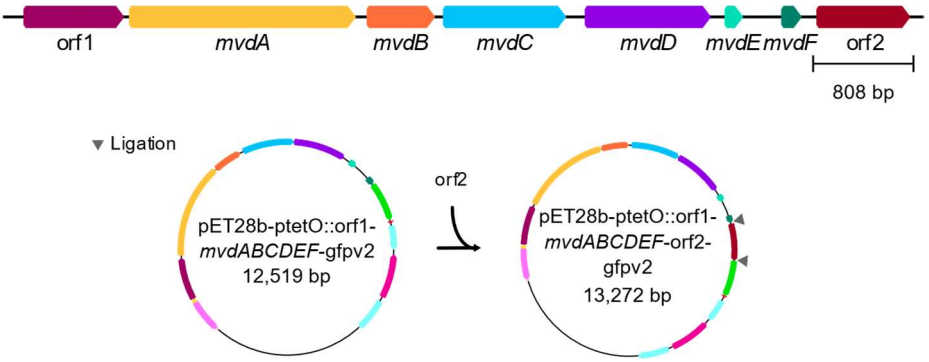


Figure 35 – Cloning strategy for orf2 from BGC 52.1. Gene orf2 was cloned into vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2.

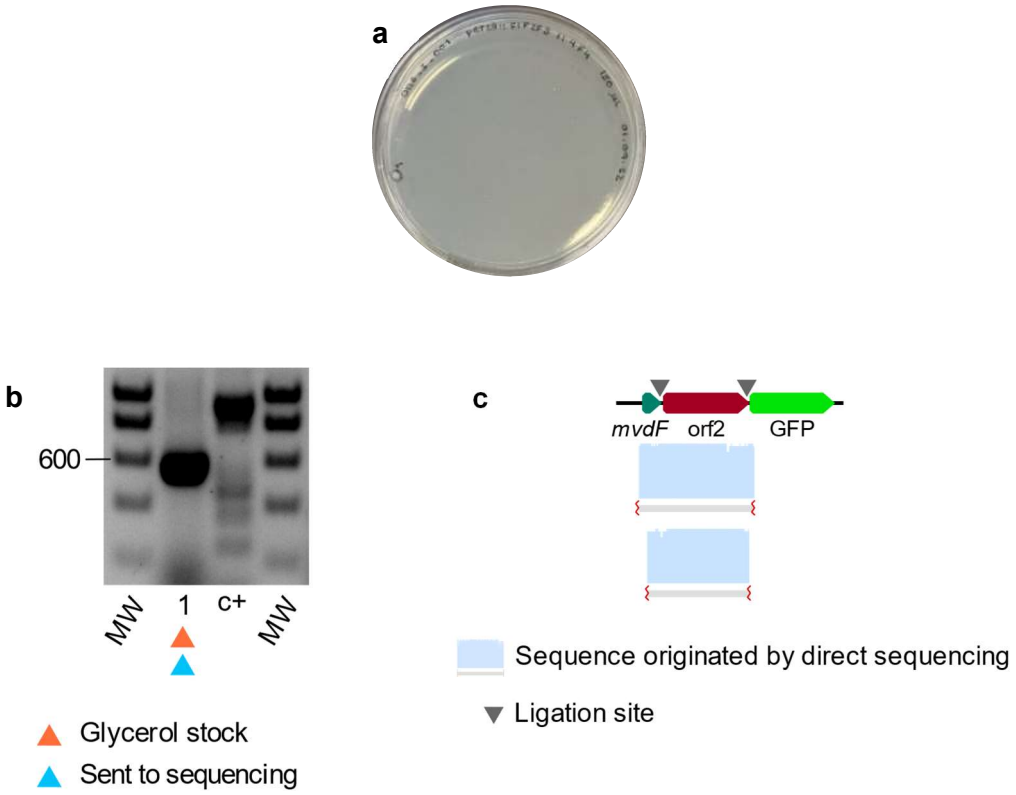


Figure 36 – Colony PCR screening for the SLIC reaction between pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 and gene orf2. **a:** Selected colony to undergo colony PCR for the ratio pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 1:4 orf2. **b:** Resulting electrophoresis gel from the colony PCR of colony 1 using primers 52.1_colony_F4 and 52.1_colony_R4 (expected size – 559 bps). **c:** Sequencing results at the ligation zones between vector backbone pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 and gene orf2 from BGC 52.1. MW: NZYDNA Ladder III (NZYTech). c+: positive control using the vector backbone.

The final 13,272 bp construct was cloned into *E. coli* BL21 (DE3). One transformant – colony 1a – was selected for heterologous expression (Fig. 37). The correct transformation with vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2 was confirmed through direct sequencing (Fig. 37). *E. coli* BL21 (DE3) cells previously transformed with pET28b-ptetO::gfpv2 (empty vector) were used as control. Cultures were incubated at 20 ºC for 3 days. Expression was induced by tetracycline supplementation. The peak at approximately min 6.12 verified in colonies transformed with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 was not present in extracts from *E. coli* carrying the vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2 (Fig. 38 a). Nonetheless, the previously detected *m/z* were all present in these cultures. The main peaks detected so far for *E. coli* transformed with the vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2 match those previously described (Fig. 38 b). However, these peaks are more abundant in extracts from *E.coli* transformed with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2. Therefore, at this point, there are no evidence of a possible role of orf2 on the biosynthesis of the compounds encoded in BGC 52.1. Given the difference on peak intensity, in the future, we will experiment other culture conditions.

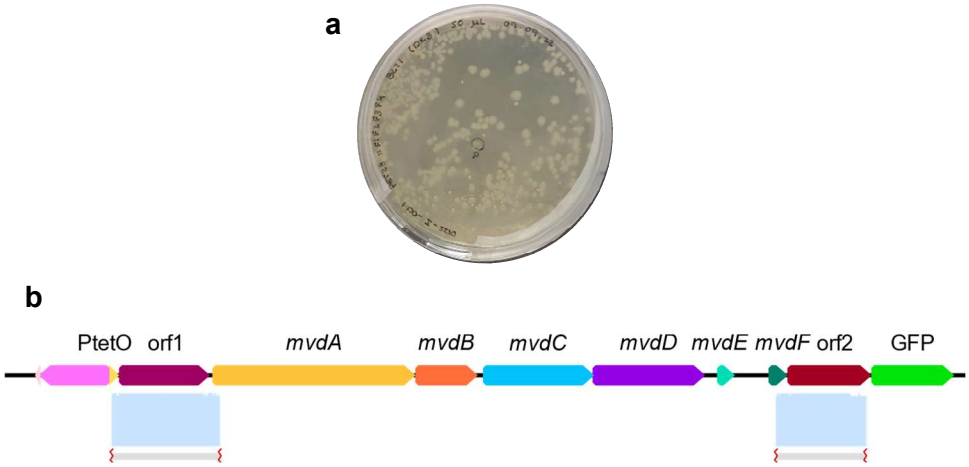


Figure 37 – Cloning of vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2 into *E.coli* BL21 (DE3). **a:** Resulting plate from the chemical transformation of *E. coli* BL21 (DE3) with vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2. **b:** Direct sequencing results.

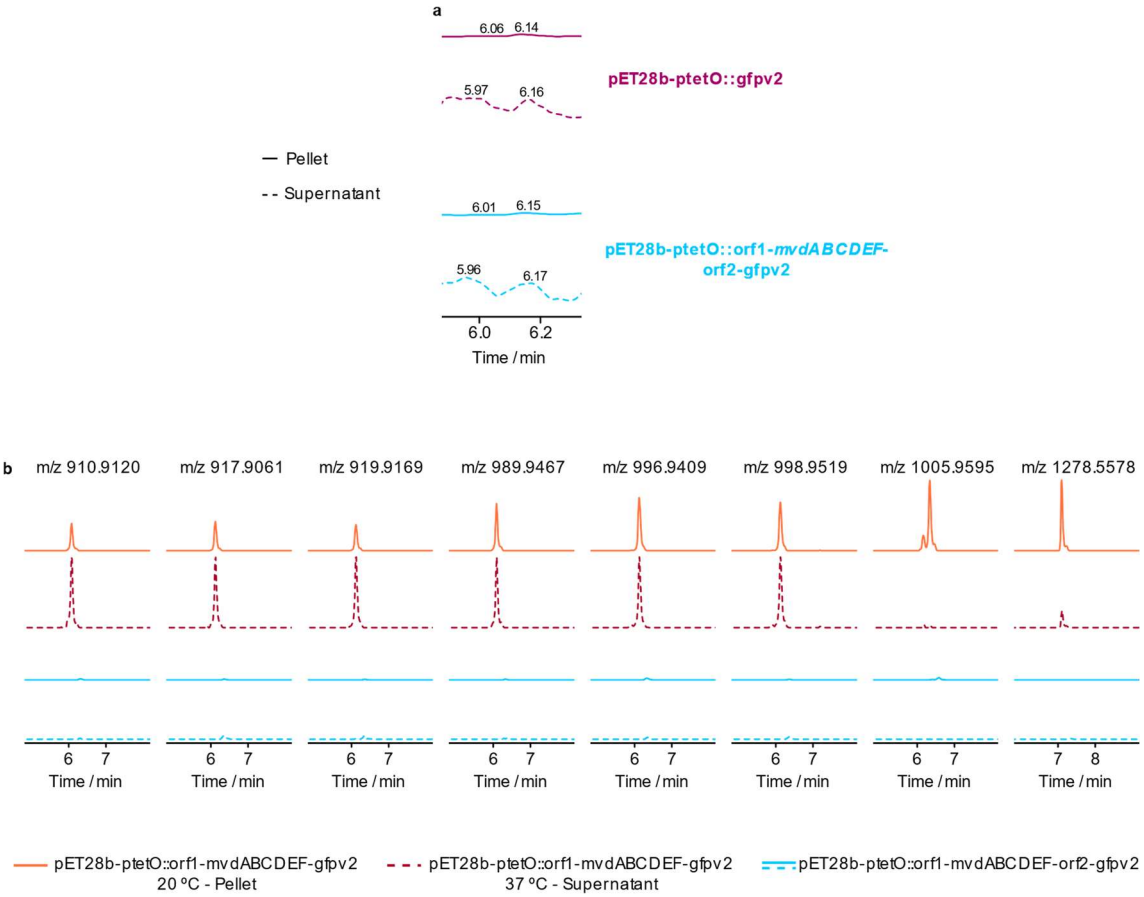


Figure 38 – Results from de MS analysis of extracts from *E.coli* carrying the vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2. **a:** TIC spectra of the pellet and supernatant extracts. The peak at approximately min 6.12 is not visible. **b:** Relative abundance of the most abundant ions of extracts from *E.coli* carrying the vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2 in comparison to the extracts with highest abundance of these *m/z* from *E.coli* carrying the vector pET28b-ptetO::orf1-*mvdABCDEF*- gfpv2.

3.6. Large scale culture of *E. coli* transformed with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2

10 L of *E. coli* holding the vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 were cultured at 20 ºC during 3 days. Expression was induced by tetracycline supplementation. At day 3 both pellet and supernatant were harvested and an organic extraction using MeOH was performed. The resulting extracts are represented in Fig. 39. Initially, due to the complexity of the supernatant extracts, only pellet extracts were going to be submitted to the SPE. However, as the recovered mass of the pellet extract was inferior to 1 g (Table 21), pellet and supernatant extracts were combined and the resulting extract was submitted to the SPE. The SPE system is demonstrated on Fig. 40. Five elutions were performed (Table 22). The cleaning fraction of the SPE column was also recovered. The resulting fractions are represented in Fig. 41. The mass recovered for each fraction is listed in Table 22. The SPE caused a loss of 375.74 mg. Given the volume of culture that was harvested and the fact that the pellet extract was not filtered, this mass could represent *E. coli* cells that were not eluted through the SPE column.

Figure 39 – Resulting extracts from the organic extraction of both supernatant (left) and pellet (right) from the large scale cultures.

Table 21 – Recovered masses from the pellet and supernatant extracts.

Extract	Weigth / mg
Supernatant Large Scale	532.74
Pellet Large Scale	670.17
Total	1202.91

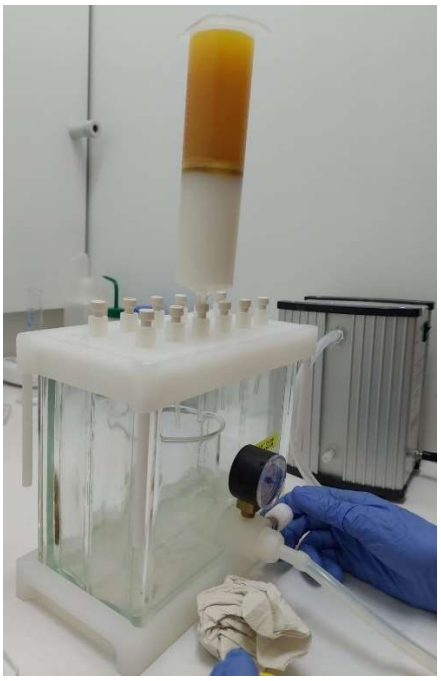


Figure 40 – SPE system. The SPE column is connected to a vacuum system to promote elution. The orange liquid is the extract dissolved on the solution from the first elution.

Table 22 – Solutions used on the SPE and recovered masses in each fraction.

Fraction	H_2O / %	MeOH / %	Number of passages	Weigth / mg
A	95	5	1	286.28
B	75	25	1	242.33
C	50	50	1	171.75
D	25	75	1	52.94
E	0	100	1	29.35
Cleaning	0	100	3	44.52
			Total	827.17

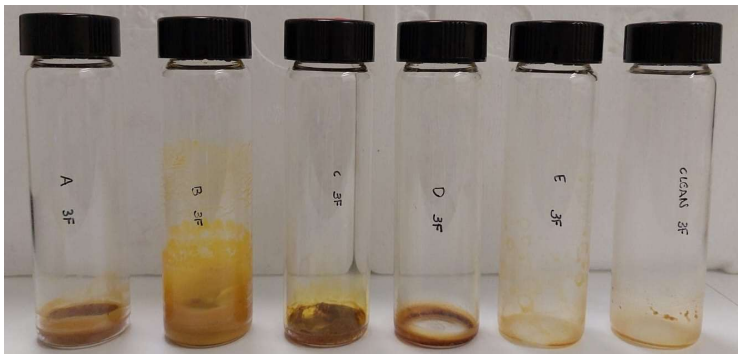


Figure 41 – Fractions resultant from the SPE.

The recovered fractions were analyzed through LC-HRESIMS. The peak at approximately min 6.12 verified in colonies transformed with vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2 was not distinguishable in any fraction (Fig. 42 a). Search for the most abundant ions revealed that all of them were more abundant in fraction C (Fig. 42 b). As such, we will continue the compound purification using fraction C. Nonetheless, a new extraction from large scale cultures might be required as this fraction mass is lower than 200 mg.

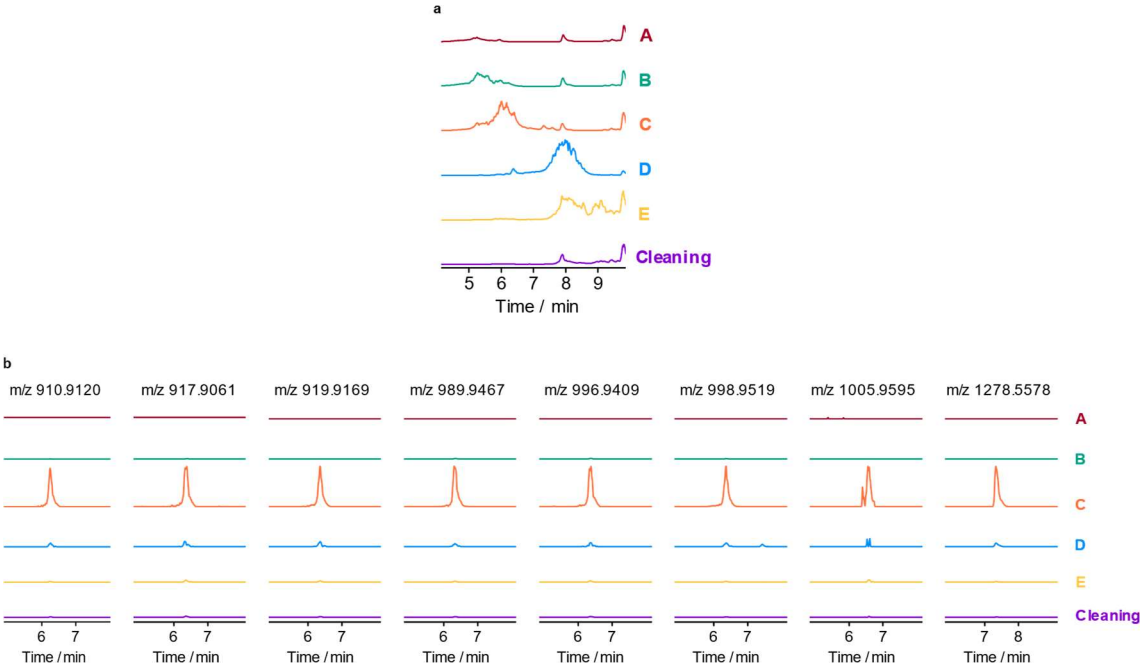


Figure 42 – Results from de MS analysis of resultant fractions from the SPE. **a:** TIC spectra of the SPE fractions. **b:** Relative abundance of the most abundant ions on each SPE fraction.

3.7. Heterologous expression of BGC 418.1 from Bin 90.1

BGC 418.1 was assembled into a modified pET-28 through DiPaC-SLIC. Three amplicons were generated through PCR for the cloning strategy (Fig. 43). The PCR conditions were optimized and the amplified genes were confirmed through direct sequencing. For each SLIC reaction, both the vector backbone and the genes were purified by gel band extraction and colony PCR was used to verify if the reactions were successful. The colonies selected for colony PCR were left to grow in plates containing LB agar medium supplemented with 50 μg mL^{-1} of kanamycin (Fig. 44-45). In each step, a positive transformant was sent to sequencing to verify the integrity of the ligation zones (Fig. 46). Colonies with no mutations were selected to proceed with the DiPaC-SLIC strategy. Gene *mvdA* was cloned into *E. coli* TOP 10 (Fig. 46 a). Genes *mvdB*-orf1-*mvdC* have a mutation in *E. coli* TOP 10 (Fig. 46 b). The mutation causes an alteration of the ORF. As such, a new colony will be selected to continue the DiPaC-SLIC protocol.

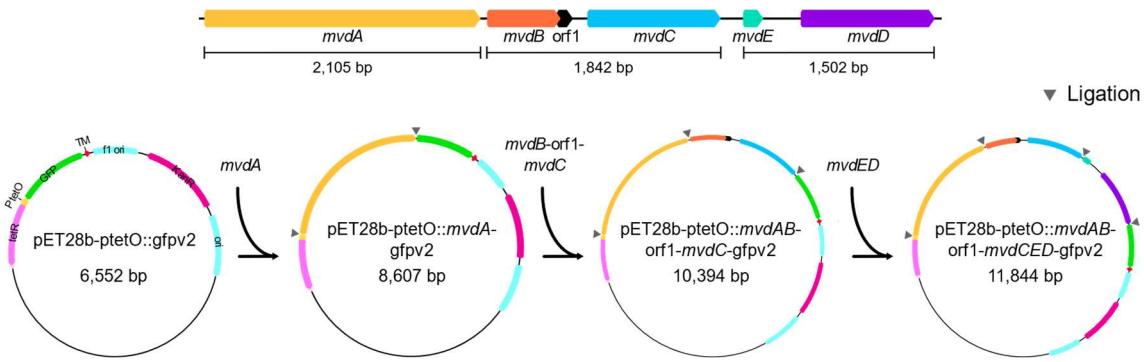


Figure 43 – Cloning strategy for BGC 418.1. The gene cluster was divided in three fragments for the cloning into the vector backbone pET28b-ptetO::gfpv2.

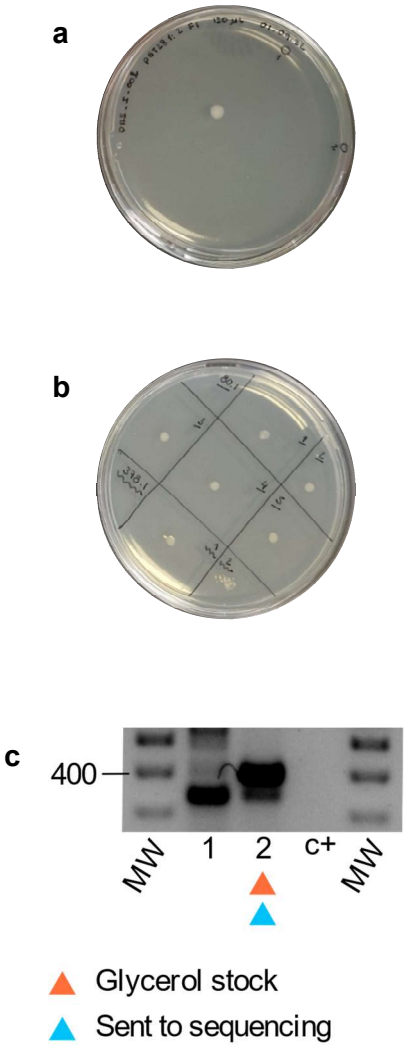


Figure 44 – Colony PCR screening for the SLIC reaction between pET28b-ptetO::gfpv2 and gene *mvdA*. **a:** Selected colonies to undergo colony PCR for the ratio pET28b-ptetO::gfpv2 1:2 *mvdA*. **b:** LB agar medium plate supplemented with 50 µg mL^{-1} of kanamycin used to grow the colonies selected for colony PCR. 378.1 corresponds to BGC 418.1. **c:** Resulting electrophoresis gel from the colony PCR of colonies 1 and 2 using primers screen_ptet_F2 and 418.1_colony_R1 (expected size – 386 bps). MW: NZYDNA Ladder III (NZYTech). c+: positive control using the vector backbone.

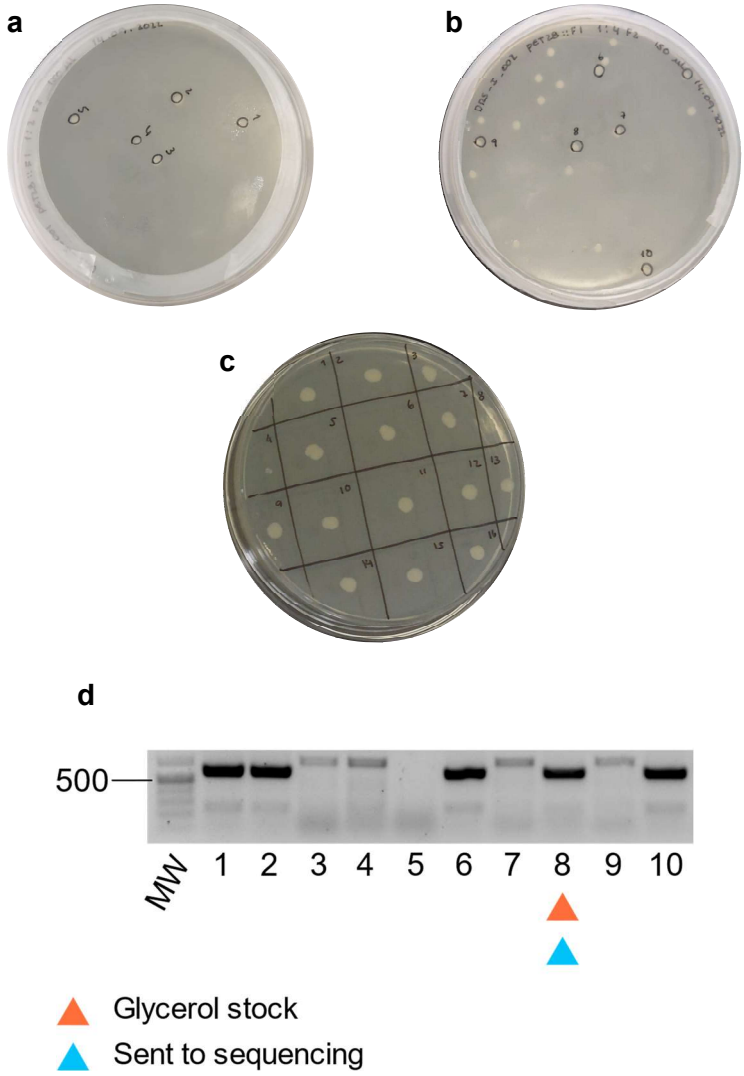


Figure 45 – Colony PCR screening for the SLIC reaction between pET28b-ptetO:: *mvdA*-gfpv2 and genes *mvdBDEF*. **a:** Selected colonies to undergo colony PCR for the ratio pET28b-ptetO:: *mvdA*-gfpv2 1:2 *mvdBDEF*. **b:** Selected colonies to undergo colony PCR for the ratio pET28b-ptetO:: *mvdA*-gfpv2 1:4 *mvdBDEF*. **c:** LB agar medium plate supplemented with 50 µg mL^{-1} of kanamycin used to grow the colonies selected for colony PCR. **d:** Resulting electrophoresis gel from the colony PCR of colonies 1 and 2 using primers 418.1_colony_F2 and 418.1_colony_R2 (expected size – 580 bps). MW: GeneRuler™ 1 kb Plus DNA Ladder (Thermo Scientific™).

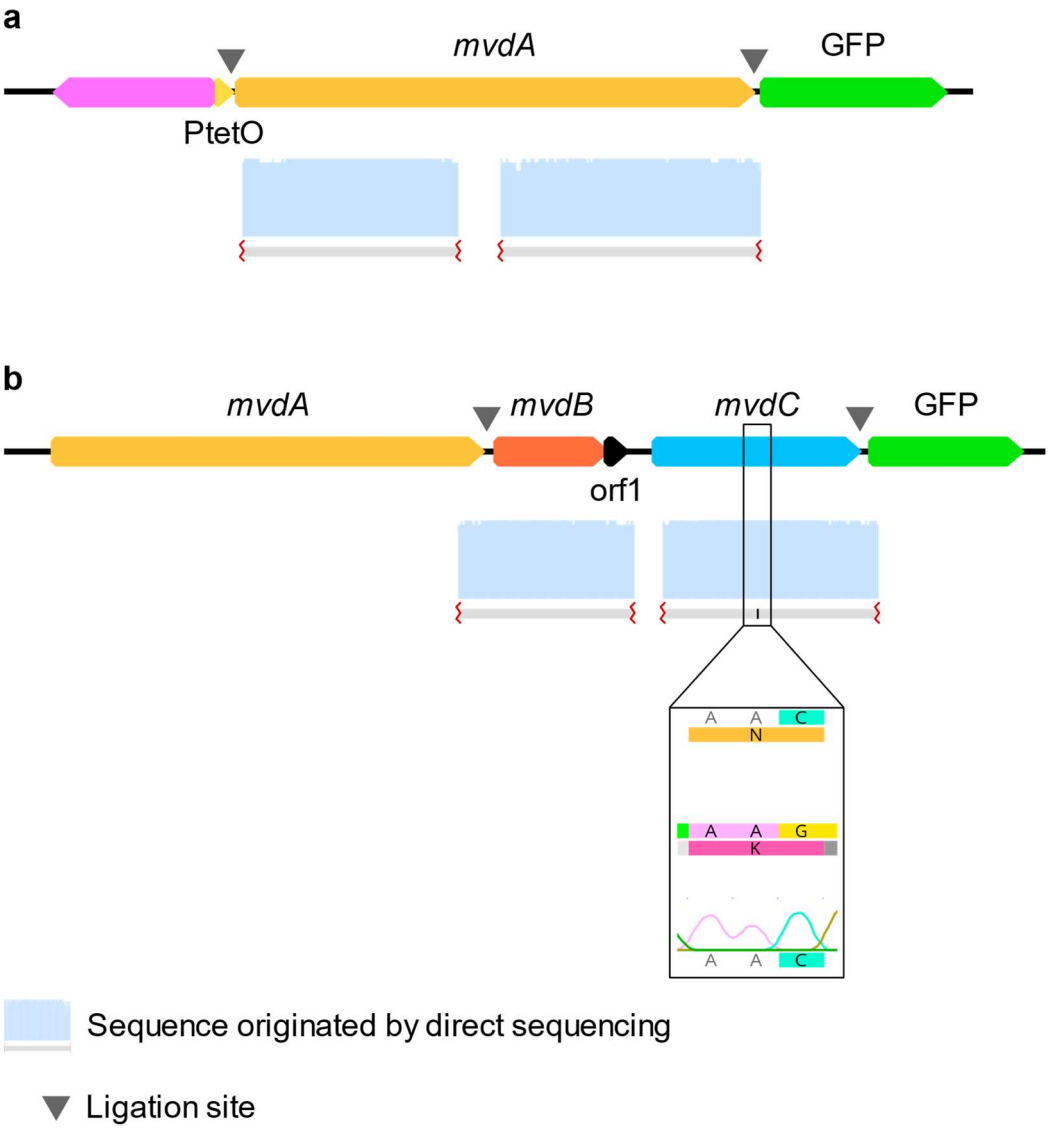


Figure 46 – Direct sequencing results. **a**: Sequencing results at the ligation zones between vector backbone pET28b-ptetO::gfpv2 and gene *mvdA* from BGC 418.1. **b**: Sequencing results at the ligation zones between vector backbone pET28b-ptetO::*mvdA*-gfpv2 and genes *mvdB*-orf1-*C* from BGC 418.1. The mutation site is highlighted. The swap from a lysine to a asparagine residue is visible on the image.

3.8. Heterologous expression of BGC 91.1 from Bin 108

BGC 91.1 was divided into three fragments for the cloning into a modified pET-28 using DiPaC-SLIC (Fig. 47). The fragments were amplified through PCR and the resulting amplicons were purified through gel band excision. The correct amplification of each set of genes was confirmed through direct sequencing (Fig. 48). Currently, we are attempting to clone the first set of genes into the vector backbone.

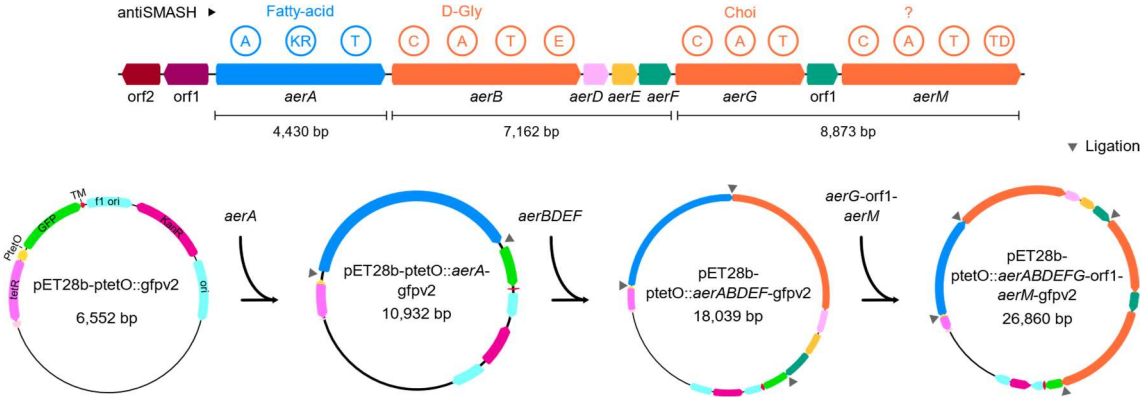


Figure 47 – Cloning strategy for BGC 91.1. The gene cluster was divided in three fragments for the cloning into the vector backbone pET28b-ptetO::gfpv2.

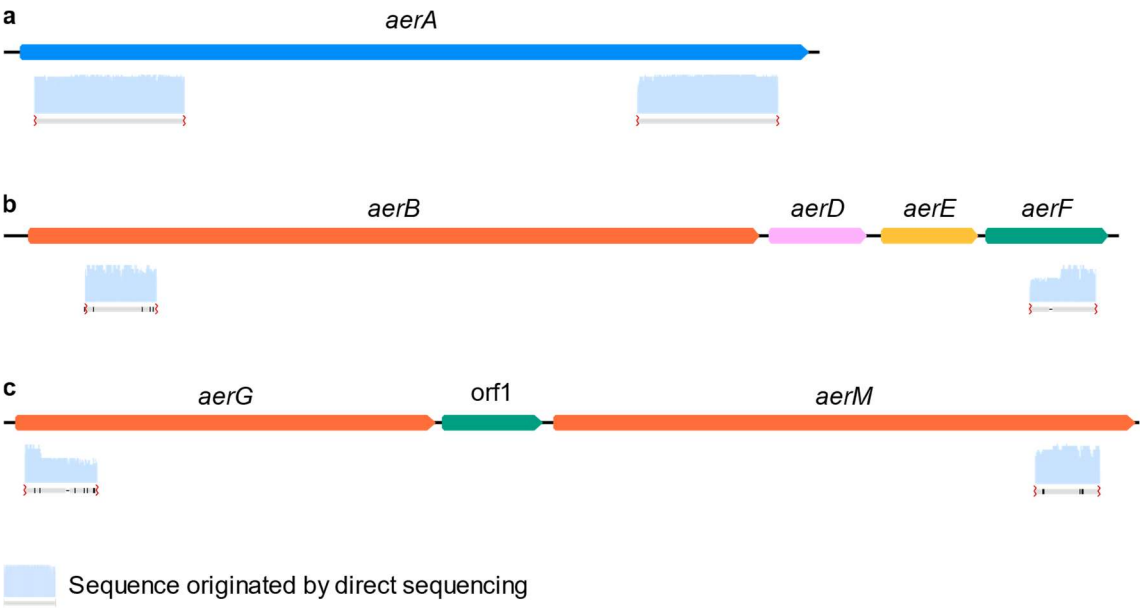


Figure 48 – Direct sequencing results. **a**: *aerA*. **b**: *aerBDEF*. **c**: *aerG*-orf1-*aerM*.

4. Conclusions

In this study, the use of a metagenomic approach allowed the recovery of several bins from a lake biofilm sample. After manual refinement, up to 25 MAGs were recovered. From this 3 represent high and medium-quality cyanobacterial MAGs. Subsequent bioinformatic analysis of the recovered cyanobacterial MAGs revealed 39 complete and near-complete BGCs. A MAG derived microviridin BGC was cloned and heterologously expressed in *E. coli* using DiPaC-SLIC. Several metabolites produced by the transformed bacteria were associated with possible peptide sequences of the microviridins encoded in the target BGC. Nonetheless, some highly expressed metabolites were not matched to a predicted peptide sequence. This BGC contains a gene with no attributed function in the biosynthesis of the microviridins (orf1). MS/MS analysis will be performed to elucidate the structure of the compound and verify if orf1 may have a role on its biosynthesis. Furthermore, compound isolation efforts have been initiated for posterior structural elucidation through Nuclear Magnetic Resonance (NMR). The complete version of the previously referred BGC includes another gene with no recognized function on microviridin biosynthesis (orf2). Analysis of extracts from *E. coli* transformed with the vector pET28b-ptetO::orf1-*mvdABCDEF*-orf2-gfpv2 demonstrated that the most abundant ions matched those from *E. coli* transformed with the vector pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2. However, due to the low abundance of these ions in comparison to what was described for *E. coli* carrying the plasmid pET28b-ptetO::orf1-*mvdABCDEF*-gfpv2, expression will be tested under different conditions. The first two sets of genes from the other microviridin BGC were already cloned into *E. coli* TOP10. However, due to a ORF changing mutation, new colonies will be screened to continue the DiPaC-SLIC strategy. The NRPS genes were successfully amplified from the recovered eDNA and the cloning attempts already started. Overall, our work demonstrates the suitability of metagenomics to access the biosynthetic potential of uncultured cyanobacteria growting as environmental biofilms, as well as the potential of DiPaC-SLIC to heterologously express cyanobacterial MAG-derived BGCs in *E. coli*.

5. References

1 Meeks, J. C. & Elhai, J. Regulation of cellular differentiation in filamentous cyanobacteria in free-living and plant-associated symbiotic growth states. *Microbiol Mol Biol Rev* **66**, 94-121, doi:10.1128/mmbr.66.1.94-121.2002 (2002).

2 Beck, C., Knoop, H., Axmann, I. M. & Steuer, R. The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. *BMC Genomics* **13**, 56, doi:10.1186/1471-2164-13-56 (2012).

3 Gillespie Doreen, E. *et al.* Isolation of Antibiotics Turbomycin A and B from a Metagenomic Library of Soil Microbial DNA. *Applied and Environmental Microbiology* **68**, 4301-4306, doi:10.1128/AEM.68.9.4301-4306.2002 (2002).

4 Garcia-Pichel, F. Cyanobacteria. *Encyclopedia of Microbiology (Third Edition)*, 107-124, doi:10.1016/B978-012373944-5.00250-9 (2009).

5 Hamouda, R. & El-Naggar, N. Cyanobacteria based microbial cell factories for production of industrial products. 277-302, doi:10.1016/B978-0-12-821477-0.00007-6 (2021).

6 Demay, J., Bernard, C., Reinhardt, A. & Marie, B. Natural Products from Cyanobacteria: Focus on Beneficial Activities. *Marine Drugs* **17**, doi:10.3390/md17060320 (2019).

7 Palinska, K. A. & Surosz, W. Taxonomy of cyanobacteria: a contribution to consensus approach. *Hydrobiologia* **740**, 1-11, doi:10.1007/s10750-014-1971-9 (2014).

8 Komárek, J. A polyphasic approach for the taxonomy of cyanobacteria: principles and applications. *European Journal of Phycology* **51**, 346-353, doi:10.1080/09670262.2016.1163738 (2016).

9 Komarek, J., Kaštovský, J., Mares, J. & Johansen, J. Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia - Praha-* **86**, 295-335 (2014).

10 Ramos, V. *et al.* Cyanobacterial diversity held in microbial biological resource centers as a biotechnological asset: the case study of the newly established LEGE culture collection. *J Appl Phycol* **30**, 1437-1451, doi:10.1007/s10811-017-1369-y (2018).

11 Pham, J. V. *et al.* A Review of the Microbial Production of Bioactive Natural Products and Biologics. *Frontiers in Microbiology* **10**, doi:10.3389/fmicb.2019.01404 (2019).

12 Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *Journal of Natural Products* **83**, 770-803, doi:10.1021/acs.jnatprod.9b01285 (2020).

13 Sertuerner. Ueber das Morphium, eine neue salzfähige Grundlage, und die Mekonsäure, als Hauptbestandtheile des Opiums. *Annalen der Physik* **55**, 56-89, doi:10.1002/andp.18170550104 (1817).

14 Ouyang, L. *et al.* Plant natural products: from traditional compounds to new emerging drugs in cancer therapy. *Cell Prolif* **47**, 506-515, doi:10.1111/cpr.12143 (2014).

15 Fleming, A. On the Antibacterial Action of Cultures of a Penicillium, with Special Reference to their Use in the Isolation of B. influenzæ. *Br J Exp Pathol* **10**, 226-236 (1929).

16 Williams, P. G. Panning for chemical gold: marine bacteria as a source of new therapeutics. *Trends Biotechnol* **27**, 45-52, doi:10.1016/j.tibtech.2008.10.005 (2009).

17 Minotti, G., Menna, P., Salvatorelli, E., Cairo, G. & Gianni, L. Anthracyclines: molecular advances and pharmacologic developments in antitumor activity and cardiotoxicity. *Pharmacol Rev* **56**, 185-229, doi:10.1124/pr.56.2.6 (2004).

18 Jose, P. A., Maharshi, A. & Jha, B. Actinobacteria in natural products research: Progress and prospects. *Microbiological Research* **246**, 126708, doi:10.1016/j.micres.2021.126708 (2021).

19 Ciferri, O. & Tiboni, O. The biochemistry and industrial potential of Spirulina. *Annu Rev Microbiol* **39**, 503-526, doi:10.1146/annurev.mi.39.100185.002443 (1985).

20 Zhao, B. *et al.* Anti-obesity effects of Spirulina platensis protein hydrolysate by modulating brain-liver axis in high-fat diet fed mice. *PLoS One* **14**, e0218543, doi:10.1371/journal.pone.0218543 (2019).

21 Fan, X., Cui, Y., Zhang, R. & Zhang, X. Purification and identification of anti-obesity peptides derived from Spirulina platensis. *Journal of Functional Foods* **47**, 350-360, doi:10.1016/j.jff.2018.05.066 (2018).

22 Luesch, H., Moore, R. E., Paul, V. J., Mooberry, S. L. & Corbett, T. H. Isolation of Dolastatin 10 from the Marine Cyanobacterium Symploca Species VP642 and Total Stereochemistry and Biological Evaluation of Its Analogue Symplostatin 1. *Journal of Natural Products* **64**, 907-910, doi:10.1021/np010049y (2001).

23 Gao, G., Wang, Y., Hua, H., Li, D. & Tang, C. Marine Antitumor Peptide Dolastatin 10: Biological Activity, Structural Modification and Synthetic Chemistry. *Mar Drugs* **19**, doi:10.3390/md19070363 (2021).

24 Chu, L., Huang, J., Muhammad, M., Deng, Z. & Gao, J. Genome mining as a biotechnological tool for the discovery of novel marine natural products. *Crit Rev Biotechnol* **40**, 571-589, doi:10.1080/07388551.2020.1751056 (2020).

25 Bachmann, B. O., Van Lanen, S. G. & Baltz, R. H. Microbial genome mining for accelerated natural products discovery: is a renaissance in the making? *J Ind Microbiol Biotechnol* **41**, 175-184, doi:10.1007/s10295-013-1389-9 (2014).

26 Stone, M. J. & Williams, D. H. On the evolution of functional secondary metabolites (natural products). *Molecular Microbiology* **6**, 29-34, doi:10.1111/j.1365-2958.1992.tb00834.x (1992).

27 Kaneko, T. *et al.* Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res* **3**, 109-136, doi:10.1093/dnares/3.3.109 (1996).

28 Alvarenga, D. O., Fiore, M. F. & Varani, A. M. A Metagenomic Approach to Cyanobacterial Genomics. *Front Microbiol* **8**, 809, doi:10.3389/fmicb.2017.00809 (2017).

29 Hirose, Y. *et al.* Genome sequencing of the NIES Cyanobacteria collection with a focus on the heterocyst-forming clade. *DNA Research* **28**, dsab024, doi:10.1093/dnares/dsab024 (2021).

30 Shih, P. M. *et al.* Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A* **110**, 1053-1058, doi:10.1073/pnas.1217107110 (2013).

31 Walsh, C. T. & Tang, Y. Chapter 3. Peptide Derived Natural Products. *Natural Product Biosynthesis: Chemical Logic and Enzymatic Machinery*, Royal Society of Chemistry (2017).

32 Miller, B. R. & Gulick, A. M. Structural Biology of Nonribosomal Peptide Synthetases. *Methods Mol Biol* **1401**, 3-29, doi:10.1007/978-1-4939-3375-4_1 (2016).

33 Vestola, J. *et al.* Hassallidins, antifungal glycolipopeptides, are widespread among cyanobacteria and are the end-product of a nonribosomal pathway. *Proceedings of the National Academy of Sciences* **111**, E1909-E1917, doi:10.1073/pnas.1320913111 (2014).

34 Fewer, D. P. *et al.* The non-ribosomal assembly and frequent occurrence of the protease inhibitors spumigins in the bloom-forming cyanobacterium Nodularia spumigena. *Molecular Microbiology* **73**, 924-937, doi:10.1111/j.1365-2958.2009.06816.x (2009).

35 Ishida, K. *et al.* Biosynthesis and structure of aeruginoside 126A and 126B, cyanobacterial peptide glycosides bearing a 2-carboxy-6-hydroxyoctahydroindole moiety. *Chemistry & biology* **14**, 565-576, doi:10.1016/j.chembiol.2007.04.006 (2007).

36 Kehr, J. C., Gatte Picchi, D. & Dittmann, E. Natural product biosyntheses in cyanobacteria: A treasure trove of unique enzymes. *Beilstein J Org Chem* **7**, 1622-1635, doi:10.3762/bjoc.7.191 (2011).

37 Walsh, C. T. & Tang, Y. Chapter 2. Polyketide Natural Products. *Natural Product Biosynthesis: Chemical Logic and Enzymatic Machinery*, Royal Society of Chemistry (2017).

38 Micallef, M. L., D'Agostino, P. M., Al-Sinawi, B., Neilan, B. A. & Moffitt, M. C. Exploring cyanobacterial genomes for natural product biosynthesis pathways. *Mar Genomics* **21**, 1-12, doi:10.1016/j.margen.2014.11.009 (2015).

39 Freitas, S. *et al.* Structure and Biosynthesis of Desmamides A–C, Lipoglycopeptides from the Endophytic Cyanobacterium Desmonostoc muscorum LEGE 12446. *Journal of Natural Products* **85**, 1704-1714, doi:10.1021/acs.jnatprod.2c00162 (2022).

40 Gehringer, M. M. *et al.* Nodularin, a cyanobacterial toxin, is synthesized in planta by symbiotic Nostoc sp. *The ISME Journal* **6**, 1834-1847, doi:10.1038/ismej.2012.25 (2012).

41 Christiansen, G., Fastner, J., Erhard, M., Börner, T. & Dittmann, E. Microcystin biosynthesis in planktothrix: genes, evolution, and manipulation. *J Bacteriol* **185**, 564-572, doi:10.1128/jb.185.2.564-572.2003 (2003).

42 Arnison, P. G. *et al.* Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. *Nat Prod Rep* **30**, 108-160, doi:10.1039/c2np20085f (2013).

43 Hemscheidt, T. K. Microviridin biosynthesis. *Methods Enzymol* **516**, 25-35, doi:10.1016/b978-0-12-394291-3.00023-x (2012).

44 Czekster, C. M., Ge, Y. & Naismith, J. H. Mechanisms of cyanobactin biosynthesis. *Curr Opin Chem Biol* **35**, 80-88, doi:10.1016/j.cbpa.2016.08.029 (2016).

45 Repka, L. M., Chekan, J. R., Nair, S. K. & van der Donk, W. A. Mechanistic Understanding of Lanthipeptide Biosynthetic Enzymes. *Chemical Reviews* **117**, 5457-5520, doi:10.1021/acs.chemrev.6b00591 (2017).

46 Kleigrewe, K., Gerwick, L., Sherman, D. H. & Gerwick, W. H. Unique marine derived cyanobacterial biosynthetic genes for chemical diversity. *Natural Product Reports* **33**, 348-364, doi:10.1039/c5np00097a (2016).

47 Yang, C. *et al.* A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Computational and Structural Biotechnology Journal* **19**, 6301-6314, doi:10.1016/j.csbj.2021.11.028 (2021).

48 Stres, B. & Kronegger, L. Shift in the paradigm towards next-generation microbiology. *FEMS Microbiology Letters* **366**, fnz159, doi:10.1093/femsle/fnz159 (2019).

49 Brady, S. F., Simmons, L., Kim, J. H. & Schmidt, E. W. Metagenomic approaches to natural products from free-living and symbiotic organisms. *Natural Product Reports* **26**, 1488-1503, doi:10.1039/B817078A (2009).

50 Ruppert, K. M., Kline, R. J. & Rahman, M. S. Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation* **17**, e00547, doi:10.1016/j.gecco.2019.e00547 (2019).

51 Sanders, J. G. *et al.* Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads. *Genome Biology* **20**, 226, doi:10.1186/s13059-019-1834-9 (2019).

52 Lin, H.-H. & Liao, Y.-C. Accurate binning of metagenomic contigs via automated clustering sequences using information of genomic signatures and marker genes. *Scientific Reports* **6**, 24175, doi:10.1038/srep24175 (2016).

53 Yu, G., Jiang, Y., Wang, J., Zhang, H. & Luo, H. BMC3C: binning metagenomic contigs using codon usage, sequence composition and read coverage. *Bioinformatics* **34**, 4172-4179, doi:10.1093/bioinformatics/bty519 (2018).

54 Wang, Z., Wang, Z., Lu, Y. Y., Sun, F. & Zhu, S. SolidBin: improving metagenome binning with semi-supervised normalized cut. *Bioinformatics* **35**, 4229-4238, doi:10.1093/bioinformatics/btz253 (2019).

55 Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**, 725-731, doi:10.1038/nbt.3893 (2017).

56 Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome Biology* **21**, 30, doi:10.1186/s13059-020-1935-5 (2020).

57 Illumina. NovaSeq™ 6000 Sequencing System Immense discovery power for deeper insights. (2022).

58 Illumina. NextSeq™ 550 System Tunable sequencing output and array scanning on a single instrument Specifiction Sheet. (2021).

59 Illumina. MiSeq™ System Speed and simplicity for targeted resequencing and small-genome sequencing Specifiction Sheet. (2022).

60 Illumina. HiSeq™ Sequencing Systems Redefining the trajectory of sequencing. Specification Sheet. (2014).

61 Ion Torrent Next-Generation Sequencing Technology. *Thermo Fisher*, https://www.thermofisher.com/pt/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html (28/08/2022).

62 Jeon, S. A. *et al.* Comparison of the MGISEQ-2000 and Illumina HiSeq 4000 sequencing platforms for RNA sequencing. *Genomics Inform* **17**, e32, doi:10.5808/GI.2019.17.3.e32 (2019).

63 Li, Q. *et al.* Reliable multiplex sequencing with rare index mis-assignment on DNB-based NGS platform. *BMC Genomics* **20**, 215, doi:10.1186/s12864-019-5569-5 (2019).

64 Zhu, F.-Y. *et al.* Comparative performance of the BGISEQ-500 and Illumina HiSeq4000 sequencing platforms for transcriptome analysis in plants. *Plant Methods* **14**, 69, doi:10.1186/s13007-018-0337-0 (2018).

65 Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59, doi:10.1038/nature07517 (2008).

66 Ali, M. M. *et al.* Rolling circle amplification: a versatile tool for chemical biology, materials science and medicine. *Chemical Society Reviews* **43**, 3324-3341, doi:10.1039/C3CS60439J (2014).

67 Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81, doi:10.1126/science.1181498 (2010).

68 How nanopore sequencing works. *Oxford Nanopore Technologies*, https://nanoporetech.com/how-it-works (28/08/2022).

69 Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138, doi:10.1126/science.1162986 (2009).

70 Chen, Y. *et al.* SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience* **7**, 1-6, doi:10.1093/gigascience/gix120 (2018).

71 Del Fabbro, C., Scalabrin, S., Morgante, M. & Giorgi, F. M. An extensive evaluation of read trimming effects on Illumina NGS data analysis. *PLoS One* **8**, e85024, doi:10.1371/journal.pone.0085024 (2013).

72 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

73 FastQC. *Babraham Bioinformatics*, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (29/08/2022).

74 FASTX-Toolkit. http://hannonlab.cshl.edu/fastx_toolkit/ (29/08/2022).

75 N50 statistics. https://www.metagenomics.wiki/tools/assembly/n50 (29/08/2022).

76 Hackl, T., Hedrich, R., Schultz, J. & Förster, F. proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. *Bioinformatics* **30**, 3004-3011, doi:10.1093/bioinformatics/btu392 (2014).

77 Bussi, Y., Kapon, R. & Reich, Z. Large-scale k-mer-based analysis of the informational properties of genomes, comparative genomics and taxonomy. *PLoS One* **16**, e0258693, doi:10.1371/journal.pone.0258693 (2021).

78 Fukasawa, Y., Ermini, L., Wang, H., Carty, K. & Cheung, M. S. LongQC: A Quality Control Tool for Third Generation Sequencing Long Read Data. *G3 (Bethesda)* **10**, 1193-1196, doi:10.1534/g3.119.400864 (2020).

79 Hufnagel, D. E., Hufford, M. B. & Seetharam, A. S. SequelTools: a suite of tools for working with PacBio Sequel raw sequence data. *BMC Bioinformatics* **21**, 429, doi:10.1186/s12859-020-03751-8 (2020).

80 BBMap Guide. *Joint Genome Institute*, https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbmap-guide/ (29/08/2022).

81 BBDuk Guide. *Joint Genome Institute*, https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/ (29/08/2022).

82 Kalyanaraman, A. Genome Assembly. *Encyclopedia of Parallel Computing*, 755-768, doi:10.1007/978-0-387-09766-4_402 (2011).

83 Haider, B. *et al.* Omega: an overlap-graph de novo assembler for metagenomics. *Bioinformatics* **30**, 2717-2722, doi:10.1093/bioinformatics/btu395 (2014).

84 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3-11, doi:10.1016/j.ymeth.2016.02.020 (2016).

85 Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res* **27**, 824-834, doi:10.1101/gr.213959.116 (2017).

86 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).

87 Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**, 1103-1110, doi:10.1038/s41592-020-00971-x (2020).

88 Chen, Y. *et al.* Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat Commun* **12**, 60, doi:10.1038/s41467-020-20236-7 (2021).

89 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

90 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

91 Nissen, J. N. *et al.* Improved metagenome binning and assembly using deep variational autoencoders. *Nature Biotechnology* **39**, 555-560, doi:10.1038/s41587-020-00777-4 (2021).

92 Imelfort, M. *et al.* GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2**, e603 (2014).

93 Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nature Methods* **11**, 1144-1146, doi:10.1038/nmeth.3103 (2014).

94 Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605-607, doi:10.1093/bioinformatics/btv638 (2016).

95 Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319, doi:10.7717/peerj.1319 (2015).

96 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).

97 Zhu, W., Lomsadze, A. & Borodovsky, M. Ab initio gene identification in metagenomic sequences. *Nucleic Acids Res* **38**, e132, doi:10.1093/nar/gkq275 (2010).

98 Zhang, S. W., Jin, X. Y. & Zhang, T. Gene Prediction in Metagenomic Fragments with Deep Learning. *Biomed Res Int* **2017**, 4740354, doi:10.1155/2017/4740354 (2017).

99 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

100 Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925-1927, doi:10.1093/bioinformatics/btz848 (2019).

101 Asnicar, F. *et al.* Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. *Nat Commun* **11**, 2500, doi:10.1038/s41467-020-16366-7 (2020).

102 Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* **2**, 1533-1542, doi:10.1038/s41564-017-0012-7 (2017).

103 Rego, A. *et al.* Secondary metabolite biosynthetic diversity in Arctic Ocean metagenomes. *Microb Genom* **7**, doi:10.1099/mgen.0.000731 (2021).

104 Rego, A. *et al.* Diversity of Bacterial Biosynthetic Genes in Maritime Antarctica. *Microorganisms* **8**, doi:10.3390/microorganisms8020279 (2020).

105 McCuaig, B., Peña-Castillo, L. & Dufour, S. C. Metagenomic analysis suggests broad metabolic potential in extracellular symbionts of the bivalve Thyasira cf. gouldi. *Animal Microbiome* **2**, 7, doi:10.1186/s42523-020-00025-9 (2020).

106 Casso, M., Turon, M., Marco, N., Pascual, M. & Turon, X. The Microbiome of the Worldwide Invasive Ascidian Didemnum vexillum. *Frontiers in Marine Science* **7**, doi:10.3389/fmars.2020.00201 (2020).

107 Iqbal, H. A., Low-Beinart, L., Obiajulu, J. U. & Brady, S. F. Natural Product Discovery through Improved Functional Metagenomics in Streptomyces. *Journal of the American Chemical Society* **138**, 9341-9344, doi:10.1021/jacs.6b02921 (2016).

108 Wu, C., Shang, Z., Lemetre, C., Ternei, M. A. & Brady, S. F. Cadasides, Calcium-Dependent Acidic Lipopeptides from the Soil Metagenome That Are Active against Multidrug-Resistant Bacteria. *Journal of the American Chemical Society* **141**, 3910-3919, doi:10.1021/jacs.8b12087 (2019).

109 Li, L. *et al.* Biosynthetic Interrogation of Soil Metagenomes Reveals Metamarin, an Uncommon Cyclomarin Congener with Activity against Mycobacterium tuberculosis. *Journal of Natural Products* **84**, 1056-1066, doi:10.1021/acs.jnatprod.0c01104 (2021).

110 Shamim, K., Mujawar, S. Y. & Mutnale, M. Chapter 12 - Metagenomics a modern approach to reveal the secrets of unculturable microbes. *Advances in Biological Science Research*, 177-195, doi:10.1016/B978-0-12-817497-5.00012-4 (2019).

111 Nierman, W. C. & Feldblyum, T. V. Genomic Library. *Encyclopedia of Genetics*, 865-872, doi:10.1006/rwgn.2001.0559 (2001).

112 Saraswathy, N. & Ramalingam, P. 4 - High capacity vectors. *Concepts and Techniques in Genomics and Proteomics*, 49-56, doi:10.1533/9781908818058.49 (2011).

113 Martínez, A. & Osburne, M. S. Chapter Seven - Preparation of Fosmid Libraries and Functional Metagenomic Analysis of Microbial Community DNA. *Methods in Enzymology* **531**, 123-142, doi:10.1016/B978-0-12-407863-5.00007-1 (2013).

114 Nora, L. C. *et al.* The art of vector engineering: towards the construction of next-generation genetic tools. *Microbial Biotechnology* **12**, 125-147, doi:10.1111/1751-7915.13318 (2019).

115 Stevenson, L. J. *et al.* Metathramycin, a new bioactive aureolic acid discovered by heterologous expression of a metagenome derived biosynthetic pathway. *RSC Chemical Biology* **2**, 556-567, doi:10.1039/D0CB00228C (2021).

116 Zhang, J. J., Tang, X. & Moore, B. S. Genetic platforms for heterologous expression of microbial natural products. *Natural product reports* **36**, 1313-1332 (2019).

117 Stöveken, J. *et al.* Successful heterologous expression of a novel chitinase identified by sequence analyses of the metagenome from a chitin-enriched soil sample. *Journal of Biotechnology* **201**, 60-68, doi:10.1016/j.jbiotec.2014.09.010 (2015).

118 Guo, C.-J. *et al.* Discovery of Reactive Microbiota-Derived Metabolites that Inhibit Host Proteases. *Cell* **168**, 517-526.e518, doi:10.1016/j.cell.2016.12.021 (2017).

119 D'Agostino, P. M. & Gulder, T. A. M. Direct Pathway Cloning Combined with Sequence- and Ligation-Independent Cloning for Fast Biosynthetic Gene Cluster Refactoring and Heterologous Expression. *ACS Synth Biol* **7**, 1702-1708, doi:10.1021/acssynbio.8b00151 (2018).

120 Eusébio, N. *et al.* Discovery and Heterologous Expression of Microginins from Microcystis aeruginosa LEGE 91341. *ACS Synthetic Biology*, doi:10.1021/acssynbio.2c00389 (2022).

121 Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421-432, doi:10.1093/bioinformatics/bty648 (2019).

122 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

123 Blin, K., Shaw, S., Kautsar, S. A., Medema, M. H. & Weber, T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Research* **49**, D639-D643, doi:10.1093/nar/gkaa978 (2021).

124 Agrawal, P., Amir, S., Deepak, Barua, D. & Mohanty, D. RiPPMiner-Genome: A Web Resource for Automated Prediction of Crosslinked Chemical Structures of RiPPs by Genome Mining. *J Mol Biol* **433**, 166887, doi:10.1016/j.jmb.2021.166887 (2021).

125 Greunke, C. *et al.* Direct Pathway Cloning (DiPaC) to unlock natural product biosynthetic potential. *Metab Eng* **47**, 334-345, doi:10.1016/j.ymben.2018.03.010 (2018).

126 Duell, E. R. *et al.* Direct pathway cloning of the sodorifen biosynthetic gene cluster and recombinant generation of its product in E. coli. *Microb Cell Fact* **18**, 32, doi:10.1186/s12934-019-1080-6 (2019).

127 Philmus, B., Christiansen, G., Yoshida, W. Y. & Hemscheidt, T. K. Post-translational modification in microviridin biosynthesis. *Chembiochem* **9**, 3066-3073, doi:10.1002/cbic.200800560 (2008).

128 Weiz, Annika R. *et al.* Leader Peptide and a Membrane Protein Scaffold Guide the Biosynthesis of the Tricyclic Peptide Microviridin. *Chemistry & Biology* **18**, 1413-1421, doi:10.1016/j.chembiol.2011.09.011 (2011).

129 Shirmast, P. *et al.* Structural characterization of a GNAT family acetyltransferase from Elizabethkingia anophelis bound to acetyl-CoA reveals a new dimeric interface. *Scientific Reports* **11**, 1274, doi:10.1038/s41598-020-79649-5 (2021).

130 do Amaral, S. C. *et al.* Current Knowledge on Microviridin from Cyanobacteria. *Marine Drugs* **19**, doi:10.3390/md19010017 (2021).

131 Fujii, K., Sivonen, K., Naganawa, E. & Harada, K.-i. Non-Toxic Peptides from Toxic Cyanobacteria, Oscillatoria agardhii. *Tetrahedron* **56**, 725-733, doi:10.1016/S0040-4020(99)01017-0 (2000).

132 Gu, W., Dong, S. H., Sarkar, S., Nair, S. K. & Schmidt, E. W. The Biochemistry and Structural Biology of Cyanobactin Pathways: Enabling Combinatorial Biosynthesis. *Methods Enzymol* **604**, 113-163, doi:10.1016/bs.mie.2018.03.002 (2018).

133 Martins, J. & Vasconcelos, V. Cyanobactins from Cyanobacteria: Current Genetic and Chemical State of Knowledge. *Marine drugs* **13**, 6910-6946, doi:10.3390/md13116910 (2015).

134 Sardar, D., Pierce, E., McIntosh, J. A. & Schmidt, E. W. Recognition Sequences and Substrate Evolution in Cyanobactin Biosynthesis. *ACS Synthetic Biology* **4**, 167-176, doi:10.1021/sb500019b (2015).

135 Leikoski, N. *et al.* Highly diverse cyanobactins in strains of the genus Anabaena. *Appl Environ Microbiol* **76**, 701-709, doi:10.1128/aem.01061-09 (2010).

136 Ahmed, M. N. *et al.* Potent Inhibitor of Human Trypsins from the Aeruginosin Family of Natural Products. *ACS Chemical Biology* **16**, 2537-2546, doi:10.1021/acschembio.1c00611 (2021).

137 Fewer, D. P. *et al.* New structural variants of aeruginosin produced by the toxic bloom forming cyanobacterium Nodularia spumigena. *PLoS One* **8**, e73618, doi:10.1371/journal.pone.0073618 (2013).

138 Zervou, S. K., Gkelis, S., Kaloudis, T., Hiskia, A. & Mazur-Marzec, H. New microginins from cyanobacteria of Greek freshwaters. *Chemosphere* **248**, 125961, doi:10.1016/j.chemosphere.2020.125961 (2020).

139 Jones, M. R. *et al.* CyanoMetDB, a comprehensive public database of secondary metabolites from cyanobacteria. *Water Research* **196**, 117017, doi:10.1016/j.watres.2021.117017 (2021).

140 US20110034680A1. (2011).

141 Rounge, T. B., Rohrlack, T., Nederbragt, A. J., Kristensen, T. & Jakobsen, K. S. A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a Planktothrix rubescens strain. *BMC Genomics* **10**, 396-396, doi:10.1186/1471-2164-10-396 (2009).

142 Eusebio, N. *et al.* Distribution and diversity of dimetal-carboxylate halogenases in cyanobacteria. *BMC Genomics* **22**, 633, doi:10.1186/s12864-021-07939-x (2021).

143 Ziemert, N., Ishida, K., Liaimer, A., Hertweck, C. & Dittmann, E. Ribosomal synthesis of tricyclic depsipeptides in bloom-forming cyanobacteria. *Angew Chem Int Ed Engl* **47**, 7756-7759, doi:10.1002/anie.200802730 (2008).

6. Supplementary Information

Supplementary Table 1 – Predicted ions for the precursor peptide MvdE with two ester bonds.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
MSKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5337.6234	1780.215076	2669.818976	5338.630676
SKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5206.5829	1736.534909	2604.298726	5207.590176
KNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5119.5508	1707.524209	2560.782676	5120.558076
NVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4991.4559	1664.825909	2496.735226	4992.463176
VKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4877.4130	1626.811609	2439.713776	4878.420276
KVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4778.3445	1593.788776	2390.179526	4779.351776
VSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4650.2496	1551.090476	2326.132076	4651.256876
SAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4551.1812	1518.067676	2276.597876	4552.188476
APKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4464.1491	1489.056976	2233.081826	4465.156376
PKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4393.1120	1465.377943	2197.563276	4394.119276
KAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4296.0593	1433.027043	2149.036926	4297.066576
AVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4167.9643	1390.328709	2084.989426	4168.971576
VPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4096.9272	1366.649676	2049.470876	4097.934476
PFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3997.8588	1333.626876	1999.936676	3998.866076
FFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3900.8060	1301.275943	1951.410276	3901.813276
FARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3753.7376	1252.253143	1877.876076	3754.744876
ARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3606.6692	1203.230343	1804.341876	3607.676476
RFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3535.6321	1179.551309	1768.823326	3536.639376
FLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3379.5310	1127.517609	1690.772776	3380.538276
LAEQAVEANNSNSAPYPTTLKYPSDWEDY	3232.4625	1078.494776	1617.238526	3233.469776
AEQAVEANNSNSAPYPTTLKYPSDWEDY	3119.3785	1040.800109	1560.696526	3120.385776
EQAVEANNSNSAPYPTTLKYPSDWEDY	3048.3414	1017.121076	1525.177976	3049.348676
QAVEANNSNSAPYPTTLKYPSDWEDY	2919.2988	974.106876	1460.656676	2920.306076
AVEANNSNSAPYPTTLKYPSDWEDY	2791.2402	931.420676	1396.627376	2792.247476
VEANNSNSAPYPTTLKYPSDWEDY	2720.2031	907.741643	1361.108826	2721.210376
EANNSNSAPYPTTLKYPSDWEDY	2621.1347	874.718843	1311.574626	2622.141976
ANNSNSAPYPTTLKYPSDWEDY	2492.0921	831.704643	1247.053326	2493.099376
NNSNSAPYPTTLKYPSDWEDY	2421.0550	808.025609	1211.534776	2422.062276
NSNSAPYPTTLKYPSDWEDY	2307.0120	770.011276	1154.513276	2308.019276
SNSAPYPTTLKYPSDWEDY	2192.9691	731.996976	1097.491826	2193.976376
NSAPYPTTLKYPSDWEDY	2105.9371	702.986309	1053.975826	2106.944376
SAPYPTTLKYPSDWEDY	1991.8941	664.971976	996.954326	1992.901376
APYPTTLKYPSDWEDY	1904.8621	635.961309	953.438326	1905.869376
PYPTTLKYPSDWEDY	1833.8250	612.282276	917.919776	1834.832276

Supplementary Table 1 – Continuation.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
YPTTLKYPSDWEDY	1736.7722	579.931343	869.393376	1737.779476
AC-YPTTLKYPSDWEDY	1778.7828	593.934876	890.398676	1779.790076

Supplementary Table 2 – Predicted ions for the precursor peptide MvdE with one ester bond between tyrosine and aspartate.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
MSKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5369.6496	1790.890476	2685.832076	5370.656876
SKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5238.6091	1747.210309	2620.311826	5239.616376
KNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5151.5771	1718.199643	2576.795826	5152.584376
NVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5023.4821	1675.501309	2512.748326	5024.489376
VKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4909.4392	1637.487009	2455.726876	4910.446476
KVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4810.3708	1604.464209	2406.192676	4811.378076
VSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4682.2758	1561.765876	2342.145176	4683.283076
SAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4583.2074	1528.743076	2292.610976	4584.214676
APKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4496.1754	1499.732409	2249.094976	4497.182676
PKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4425.1382	1476.053343	2213.576376	4426.145476
KAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4328.0855	1443.702443	2165.050026	4329.092776
AVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4199.9905	1401.004109	2101.002526	4200.997776
VPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4128.9534	1377.325076	2065.483976	4129.960676
PFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4029.8850	1344.302276	2015.949776	4030.892276
FFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3932.8322	1311.951343	1967.423376	3933.839476
FARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3785.7638	1262.928543	1893.889176	3786.771076
ARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3638.6954	1213.905743	1820.354976	3639.702676
RFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3567.6583	1190.226709	1784.836426	3568.665576
FLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3411.5572	1138.193009	1706.785876	3412.564476
LAEQAVEANNSNSAPYPTTLKYPSDWEDY	3264.4888	1089.170209	1633.251676	3265.496076
AEQAVEANNSNSAPYPTTLKYPSDWEDY	3151.4047	1051.475509	1576.709626	3152.411976
EQAVEANNSNSAPYPTTLKYPSDWEDY	3080.3676	1027.796476	1541.191076	3081.374876
QAVEANNSNSAPYPTTLKYPSDWEDY	2951.3250	984.782276	1476.669776	2952.332276
AVEANNSNSAPYPTTLKYPSDWEDY	2823.2664	942.096076	1412.640476	2824.273676
VEANNSNSAPYPTTLKYPSDWEDY	2752.2293	918.417043	1377.121926	2753.236576
EANNSNSAPYPTTLKYPSDWEDY	2653.1609	885.394243	1327.587726	2654.168176
ANNSNSAPYPTTLKYPSDWEDY	2524.1183	842.380043	1263.066426	2525.125576
NNSNSAPYPTTLKYPSDWEDY	2453.0812	818.701009	1227.547876	2454.088476
NSNSAPYPTTLKYPSDWEDY	2339.0382	780.686676	1170.526376	2340.045476

Supplementary Table 2 – Continuation.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
SNSAPYPTTLKYPSDWEDY	2224.9953	742.672376	1113.504926	2226.002576
NSAPYPTTLKYPSDWEDY	2137.9633	713.661709	1069.988926	2138.970576
SAPYPTTLKYPSDWEDY	2023.9204	675.647409	1012.967476	2024.927676
APYPTTLKYPSDWEDY	1936.8883	646.636709	969.451426	1937.895576
PYPTTLKYPSDWEDY	1865.8512	622.957676	933.932876	1866.858476
YPTTLKYPSDWEDY	1768.7985	590.606776	885.406526	1769.805776
AC-YPTTLKYPSDWEDY	1810.8090	604.610276	906.411776	1811.816276

Supplementary Table 3 – Predicted ions for the precursor peptide MvdE with one ester bond between serine and glutamate.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
MSKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5355.6339	1786.218576	2678.824226	5356.641176
SKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5224.5934	1742.538409	2613.303976	5225.600676
KNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5137.5614	1713.527743	2569.787976	5138.568676
NVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5009.4664	1670.829409	2505.740476	5010.473676
VKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4895.4235	1632.815109	2448.719026	4896.430776
KVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4796.3551	1599.792309	2399.184826	4797.362376
VSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4668.2601	1557.093976	2335.137326	4669.267376
SAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4569.1917	1524.071176	2285.603126	4570.198976
APKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4482.1597	1495.060509	2242.087126	4483.166976
PKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4411.1226	1471.381476	2206.568576	4412.129876
KAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4314.0698	1439.030543	2158.042176	4315.077076
AVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4185.9749	1396.332243	2093.994726	4186.982176
VPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4114.9377	1372.653176	2058.476126	4115.944976
PFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4015.8693	1339.630376	2008.941926	4016.876576
FFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3918.8166	1307.279476	1960.415576	3919.823876
FARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3771.7482	1258.256676	1886.881376	3772.755476
ARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3624.6797	1209.233843	1813.347126	3625.686976
RFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3553.6426	1185.554809	1777.828576	3554.649876
FLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3397.5415	1133.521109	1699.778026	3398.548776
LAEQAVEANNSNSAPYPTTLKYPSDWEDY	3250.4731	1084.498309	1626.243826	3251.480376
AEQAVEANNSNSAPYPTTLKYPSDWEDY	3137.3890	1046.803609	1569.701776	3138.396276
EQAVEANNSNSAPYPTTLKYPSDWEDY	3066.3519	1023.124576	1534.183226	3067.359176
QAVEANNSNSAPYPTTLKYPSDWEDY	2037.3093	680.110376	1019.661926	2038.316576
AVEANNSNSAPYPTTLKYPSDWEDY	2809.2508	937.424209	1405.632676	2810.258076
VEANNSNSAPYPTTLKYPSDWEDY	2738.2136	913.745143	1370.114076	2739.220876
EANNSNSAPYPTTLKYPSDWEDY	2639.1452	880.722343	1320.579876	2640.152476

Supplementary Table 3 – Continuation.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
ANNSNSAPYPTTLKYPSDWEDY	2510.1026	837.708143	1256.058576	2511.109876
NNSNSAPYPTTLKYPSDWEDY	2439.0655	814.029109	1220.540026	2440.072776
NSNSAPYPTTLKYPSDWEDY	2325.0226	776.014809	1163.518576	2326.029876
SNSAPYPTTLKYPSDWEDY	2210.9797	738.000509	1106.497126	2211.986976
NSAPYPTTLKYPSDWEDY	2123.9476	708.989809	1062.981076	2124.954876
SAPYPTTLKYPSDWEDY	2009.9047	670.975509	1005.959626	2010.911976
APYPTTLKYPSDWEDY	1922.8727	641.964843	962.443626	1923.879976
PYPTTLKYPSDWEDY	1851.8356	618.285809	926.925076	1852.842876
YPTTLKYPSDWEDY	1754.7828	585.934876	878.398676	1755.790076
AC-YPTTLKYPSDWEDY	1796.7934	599.938409	899.403976	1797.800676

Supplementary Table 4 – Predicted ions for the precursor peptide MvdE without ester bonds.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
MSKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5387.6601	1796.893976	2694.837326	5388.667376
SKNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5256.6197	1753.213843	2629.317126	5257.626976
KNVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5169.5876	1724.203143	2585.801076	5170.594876
NVKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	5041.4927	1681.504843	2521.753626	5042.499976
VKVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4927.4497	1643.490509	2464.732126	4928.456976
KVSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4828.3813	1610.467709	2415.197926	4829.388576
VSAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4700.2864	1567.769409	2351.150476	4701.293676
SAPKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4601.2179	1534.746576	2301.616226	4602.225176
APKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4514.1859	1505.735909	2258.100226	4515.193176
PKAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4443.1488	1482.056876	2222.581676	4444.156076
KAVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4346.096	1449.705943	2174.055276	4347.103276
AVPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4218.0011	1407.007643	2110.007826	4219.008376
VPFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4146.9640	1383.328609	2074.489276	4147.971276
PFFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	4047.8955	1350.305776	2024.955026	4048.902776
FFARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3950.8428	1317.954876	1976.428676	3951.850076
FARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3803.7744	1268.932076	1902.894476	3804.781676
ARFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3656.7060	1219.909276	1829.360276	3657.713276
RFLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3585.6688	1196.230209	1793.841676	3586.676076
FLAEQAVEANNSNSAPYPTTLKYPSDWEDY	3429.5677	1144.196509	1715.791126	3430.574976
LAEQAVEANNSNSAPYPTTLKYPSDWEDY	3282.4993	1095.173709	1642.256926	3283.506576
AEQAVEANNSNSAPYPTTLKYPSDWEDY	3169.4153	1057.479043	1585.714926	3170.422576
EQAVEANNSNSAPYPTTLKYPSDWEDY	3098.3781	1033.799976	1550.196326	3099.385376
QAVEANNSNSAPYPTTLKYPSDWEDY	2969.3355	990.785776	1485.675026	2970.342776
AVEANNSNSAPYPTTLKYPSDWEDY	2841.2770	948.099609	1421.645776	2842.284276

Supplementary Table 4 – Continuation.

MvdE sequence	Exact mass	M+3H	M+2H	M+H
VEANNSNSAPYPTTLKYPSDWEDY	2770.2399	924.420576	1386.127226	2771.247176
EANNSNSAPYPTTLKYPSDWEDY	2671.1714	891.397743	1336.592976	2672.178676
ANNSNSAPYPTTLKYPSDWEDY	2542.1288	848.383543	1272.071676	2543.136076
NNSNSAPYPTTLKYPSDWEDY	2471.0917	824.704509	1236.553126	2472.098976
NSNSAPYPTTLKYPSDWEDY	2357.0488	786.690209	1179.531676	2358.056076
SNSAPYPTTLKYPSDWEDY	2243.0059	748.675909	1122.510226	2244.013176
NSAPYPTTLKYPSDWEDY	2155.9739	719.665243	1078.994226	2156.981176
SAPYPTTLKYPSDWEDY	2041.9309	681.650909	1021.972726	2042.938176
APYPTTLKYPSDWEDY	1954.8989	652.640243	978.456726	1955.906176
PYPTTLKYPSDWEDY	1883.8618	628.961209	942.938176	1884.869076
YPTTLKYPSDWEDY	1786.8090	596.610276	894.411776	1787.816276
AC-YPTTLKYPSDWEDY	1828.8196	610.613809	915.417076	1829.826876

Supplementary Table 5 – Predicted ions for the precursor peptide MvdF with two ester bonds.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
MSKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5631.6457	1878.222509	2816.830126	5632.652976
SKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5500.6052	1834.542343	2751.309876	5501.612476
KNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5413.5732	1805.531676	2707.793876	5414.580476
NIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5285.4782	1762.833343	2643.746376	5286.485476
IKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5171.4353	1724.819043	2586.724926	5172.442576
KVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5058.3512	1687.124343	2530.182876	5059.358476
VSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4930.2562	1644.426009	2466.135376	4931.263476
STGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4831.1878	1611.403209	2416.601176	4832.195076
TGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4744.1558	1582.392543	2373.085176	4745.163076
GSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4643.1081	1548.709976	2322.561326	4644.115376
SAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4586.0867	1529.702843	2294.050626	4587.093976
AVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4499.0546	1500.692143	2250.534576	4500.061876
VPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4428.0175	1477.013109	2215.016026	4429.024776
PFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4328.9491	1443.990309	2165.481826	4329.956376
FFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4231.8963	1411.639376	2116.955426	4232.903576
FARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4084.8279	1362.616576	2043.421226	4085.835176
ARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3937.7595	1313.593776	1969.887026	3938.766776
RFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3866.7224	1289.914743	1934.368476	3867.729676
FLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3710.6213	1237.881043	1856.317926	3711.628576
LSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3563.5529	1188.858243	1782.783726	3564.560176
SEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3450.4688	1151.163543	1726.241676	3451.476076
EQDTETGDSTSTDIPTIWTFKWPSDWEDS	3363.4368	1122.152876	1682.725676	3364.444076

Supplementary Table 5 – Continuation.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
QDTETGDSTSTDIPTIWTFKWPSDWEDS	3234.3942	1079.138676	1618.204376	3235.401476
DTETGDSTSTDIPTIWTFKWPSDWEDS	3106.3356	1036.452476	1554.175076	3107.342876
TETGDSTSTDIPTIWTFKWPSDWEDS	2991.3087	998.110176	1496.661626	2992.315976
ETGDSTSTDIPTIWTFKWPSDWEDS	2890.2610	964.427609	1446.137776	2891.268276
TGDSTSTDIPTIWTFKWPSDWEDS	2761.2184	921.413409	1381.616476	2762.225676
GDSTSTDIPTIWTFKWPSDWEDS	2660.1707	887.730843	1331.092626	2661.177976
DSTSTDIPTIWTFKWPSDWEDS	2603.1492	868.723676	1302.581876	2604.156476
STSTDIPTIWTFKWPSDWEDS	2488.1223	830.381376	1245.068426	2489.129576
TSTDIPTIWTFKWPSDWEDS	2401.0903	801.370709	1201.552426	2402.097576
STDIPTIWTFKWPSDWEDS	2300.0426	767.688143	1151.028576	2301.049876
TDIPTIWTFKWPSDWEDS	2213.0106	738.677476	1107.512576	2214.017876
DIPTIWTFKWPSDWEDS	2111.9629	704.994909	1056.988726	2112.970176
IPTIWTFKWPSDWEDS	1996.9359	666.652576	999.475226	1997.943176
PTIWTFKWPSDWEDS	1883.8519	628.957909	942.933226	1884.859176
TIWTFKWPSDWEDS	1786.7991	596.606976	894.406826	1787.806376
AC-TIWTFKWPSDWEDS	1828.8097	610.610509	915.412126	1829.816976

Supplementary Table 6 – Predicted ions for the precursor peptide MvdF with one ester bond between tyrosine and aspartate.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
MSKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5663.6719	1888.897909	2832.843226	5664.679176
SKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5532.6314	1845.217743	2767.322976	5533.638676
KNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5445.5994	1816.207076	2723.806976	5446.606676
NIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5317.5044	1773.508743	2659.759476	5318.511676
IKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5203.4615	1735.494443	2602.738026	5204.468776
KVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5090.3774	1697.799743	2546.195976	5091.384676
VSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4962.2825	1655.101443	2482.148526	4963.289776
STGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4863.2140	1622.078609	2432.614276	4864.221276
TGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4776.1820	1593.067943	2389.098276	4777.189276
GSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4675.1343	1559.385376	2338.574426	4676.141576
SAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4618.1129	1540.378243	2310.063726	4619.120176
AVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4531.0808	1511.367543	2266.547676	4532.088076
VPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4460.0437	1487.688509	2231.029126	4461.050976
PFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4360.9753	1454.665709	2181.494926	4361.982576
FFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4263.9226	1422.314809	2132.968576	4264.929876
FARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4116.8541	1373.291976	2059.434326	4117.861376
ARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3969.7857	1324.269176	1985.900126	3970.792976

Supplementary Table 6 – Continuation.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
RFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3898.7486	1300.590143	1950.381576	3899.755876
FLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3742.6475	1248.556443	1872.331026	3743.654776
LSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3595.5791	1199.533643	1798.796826	3596.586376
SEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3482.4950	1161.838943	1742.254776	3483.502276
EQDTETGDSTSTDIPTIWTFKWPSDWEDS	3395.4630	1132.828276	1698.738776	3396.470276
QDTETGDSTSTDIPTIWTFKWPSDWEDS	3266.4204	1089.814076	1634.217476	3267.427676
DTETGDSTSTDIPTIWTFKWPSDWEDS	3138.3618	1047.127876	1570.188176	3139.369076
TETGDSTSTDIPTIWTFKWPSDWEDS	3023.3349	1008.785576	1512.674726	3024.342176
ETGDSTSTDIPTIWTFKWPSDWEDS	2922.2872	975.103009	1462.150876	2923.294476
TGDSTSTDIPTIWTFKWPSDWEDS	2793.2446	932.088809	1397.629576	2794.251876
GDSTSTDIPTIWTFKWPSDWEDS	2692.1969	898.406243	1347.105726	2693.204176
DSTSTDIPTIWTFKWPSDWEDS	2635.1755	879.399109	1318.595026	2636.182776
STSTDIPTIWTFKWPSDWEDS	2520.1485	841.056776	1261.081526	2521.155776
TSTDIPTIWTFKWPSDWEDS	2433.1165	812.046109	1217.565526	2434.123776
STDIPTIWTFKWPSDWEDS	2332.0688	778.363543	1167.041676	2333.076076
TDIPTIWTFKWPSDWEDS	2245.0368	749.352876	1123.525676	2246.044076
DIPTIWTFKWPSDWEDS	2143.9891	715.670309	1073.001826	2144.996376
IPTIWTFKWPSDWEDS	2028.9622	677.328009	1015.488376	2029.969476
PTIWTFKWPSDWEDS	1915.8781	639.633309	958.946326	1916.885376
TIWTFKWPSDWEDS	1818.8253	607.282376	910.419926	1819.832576
AC-TIWTFKWPSDWEDS	1860.8359	621.285909	931.425226	1861.843176

Supplementary Table 7 – Predicted ions for the precursor peptide MvdF with one ester bond between serine and glutamate.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
MSKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5649.6562	1884.226009	2825.835376	5650.663476
SKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5518.6158	1822.150490	2760.315176	5519.623076
KNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5431.5837	1793.429897	2716.799126	5432.590976
NIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5303.4888	1751.158580	2652.751676	5304.496076
IKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5189.4458	1713.524390	2595.730176	5190.453076
KVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5076.3618	1676.206670	2539.188176	5077.369076
VSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4948.2668	1633.935320	2475.140676	4949.274076
STGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4849.1984	1601.242748	2425.606476	4850.205676
TGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4762.1664	1572.522188	2382.090476	4763.173676
GSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4661.1187	1539.176447	2331.566626	4662.125976
SAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4604.0972	1520.359352	2303.055876	4605.104476
AVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4517.0652	1491.638792	2259.539876	4518.072476

Supplementary Table 7 – Continuation.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
VPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4446.0281	1468.196549	2224.021326	4447.035376
PFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4346.9597	1435.503977	2174.487126	4347.966976
FFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4249.9069	1403.476553	2125.960726	4250.914176
FARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4102.8385	1354.943981	2052.426526	4103.845776
ARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3955.7701	1306.411409	1978.892326	3956.777376
RFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3884.7330	1282.969166	1943.373776	3885.740276
FLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3728.6318	1231.455770	1865.323176	3729.639076
LSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3581.5634	1182.923198	1791.788976	3582.570676
SEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3468.4794	1145.605478	1735.246976	3469.486676
EQDTETGDSTSTDIPTIWTFKWPSDWEDS	3381.4473	1116.884885	1691.730926	3382.454576
QDTETGDSTSTDIPTIWTFKWPSDWEDS	3252.4047	1074.300827	1627.209626	3253.411976
DTETGDSTSTDIPTIWTFKWPSDWEDS	3124.3462	1032.041522	1563.180376	3125.353476
TETGDSTSTDIPTIWTFKWPSDWEDS	3009.3192	994.082612	1505.666876	3010.326476
ETGDSTSTDIPTIWTFKWPSDWEDS	2908.2715	960.736871	1455.143026	2909.278776
TGDSTSTDIPTIWTFKWPSDWEDS	2779.2290	918.152846	1390.621776	2780.236276
GDSTSTDIPTIWTFKWPSDWEDS	2678.1813	884.807105	1340.097926	2679.188576
DSTSTDIPTIWTFKWPSDWEDS	2621.1598	865.990010	1311.587176	2622.167076
STSTDIPTIWTFKWPSDWEDS	2506.1329	828.031133	1254.073726	2507.140176
TSTDIPTIWTFKWPSDWEDS	2419.1008	799.310540	1210.557676	2420.108076
STDIPTIWTFKWPSDWEDS	2318.0532	765.964832	1160.033876	2319.060476
TDIPTIWTFKWPSDWEDS	2231.0211	737.244239	1116.517826	2232.028376
DIPTIWTFKWPSDWEDS	2129.9735	703.898531	1065.994026	2130.980776
IPTIWTFKWPSDWEDS	2014.9465	665.939621	1008.480526	2015.953776
PTIWTFKWPSDWEDS	1901.8625	628.621901	951.938526	1902.869776
TIWTFKWPSDWEDS	1804.8097	596.594477	903.412126	1805.816976
AC-TIWTFKWPSDWEDS	1846.8203	610.457975	924.417426	1847.827576

Supplementary Table 8 – Predicted ions for the precursor peptide MvdF without ester bonds.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
MSKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5681.6825	1894.901443	2841.848526	5682.689776
SKNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5550.6420	1851.221276	2776.328276	5551.649276
KNIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5463.6099	1822.210576	2732.812226	5464.617176
NIKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5335.5150	1779.512276	2668.764776	5336.522276
IKVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5221.4720	1741.497943	2611.743276	5222.479276
KVSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	5108.3880	1703.803276	2555.201276	5109.395276
VSTGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4980.2930	1661.104943	2491.153776	4981.300276
STGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4881.2246	1628.082143	2441.619576	4882.231876

Supplementary Table 8 – Continuation.

MvdF sequence	Exact mass	M+3H	M+2H	M+H
TGSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4794.1926	1599.071476	2398.103576	4795.199876
GSAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4693.1449	1565.388909	2347.579726	4694.152176
SAVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4636.1234	1546.381743	2319.068976	4637.130676
AVPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4549.0914	1517.371076	2275.552976	4550.098676
VPFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4478.0543	1493.692043	2240.034426	4479.061576
PFFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4378.9859	1460.669243	2190.500226	4379.993176
FFARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4281.9331	1428.318309	2141.973826	4282.940376
FARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	4134.8647	1379.295509	2068.439626	4135.871976
ARFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3987.7963	1330.272709	1994.905426	3988.803576
RFLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3916.7592	1306.593676	1959.386876	3917.766476
FLSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3760.6581	1254.559976	1881.336326	3761.665376
LSEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3613.5896	1205.537143	1807.802076	3614.596876
SEQDTETGDSTSTDIPTIWTFKWPSDWEDS	3500.5056	1167.842476	1751.260076	3501.512876
EQDTETGDSTSTDIPTIWTFKWPSDWEDS	3413.4736	1138.831809	1707.744076	3414.480876
QDTETGDSTSTDIPTIWTFKWPSDWEDS	3284.4310	1095.817609	1643.222776	3285.438276
DTETGDSTSTDIPTIWTFKWPSDWEDS	3156.3724	1053.131409	1579.193476	3157.379676
TETGDSTSTDIPTIWTFKWPSDWEDS	3041.3454	1014.789076	1521.679976	3042.352676
ETGDSTSTDIPTIWTFKWPSDWEDS	2940.2978	981.106543	1471.156176	2941.305076
TGDSTSTDIPTIWTFKWPSDWEDS	2811.2552	938.092343	1406.634876	2812.262476
GDSTSTDIPTIWTFKWPSDWEDS	2710.2075	904.409776	1356.111026	2711.214776
DSTSTDIPTIWTFKWPSDWEDS	2653.1860	885.402609	1327.600276	2654.193276
STSTDIPTIWTFKWPSDWEDS	2538.1591	847.060309	1270.086826	2539.166376
TSTDIPTIWTFKWPSDWEDS	2451.1271	818.049643	1226.570826	2452.134376
STDIPTIWTFKWPSDWEDS	2350.0794	784.367076	1176.046976	2351.086676
TDIPTIWTFKWPSDWEDS	2263.0474	755.356409	1132.530976	2264.054676
DIPTIWTFKWPSDWEDS	2161.9997	721.673843	1082.007126	2163.006976
IPTIWTFKWPSDWEDS	2046.9727	683.331509	1024.493626	2047.979976
PTIWTFKWPSDWEDS	1933.8887	645.636843	967.951626	1934.895976
TIWTFKWPSDWEDS	1836.8359	613.285909	919.425226	1837.843176
AC-TIWTFKWPSDWEDS	1878.8465	627.289443	940.430526	1879.853776