Eusebio et al. BMC Genomics (2021) 22:633
https://doi.org/10.1186/s12864-021-07939-x

BMC Genomics

RESEARCH Open Access

Distribution and diversity of dimetal-carboxylate halogenases in cyanobacteria

Nadia Eusebio[1], Adriana Rego[1], Nathaniel R. Glasser[2], Raquel Castelo-Branco[1], Emily P. Balskus[2*] and Pedro N. Leão[1*]

Abstract

Background: Halogenation is a recurring feature in natural products, especially those from marine organisms. The selectivity with which halogenating enzymes act on their substrates renders halogenases interesting targets for biocatalyst development. Recently, CylC – the first predicted dimetal-carboxylate halogenase to be characterized – was shown to regio- and stereoselectively install a chlorine atom onto an unactivated carbon center during cylindrocyclophane biosynthesis. Homologs of CylC are also found in other characterized cyanobacterial secondary metabolite biosynthetic gene clusters. Due to its novelty in biological catalysis, selectivity and ability to perform C-H activation, this halogenase class is of considerable fundamental and applied interest. The study of CylC-like enzymes will provide insights into substrate scope, mechanism and catalytic partners, and will also enable engineering these biocatalysts for similar or additional C-H activating functions. Still, little is known regarding the diversity and distribution of these enzymes.

Results: In this study, we used both genome mining and PCR-based screening to explore the genetic diversity of CylC homologs and their distribution in bacteria. While we found non-cyanobacterial homologs of these enzymes to be rare, we identified a large number of genes encoding CylC-like enzymes in publicly available cyanobacterial genomes and in our in-house culture collection of cyanobacteria. Genes encoding CylC homologs are widely distributed throughout the cyanobacterial tree of life, within biosynthetic gene clusters of distinct architectures (combination of unique gene groups). These enzymes are found in a variety of biosynthetic contexts, which include fatty-acid activating enzymes, type I or type III polyketide synthases, dialkylresorcinol-generating enzymes, monooxygenases or Rieske proteins. Our study also reveals that dimetal-carboxylate halogenases are among the most abundant types of halogenating enzymes in the phylum Cyanobacteria.

Conclusions: Our data show that dimetal-carboxylate halogenases are widely distributed throughout the Cyanobacteria phylum and that BGCs encoding CylC homologs are diverse and mostly uncharacterized. This work will help guide the search for new halogenating biocatalysts and natural product scaffolds.

Keywords: Halogenases, Cyanobacteria, Natural products, Biocatalysis

* Correspondence: balskus@chemistry.harvard.edu; pleao@ciimar.up.pt
[2]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA
[1]Interdisciplinary Centre of Marine and Environmental Research (CIIMAR/CIMAR), University of Porto, Matosinhos, Portugal

Introduction

Nature is a rich source of new compounds that fuel innovation in the pharmaceutical and agriculture sectors [1]. The remarkable diversity of natural products (NPs) results from a similarly diverse pool of biosynthetic enzymes [2]. These often are highly selective and efficient, carrying out demanding reactions in aqueous media, and therefore are interesting starting points for the development of industrially relevant biocatalysts [2]. Faster and more accessible DNA sequencing technologies have enabled, in the past decade, a large number of genomics and metagenomics projects focused on the microbial world [3]. The resulting sequence data holds immense opportunities for the discovery of new microbial enzymes and their associated NPs [4].

Halogenation is a widely used and well-established reaction in synthetic and industrial chemistry [5], which can have significant consequences for the bioactivity, bioavailability and metabolic activity of a compound [5–7]. Halogenating biocatalysts are thus highly desirable for biotechnological purposes [6, 8]. The mechanistic aspects of biological halogenation can also inspire the development of organometallic catalysts [9]. Nature has evolved multiple strategies to incorporate halogen atoms into small molecules [6], as illustrated by the structural diversity of thousands of currently known halogenated NPs, which include drugs and agrochemicals [10, 11]. Until the early 1990's, haloperoxidases were the only known halogenating enzymes. Research on the biosynthesis of halogenated metabolites eventually revealed a more diverse range of halogenases with different mechanisms. Currently, biological halogenation is known to proceed by distinct electrophilic, nucleophilic or radical mechanisms [6]. Electrophilic halogenation is characteristic of the flavin-dependent halogenases and the heme- and vanadium-dependent haloperoxidases, which catalyze the installation of C-I, C-Br or C-Cl bonds onto electron-rich substrates. Two families of nucleophilic halogenases are known, the halide methyltransferases and SAM halogenases. Both utilize *S*-adenosylmethionine (SAM) as an electrophilic co-factor or as a co-substrate and halide anions as nucleophiles. Notably, these are the only halogenases capable of generating C-F bonds [12]. Finally, radical halogenation has only been described for nonheme-iron/2-oxo-glutarate (2OG)-dependent enzymes. This type of halogenation allows the selective insertion of a halogen into a non-activated, aliphatic C-H bond. A recent review by Agarwal et al. (2017) thoroughly covers the topic of enzymatic halogenation.

Cyanobacteria are a rich source of halogenases among bacteria, in particular for nonheme iron/2OG-dependent and flavin-dependent halogenases (Fig. 1a and b). AmbO5 and WelO5 are cyanobacterial enzymes that belong to the nonheme iron/2OG-dependent halogenase family [13–15]. AmbO5 is an aliphatic halogenase capable of site-selectively modifying ambiguine, fischerindole and hapalindole alkaloids [13, 14]. The close homolog (79 % sequence identity) WelO5 is capable of performing analogous halogenations in hapalindole-type alkaloids and it is involved in the biosynthesis of welwintindolinone [14, 16]. BarB1 and BarB2 are also nonheme iron/2OG-dependent halogenases that catalyze trichlorination of a methyl group from a leucine substrate attached to the peptidyl carrier protein BarA in the biosynthesis of barbamide [17–19]. Other halogenases from this enzyme family include JamE, CurA, and HctB. JamE and CurA catalyse halogenations in intermediate steps of the biosynthesis of jamaicamide and curacin A, respectively [20, 21], while HctB is a fatty acid halogenase responsible for chlorination in hectochlorin assembly [22]. Flavin-dependent halogenases – for which characterized examples include the tryptophan-7-chlorinase PrnA from *Pseudomonas fluorescens* [23], the phenol brominase Bmp5 from members of the bacterial genera *Marinomonas* and *Pseudoalteromonas* [24, 25] and the recently described chlorinase AoiQ which functionalizes the sp^3-hybridized carbon atoms of 1,3-diketones [26] – have also been reported in cyanobacterial biosynthetic pathways. While, to our knowledge, a cyanobacterial flavin-dependent halogenase has not been characterized *in vitro*, ApdC and McnD are predicted to be FAD-dependent enzymes responsible for the halogenation of tyrosine in cyanopeptolin-type peptides. It has been proposed that these enzymes chlorinate, respectively, anabaenopeptilides in *Anabaena* and micropeptins in *Microcystis* strains [27–30]. AerJ is another example of a predicted cyanobacterial FAD-dependent chlorinase, likely involved in the modification of a tyrosine-derived moiety in aeruginosin biosynthesis in *Planktothrix* and *Microcystis* strains [29].

Recent efforts to characterize the biosynthesis of structurally unusual cyanobacterial natural products have uncovered a distinct class of halogenating enzymes (Fig. 1c). Using a genome mining approach, Nakamura et al. (2012) discovered the biosynthetic gene cluster (BGC) responsible for the biosynthesis of the cylindrocyclophanes in the cyanobacterium *Cylindrospermum licheniforme* ATCC 29412 [31]. These natural paracyclophanes were found to be assembled from two identical chlorinated alkylresorcinol units [32]. The paracyclophane macrocycle is created by forming two C-C bonds using a Friedel–Crafts-like alkylation reaction catalyzed by the enzyme CylK [32] (Fig. 1c). Therefore, although many cylindrocyclophanes are not halogenated, their biosynthesis involves a halogenated intermediate [31, 32], a process termed a cryptic halogenation [33]. Nakamura et al. (2017) showed that the CylC enzyme was

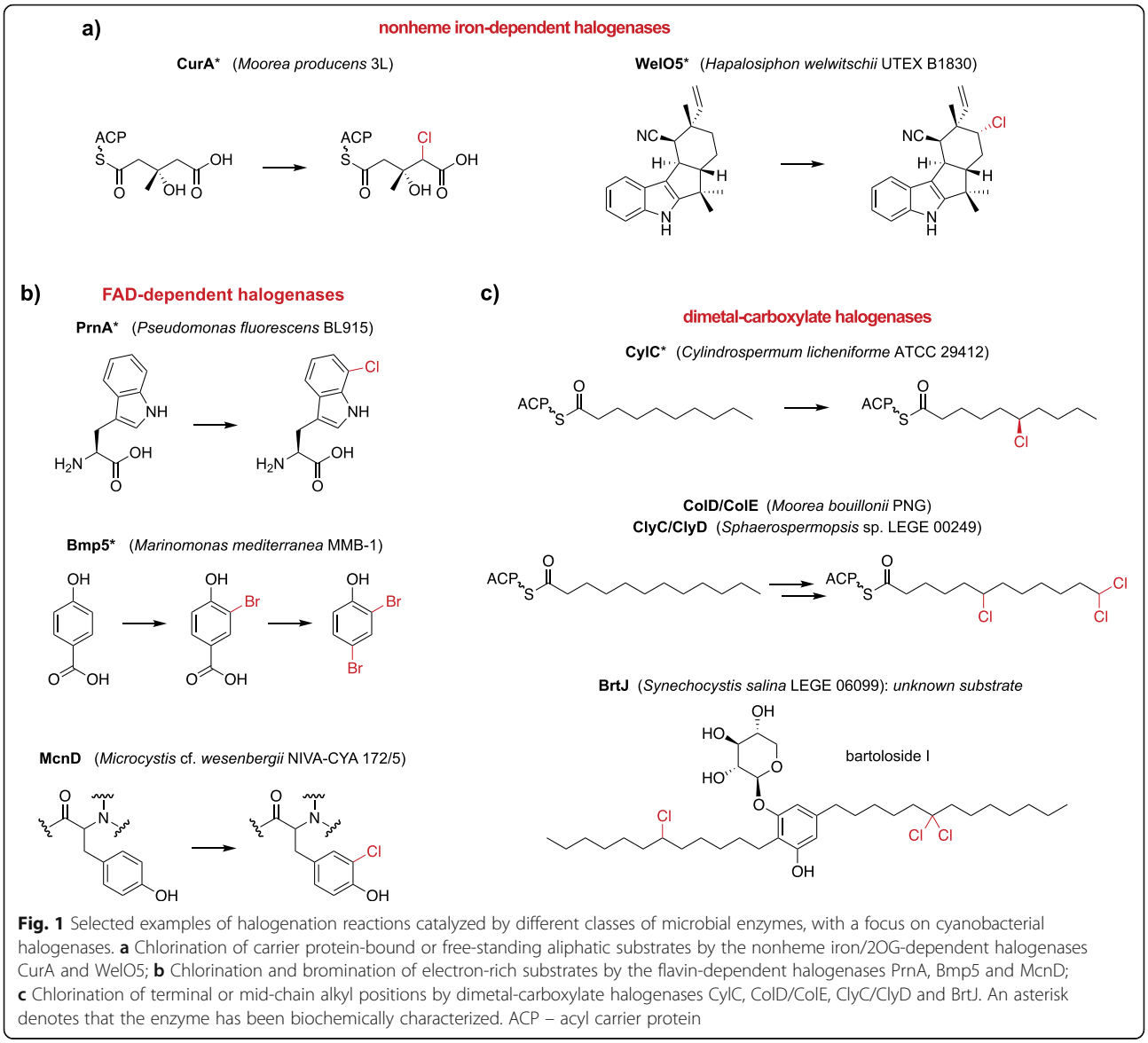

Fig. 1 Selected examples of halogenation reactions catalyzed by different classes of microbial enzymes, with a focus on cyanobacterial halogenases. **a** Chlorination of carrier protein-bound or free-standing aliphatic substrates by the nonheme iron/2OG-dependent halogenases CurA and WelO5; **b** Chlorination and bromination of electron-rich substrates by the flavin-dependent halogenases PrnA, Bmp5 and McnD; **c** Chlorination of terminal or mid-chain alkyl positions by dimetal-carboxylate halogenases CylC, ColD/ColE, ClyC/ClyD and BrtJ. An asterisk denotes that the enzyme has been biochemically characterized. ACP – acyl carrier protein

responsible for regio- and stereoselectively installing a chlorine atom onto the fatty acid-derived sp^3 carbon center of a biosynthetic intermediate that is subsequently elaborated to the key alkylresorcinol monomer (Fig. 1c). To date, CylC is the only characterized dimetal-carboxylate halogenase (this classification is based on both biochemical evidence and similarity to other diiron-carboxylate proteins) [32]. Homologs of CylC have been found in the BGCs of the columbamides [34], bartolosides [35], microginin [32], puwainaphycins/minutissamides [36], and chlorosphaerolactylates [37], all of which produce halogenated metabolites. CylC-type enzymes bear low sequence homology to dimetal desaturases and *N*-oxygenases [32], functionalize C-H bonds in aliphatic moieties at either terminal or mid-chain positions, and are likely able to carry out gem-dichlorination [34, 35]. The reactivity displayed by CylC and its homologs is of interest for biocatalysis, in particular because this type of carbon center activation is often inaccessible to organic synthesis [16, 38]. An understanding of the molecular basis for the halogenation of different positions and for chain-length preference will also be of value for biocatalytic applications. Hence, accessing novel variants of CylC enzymes will facilitate the functional characterization of this class of halogenases, mechanistic studies, and biocatalyst development.

Here, we provide an in-depth analysis of the diversity, distribution and context of CylC homologs in microbial genomes. Using both publicly available genomes and our in-house culture collection of cyanobacteria (LEGEcc), we report that CylC enzymes are common in

cyanobacterial genomes, found in numbers comparable to those of flavin-dependent or nonheme iron/2OG-dependent halogenases. We additionally show that CylC homologs are distributed throughout the cyanobacterial phylogeny and are, to a great extent, part of cryptic BGCs with diverse architectures, underlining the potential for NP discovery associated with this new halogenase class.

Methods

Sequence similarity networks and Genomic Neighborhood Diagrams

Sequence similarity networks (SSNs) were generated using the EFI-EST sever, following a "Sequence BLAST (Basic Local Alignment Search Tool)" of CylC (AFV96137) as input [39], using negative log e-values of 2 and 40 for UniProt BLAST retrieval and SSN edge calculation, respectively. This SSN edge calculation cutoff was found to segregate the homologs into different SSN clusters, less stringent cutoff values resulted in a single SSN cluster. The 153 retrieved sequences and the query sequence were then used to generate the SSNs with an alignment score threshold of 42 and a minimum length of 90. The networks were visualized in Cytoscape (v3.80). The full SSN obtained in the previous step was used to generate Genomic Neighborhood Diagrams (GNDs) using the EFI-GNT tool [39]. A Neighborhood Size of 10 was used and the Lower Limit for Co-occurrence was 20 %. The resulting GNDs were visualized in Cytoscape (Fig. 2).

Cyanobacterial strains and growth conditions

Freshwater and marine cyanobacteria strains from Blue Biotechnology and Ecotoxicology Culture Collection (LEGEcc) (CIIMAR, University of Porto) were grown in 50 mL Z8 medium [40] or 50 mL Z8 2.5 % sea salts (Tropic Marin) with vitamin B_{12}, with orbital shaking (~ 200 rpm) under a regimen of 16 h light (25 µmol photons m-2 s -1)/8 h dark at 25 °C.

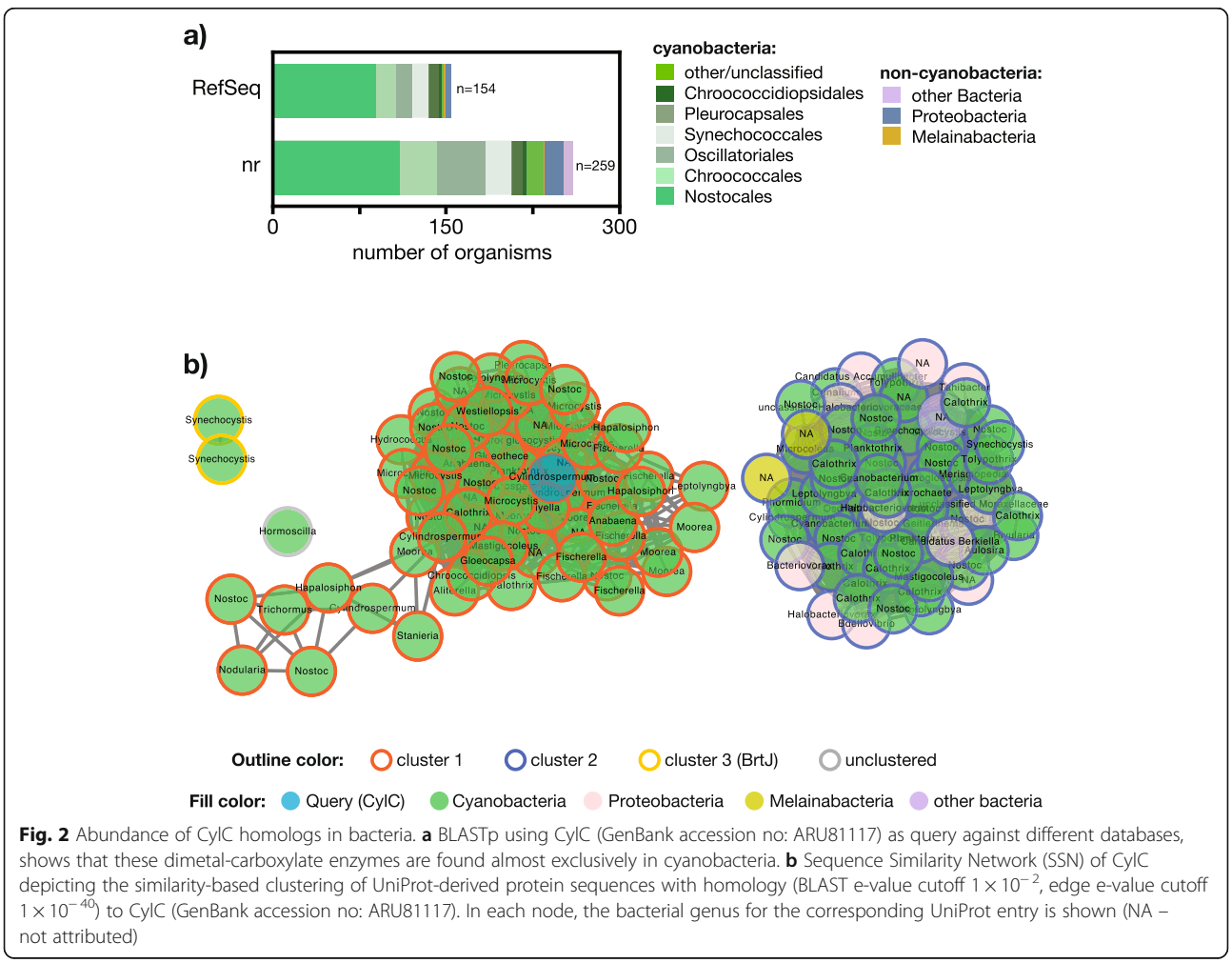

Fig. 2 Abundance of CylC homologs in bacteria. **a** BLASTp using CylC (GenBank accession no: ARU81117) as query against different databases, shows that these dimetal-carboxylate enzymes are found almost exclusively in cyanobacteria. **b** Sequence Similarity Network (SSN) of CylC depicting the similarity-based clustering of UniProt-derived protein sequences with homology (BLAST e-value cutoff 1×10^{-2}, edge e-value cutoff 1×10^{-40}) to CylC (GenBank accession no: ARU81117). In each node, the bacterial genus for the corresponding UniProt entry is shown (NA – not attributed)

Genomic DNA extraction

50 mL of each cyanobacterial strain were centrifuged at 7000 ×*g* for 10 min. The cell pellets were used for genomic DNA (gDNA) extraction using the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific) or NZY Plant/Fungi gDNA Isolation kit (Nzytech), according to the manufacturer's instructions.

Primer design

BLAST searches using CylC [*Cylindrospermum licheniforme* UTEX B 2014] as query identified related genes (for tBLASTn: 31–93 % amino acid identity). We discarded nucleotide hits with a length < 210 and e-values $< 1 \times 10^{-10}$. The complete sequences (56 *cylC* homolog sequences, Table S1) were collected from NCBI and aligned using MUltiple Sequence Comparison by Log-Expectation (MUSCLE) [41]. Phylogenetic analysis of the hits was performed using FastTree GTR with a rate of 100. *Streptomyces thioluteus aurF*, encoding a distant dimetal-carboxylate protein [32] was used as an outgroup (AJ575648.1:4858–5868). We divided the phylogeny of *cylC* homologs in five groups with moderate similarity (Fig. S1). The regions of higher similarity within each group were selected for degenerate primer design (Table 1).

PCR conditions

The PCRs to detect *cylC* homologs were conducted in a final volume of 20 μL, containing 6.9 μL of ultrapure water, 4.0 μL of 5× GoTaq Buffer (Promega), 2.0 μL of $MgCl_2$, 1.0 μL of dNTPs, 2.0 μL of reverse and 2.0 μL of forward primer (each at 10 μM), 0.1 μL of GoTaq and 2.0 μL of cyanobacterial gDNA (20–100 ng/μL). PCR thermocycling conditions were: denaturation for 5 min at 95 °C; 35 cycles with denaturation for 1 min at 95 °C, primer annealing for 30 s at different temperatures (55 °C for group A; 57 °C for group B; 55 °C for group C; 51 °C for group D; 51 °C for group X) and extension for 1 min at 72 °C; and final extension for 10 min at 72 °C.

When not already available, the 16S rRNA gene for a tested strain was amplified by PCR, using standard primers for amplification (CYA106F 5' CGG ACG GGT GAG TAA CGC GTG A 3' and CYA785R 5' GAC TAC WGG GGT ATC TAA TCC 3'). The PCR reactions were conducted in a final volume of 20 μL, containing 6.9 μL of ultrapure water, 4.0 μL of 5× GoTaq Buffer, 2.0 μL of $MgCl_2$, 1.0 μL of dNTPs, 2.0 μL of primer reverse and 2.0 μL of primer forward (each one at 10 μM), 0.1 μL of GoTaq and 2.0 μL of cyanobacterial DNA (5–10 ng/μL). PCR thermocycling conditions were: denaturation for 5 min at 95 °C; 35 cycles with denaturation for 1 min at 95 °C, primer annealing for 30 s at 52 °C and extension for 1 min at 72 °C; and final extension for 10 min at 72 °C.

Amplicon sizes were confirmed after separation in a 1.0 % agarose gel.

Cloning and sequencing

The *cylC* homolog and 16S rRNA gene sequences were obtained either directly from the NCBI or through sequencing. To obtain high quality sequences, TOPO PCR cloning (Invitrogen) was used. The TOPO cloning reaction was conducted in a final volume of 3 μL, containing 1 μL of fresh PCR product, 1 μL of salt solution, 0.5 μL of TOPO vector and 0.5 μL of water. The reaction was incubated for 20 min at room temperature. 3 μL of TOPO reaction were added into a microcentrifuge tube containing chemically competent *E. coli* (Top10, Life Technologies) cells. After 30 min of incubation on ice, the cells were placed for 30 s at 42 °C without shaking and were then immediately transferred to ice. 250 μL of room temperature SOC medium were added to the previous mixture and the tube was horizontally shaken at 37 °C for 1 h (180 rpm). 60 μL of the different cloning reactions were spread onto LB ampicillin/X-gal plates and incubated overnight at 37 °C.

Two or three positive colonies from each reaction were tested by colony-PCR. The PCR was conducted in

Table 1 Degenerate primers

Code	Sequence	Expected amplicon size (bp)	Tm (°C)
AF	CAAAAAATHGCDCTYAAYC	788–986	55
AR	TGDAADCCTTCRTGTTC		
BF	CACAAAAAHTWGCTCTYAAYC	673–715	57
BR	GTKGTRTGGWARGATTCATC		
CF	AATCAWCTTTAYTGGGTRGC	506–509	55
CR	AARAARTGAAARCTYTCRTC		
DF	AATCAAACYAGYGCWGC	299	51
DR	GTRAAATAYTGACAAGC		
XF	ATCWRGAAACCARTSAAGA	449–591	51
XR	CATCAAAAACTTTYYGTARRC		

a final volume of 20 μL, containing 10.9 μL of ultrapure water, 4.0 μL of 5x GoTaq Buffer, 2.0 μL of $MgCl_2$, 1.0 μL of dNTPs, 1.0 μL of reverse pUCR and 1.0 μL of forward pUCF primers (each at 20 μM), 0.1 μL of GoTaq and the target colony. PCR thermocycling conditions were: denaturation for 5 min at 95 °C; 35 cycles with denaturation for 1 min at 95 °C, primer annealing for 30 s at 50 °C and extension for 1 min at 72 °C; and final extension for 10 min at 72 °C. Amplicon sizes were confirmed after separation in an 1.0 % agarose gel. Selected colonies were incubated overnight at 37 °C (180 rpm), in 5 mL of LB supplemented with 100 μg mL^{-1} ampicillin. The plasmids containing the amplified PCR products were extracted (NZYMiniprep) and Sanger sequenced using pUC primers.

Cyanobacteria genome sequencing

Many of the LEGEcc strains are non-axenic, and so before extraction of gDNA for genome sequencing, an evaluation of the amount of heterotrophic contaminant bacteria in cyanobacterial cultures was performed by plating onto Z8 or Z8 with added 2.5 % sea salts (Tropic Marin) and vitamin B_{12} (10 μg/L) agar medium (depending the original environment) supplemented with casamino acids (0.02 % wt/vol) and glucose (0.2 % wt/vol) [42]. The plates were incubated for 2–4 days at 25 °C in the dark and examined for bacterial growth. Those cultures with minimal contamination were used for DNA extraction for genome sequencing. The selection of DNA extraction methodology used was based on morphological features of each strain (coccoid or non-coccoid strains). Total genomic DNA was isolated from a fresh or frozen pellet of 50 mL culture using the commercial PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific) (for coccoid strains) or the NZY Plant/Fungi gDNA Isolation kit (NZYTech) (for non-coccoid strains) or using a CTAB-chloroform/isoamyl alcohol-based protocol [43] (if the previously indicated protocols failed). The latter included a homogenization step (grinding cells using a mortar and pestle with liquid nitrogen) before extraction using the standard kit protocol. The quality of the gDNA was evaluated in a DS-11 FX Spectrophotometer (DeNovix) and 1 % agarose gel electrophoresis, before genome sequencing, which was performed elsewhere (Era7, Spain and MicrobesNG, UK) using 2 × 250 bp paired-end libraries and the Illumina platform (except for *Synechocystis* sp. LEGE 06099, whose genome was sequenced using the Ion Torrent PGM platform). A standard pipeline including the identification of the closest reference genomes for reading mapping using Kraken 2 [44] and BWA-MEM to check the quality of the reads [45] was carried out, while *de novo* assembly was performed using SPAdes [46]. The genomic data obtained for each strain was treated as a metagenome. The contigs obtained as previously mentioned were analyzed using the binning tool MaxBin 2.0 [47] and checked manually in order to obtain only cyanobacterial contigs. The draft genomes were annotated using the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) [48] and submitted to GenBank under the BioProject number PRJNA667061. In the case of *Hyella patelloides* LEGE 07179 and *Sphaerospermopsis* sp. LEGE 00249 the assemblies had been previously deposited in NCBI under the BioSample numbers SAMEA4964519 and SAMN15758549, respectively.

Genomic context of CylC homologs

BLASTp searches using CylC [*Cylindrospermum licheniforme* UTEX B 2014] as query identified related CylC homologs within the publicly available cyanobacterial genomes and in the genomes of LEGEcc strains. We annotated the genomic context for each CylC homolog using antiSMASH v5.0 [49] and manual annotation through BLASTp of selected proteins. Some BGCs were not identified by antiSMASH and were manually annotated using BLASTp searches.

Phylogenetic analysis

Nucleotide sequences of *cylC* homologs obtained from the NCBI and from genome sequencing in this study, were aligned using MUSCLE from within the Geneious R11.0 software package (Biomatters). The nucleotide sequence of the distantly-related dimetal-carboxylate protein AurF [32] from *Streptomyces thioluteus* (AJ575648.1:4858–5868) was used as an outgroup. The alignments, trimmed to their core 788, 673, 506, 299 and 499 positions (for group A, B, C, D and X, respectively), were used for phylogenetic analysis, which was performed using FastTree 2 (from within Geneious), using a GTR substitution model (from jmodeltest, [50]) with a rate of 100 (Fig. S2).

For the phylogenetic analysis based on the 16S rRNA gene (Fig. 3, Fig. S3), the corresponding nucleotide sequences were retrieved from the NCBI (from public available genomes until March 16, 2020) or from sequence data (amplicon or genome) obtained in this study. The sequences were aligned as detailed for *cylC* homologs and trimmed to the core shared positions (656). A RAxML-HPC2 phylogenetic tree inference using maximum likelihood/rapid bootstrapping run on XSEDE (8.2.12) with 1000 bootstrap iterations in the Cipres platform [51] was performed.

The amino acid sequences of CylC homologs were aligned using MUSCLE from within the Geneious software package (Biomatters). The alignments were trimmed to their core 333 residues and used for phylogenetic analysis, which was performed using RAxML-HPC2 phylogenetic tree inference using maximum

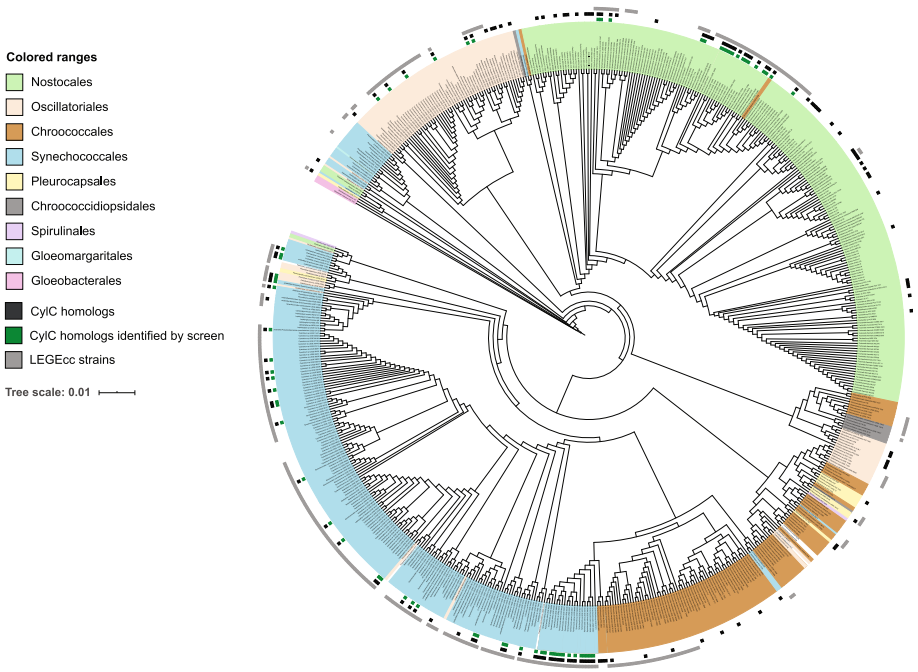

Fig. 3 RAxML cladogram of the 16S rRNA gene of LEGEcc strains (grey squares) and from cyanobacterial strains with NCBI-deposited reference genomes, screened in this study. Taxonomy is presented at the order level (colored rectangles). Strains whose genomes encode CylC homologs are denoted by black squares. Green squares indicate that at least one homolog was detected by PCR-screening and verified by retrieving the sequence of the corresponding amplicon by cloning followed by Sanger sequencing. *Gloeobacter violaceus* PCC 7421 served as an outgroup. A version of this cladogram including the bootstrap values for 1000 replications is provided as Supplementary Material

likelihood/rapid bootstrapping run on XSEDE (8.2.12) with 1000 bootstrap iterations in the Cipres platform [51] (Fig. 4c).

CORASON analysis

CORASON, a bioinformatic tool that computes multilocus phylogenies of BGCs within and across gene cluster families [52], was used to analyze cyanobacterial genomes collected from the NCBI and the LEGEcc genomes (Table S2). In total 2059 cyanobacterial genomes recovered from NCBI and 56 additional LEGE genomes were used in the analysis. The amino acid sequences of CurA (AAT70096.1), WelO5 (AHI58816.1), McnD (CCI20780.1), Bmp5 (WP_008184789.1), PrnA (WP_044451271.1) and CylC (ARU81117.1) were used as query and, for each enzyme, a reference genome was selected (Table S2). To increase the phylogenetic resolution, selected genomes were removed from the analysis of enzymes CylC, PrnA, CurA, McnD and Bmp5 (Table S2). Additionally, for the CylC analysis, a few BGCs were manually extracted and included in the analysis (Table S2) since they were not detected by CORASON.

Prevalence of halogenases in cyanobacterial genomes

Representative proteins of each class were used as query in each search: CylC (ARU81117.1), BrtJ (AKV71855.1), "Mic" (WP_002752271.1) - the halogenase in the putative microginin gene cluster, ColD (AKQ09581.1), ColE (AKQ09582.1), NocO (AKL71648.1), NocN (AKL71647.1) for dimetal-carboxylate halogenases; PrnA (WP_044451271.1), Bmp5 (WP_008184789.1), and McnD (CCI20780.1) for flavin-dependent halogenases; the halogenase domains from CurA (AAT70096.1), and the halogenases Barb1 (AAN32975.1), HctB (AAY42394.1), WelO5 (AHI58816.1) and AmbO5 (AKP23998.1) for nonheme iron-dependent halogenases). Non-redundant sequences obtained for these searches using a 1×10^{-20} e-value cutoff, which represents a percentage identity between the query and target protein superior to 30 %, were considered to share the same function as the query.

Results

CylC-like halogenases are mostly found in cyanobacteria

To investigate the distribution of CylC homologs encoded in microbial genomes, we first searched the reference protein (RefSeq) or non-redundant protein sequences (nr) databases (NCBI) for homologs of CylC or BrtJ, using BLASTp (min 25 % identity, 9.9×10^{-20} E-value and 50 % coverage). A total of 128 and 246 homologous unique protein sequences were retrieved using the RefSeq or nr databases, respectively; in both cases,

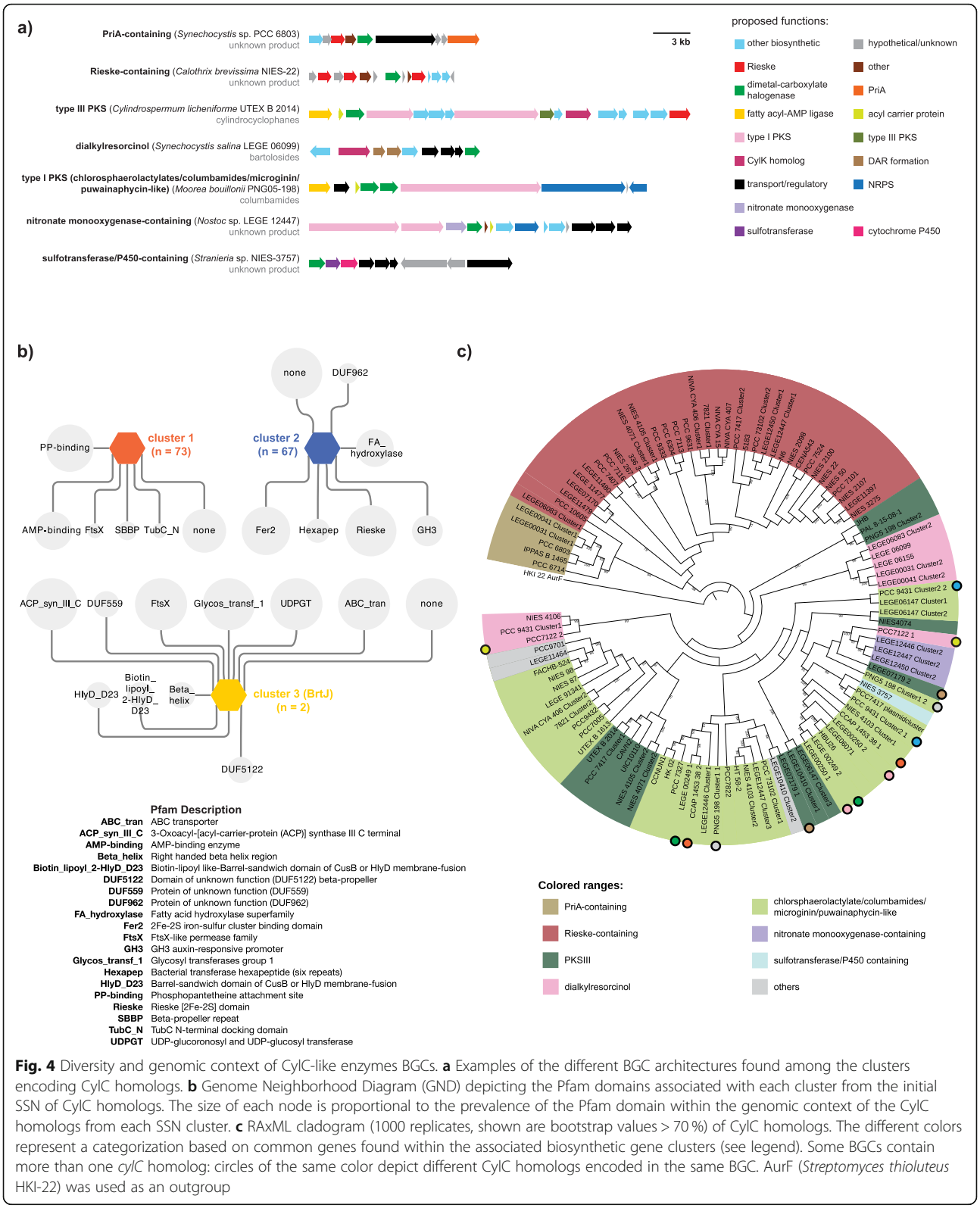

Fig. 4 Diversity and genomic context of CylC-like enzymes BGCs. **a** Examples of the different BGC architectures found among the clusters encoding CylC homologs. **b** Genome Neighborhood Diagram (GND) depicting the Pfam domains associated with each cluster from the initial SSN of CylC homologs. The size of each node is proportional to the prevalence of the Pfam domain within the genomic context of the CylC homologs from each SSN cluster. **c** RAxML cladogram (1000 replicates, shown are bootstrap values > 70 %) of CylC homologs. The different colors represent a categorization based on common genes found within the associated biosynthetic gene clusters (see legend). Some BGCs contain more than one *cylC* homolog: circles of the same color depict different CylC homologs encoded in the same BGC. AurF (*Streptomyces thioluteus* HKI-22) was used as an outgroup

sequences were primarily from cyanobacteria (96 and 88 %, respectively) (Fig. 2a). We then used the Enzyme Similarity Tool of the Enzyme Function Initiative (EFI-EST) [39] to evaluate the sequence landscape of dimetal-carboxylate halogenases. Using CylC as query, we obtained a SSN (sequence similarity network) composed of

154 sequences (153 retrieved from the UniProt database and the query sequence) [53] (Fig. 2b). The SSN featured two major clusters, one containing homologs from diverse cyanobacterial genera, the other composed of homologs from several cyanobacteria, with a few from proteobacteria (mostly deltaproteobacteria) and two from the cyanobacteria sister-phylum Melainabacteria. A third SSN cluster was composed only by the previously reported BrtJ enzymes and, finally, a homolog from the cyanobacterial genus *Hormoscilla* remained unclustered. We were unable to recover any SSN that included clusters containing other characterized enzyme functions.

CylC homologs are widely distributed throughout the phylum Cyanobacteria

With the intent of accessing a wide diversity of CylC homolog sequences, we decided to use a degenerate-primer PCR strategy to discover additional homologs in cyanobacteria from the LEGEcc culture collection [54], because the phylum Cyanobacteria is diverse and still underrepresented in terms of genome data [55–60]. The LEGEcc maintains cultures isolated from diverse freshwater and marine environments, mostly in Portugal, and, for example, contains all known bartoloside-producing strains [35]. We considered that strains with CylC homologs detected by PCR could then be selected for genome sequencing and subsequent recovery of full-length sequences and genomic contexts. To carry out the PCR screen, primers were designed based on 54 nucleotide sequences retrieved from the NCBI (these in turn were selected to represent the phylogenetic diversity of CylC homologs) (Fig. S1). Due to the lack of highly conserved nucleotide sequences among all homologs considered, we divided the nucleotide alignment into five groups and designed a degenerate primer pair for each. Upon screening 326 strains from LEGEcc using the five primer pairs, we retrieved 89 sequences encoding CylC homologs, confirmed through cloning and Sanger sequencing of the obtained amplicons. We were unable to directly analyze the diversity of the entire set of LEGEcc-derived *cylC* amplicons due to low overlap between sequences obtained with different primers. As such, we performed a phylogenetic analysis of the diversity retrieved with each primer pair (Fig. S2), by aligning the PCR-derived sequences with a set of diverse *cylC* genes retrieved from the NCBI. For some strains, our PCR screen retrieved more than one homolog using different primer pairs (e.g. *Nostoc* sp. LEGE 12451 or *Planktothrix mougeotii* LEGE 07231).

To access the full-length sequences of CylC homologs identified among LEGEcc strains, as well as their genomic context, we undertook a genome-sequencing effort informed by our PCR screen. We selected 21 strains for genome sequencing, which represents the diversity of CylC homologs observed in the different PCR screening groups. The resulting genome data was used to generate a local BLAST database and the homologs were located within the genomes. In some cases, additional homologs that were not detected in the PCR screen were identified. Overall, 33 full-length genes encoding CylC homologs were retrieved from LEGEcc strains.

To explore the phylogenetic distribution of CylC homologs encoded in publicly available reference genomes and the herein sequenced LEGEcc genomes, we aligned the 16 S rRNA genes from 648 strains with RefSeq genomes and the LEGEcc strains that were screened by PCR in this study. Using this dataset, we performed a phylogenetic analysis which indicated that CylC homologs are broadly distributed through five Cyanobacterial orders: Nostocales, Oscillatoriales, Chroococcales, Synechococcales and Pleurocapsales (Fig. 3, Fig. S3). We could not detect CylC homologs in the genomes of picocyanobacterial strains (genera *Synechococcus* and *Prochlorococcus*), which are overrepresented among currently available cyanobacterial genomes. It was also noteworthy that the cyanobacterial orders for which we did not find CylC homologs (Chroococcidiopsidales, Spirulinales, Gloeomargaritales and Gloeobacterales) are poorly represented in our dataset (Fig. 3, Fig. S3). However, our previous BLASTp search against the nr database did retrieve two close homologs in a couple of Chroococcidiopsidales strains (genera *Aliterella* and *Chroococcidiopsis*) and a more distant homolog in a *Gloeobacter* strain (Gloeobacterales) (Table S3).

Diversity of BGCs encoding CylC homologs

To characterize the biosynthetic diversity of BGCs encoding CylC homologs, which were found in 78 cyanobacterial genomes (21 from LEGEcc and 57 from RefSeq) from different orders, we first submitted these genome sequences for antiSMASH [49] analysis. 55 CylC-encoding BGCs were detected, which were classified as resorcinol, NRPS, PKS, or hybrid NRPS-PKS. Given the number of CylC homolog-encoding genes detected in these genomes (105), we considered that several BGCs might have not been identified with antiSMASH. Therefore, we performed manual annotation of the genomic contexts of the CylC homologs and were able to identify 40 additional BGCs (i.e. a total of 95 BGCs). Upon analysis of the entire set of CylC-encoding BGCs, we classified the BGCs in seven major categories, based on their overall architecture, which we designated as follows (listed in decreasing abundance): Rieske-containing ($n = 36$), type I PKS (chlorosphaerolactylate/columbamide/microginin/puwainaphycin-like, $n = 29$), type III PKS ($n = 13$), dialkylresorcinol ($n = 8$), PriA-containing ($n = 5$), nitronate monooxygenase-containing ($n = 3$) and cytochrome P450/

sulfotransferase-containing ($n = 1$) (Fig. 4a, Figs. S4, S5, S6, S7, S8, S9 and S10). Three BGCs were excluded from our classification since they were only partially sequenced (Fig. S11). Examples of each of the cluster architectures are presented in Fig. 4a and schematic representations of each of the classified BGCs are presented in Supplementary Figures S4, S5, S6, S7, S8, S9 and S10. It should be stressed that within several of these seven major categories, there is still considerable BGC architecture diversity, notably within the dialkylresorcinol, type I and type III PKS BGCs. Rieske-containing BGCs are not associated with any known NP and encode between two and four proteins with Rieske domains. Most contain a sterol desaturase family protein, feature a single CylC homolog and are chiefly found among Nostocales and Oscillatoriales (Fig. S4). PriA-containing BGCs encode, apart from the Primosomal protein N' (PriA), a set of additional diguanylate cyclase/phosphodiesterase, aromatic ring-hydroxylating dioxygenase subunit alpha and a ferritin-like protein and were only detected in *Synechocystis* spp. (Fig. S5). These are similar to the Rieske-containing BGCs; however, in strains harboring PriA-containing BGCs, the additional functionalities that are found in the Rieske-containing BGCs can be found dispersed throughout the genome (Table S4). In our dataset, a single sulfotransferase/P450 containing BGC was detected in *Stanieria* sp. and was unrelated to the above-mentioned architectures (Fig. S6). Type I PKS BGCs encode clusters similar to those of the chlorosphaerolactylates, columbamides, microginins and puwainaphycins and typically feature a fatty acyl-AMP ligase (FAAL) and an acyl carrier protein upstream of one or two CylC homologs and a type I PKS downstream of the CylC homolog(s). These were found in Nostocales and Oscillatoriales strains (Fig. S7). Taken together with the known NP structures associated with these BGCs [34, 61, 62], we can expect that the encoded metabolites feature halogenated fatty acids in terminal or mid-chain positions. BGCs of the dialkylresorcinol type, which contain DarA and DarB homologs [35, 63], including several bartoloside-like clusters (found only in LEGEcc strains), were detected in Nostocales, Pleurocapsales and Chroococcales (Fig. S8). Type III PKS BGCs encoding CylC homologs, which include a variety of cyclophane BGCs, were detected in the Nostocales, Oscillatoriales and Pleurocapsales (Fig. S9). Finally, nitronate monooxygenase-containing BGCs, which are not associated with any known NP, were only found in Nostocales strains from the LEGEcc and featured also genes encoding PKSI, ferredoxin, ACP or glycosyl transferase (Fig. S10).

A less BGC-centric perspective of the genomic context of CylC homologs could be obtained through the Genome Neighborhood Tool of the EFI (EFI-GNT, [64]). Using the previously generated SSN as input, we analyzed the resulting Genomic Neighborhood Diagrams (Fig. 4b), which indicated that the three SSN clusters had entirely different genomic contexts (herein defined as 10 upstream and 10 downstream genes from the *cylC* homolog). The SSN cluster that encompasses CylC and its closest homologs indicates that these enzymes associate most often with PP-binding (ACP/PCPs) and AMP-binding (such as FAALs) proteins. Regarding the SSN cluster that includes both cyanobacterial and non-cyanobacterial CylC homologs, their genomic contexts most prominently feature Rieske/[2Fe-2S] cluster proteins as well as fatty acid hydroxylase family enzymes. The cyanobacterial homologs in this SSN cluster are exclusively encoded in Rieske or PriA-containing BGCs. Such homologs may not require a phosphopantetheine tethered substrate as no substrate activation or carrier proteins/domains are found in their genomic neighborhoods. However, acyl transferases and desaturases are encoded in most of these BGCs, so we propose these enzymes may act on central fatty acid metabolism intermediates or their derivatives. Finally, the BrtJ SSN cluster, composed only of the two reported BrtJ enzymes, shows entirely different surrounding genes, obviously corresponding to the *brt* genes. Overall, there is a considerable number of proteins with unknown function found in the vicinity of dimetal-carboxylate halogenases, suggesting that uncharted biochemistry is associated with these enzymes.

Since SSN analysis generated only three clusters of CylC homologs, we next investigated the genetic relatedness among these enzymes and how it correlates to BGC architecture. We performed a phylogenetic analysis of the CylC homologs from the 98 classified and 3 unclassified BGCs (Fig. 4c). Our analysis indicated that PriA-containing and Rieske-containing BGCs formed a well-supported clade. Its sister clade contained homologs from the remaining BGCs. Within this larger clade, homologs associated with the type I PKS, dialkylresorcinol or type III PKS BGCs were found to be polyphyletic. In some cases, the same BGC contained distantly related CylC homologs (e.g. *Hyella patelloides* LEGE 07179, *Anabaena cylindrica* PCC 7122) (Fig. 4c). This analysis also revealed that several strains (Fig. 4c) encode two or three phylogenetically distant CylC homologs in different BGCs.

CylC enzymes and other cyanobacterial halogenases

We sought to understand how CylC-type halogenases compare to other halogenating enzyme classes found in cyanobacteria in terms of prevalence and association with BGCs. To this end, we carried out a CORASON [52] analysis of publicly available cyanobacterial genomes

(including non-reference genomes) and the herein acquired genome data from LEGEcc strains (a total of 2,115 cyanobacterial genomes). We used different cyanobacterial halogenases as input, namely CylC, McnD, PrnA, Bmp5, the 2OG-Fe(II) oxygenase domains from CurA and BarB1. CORASON attempts to retrieve genome context by exploring gene cluster diversity linked to enzyme phylogenies [52]. The CORASON analysis retrieved 117 (5.6 %) dimetal-carboxylate halogenases, 61 (2.9 %) nonheme iron-dependent halogenases and 226 (10.7 %) flavin dependent halogenases from the cyanobacterial genomes (Fig. 5a). Using the protein homologs detected in BGCs by CORASON, a sequence alignment was performed for dimetal-carboxylate, nonheme iron/2OG-dependent and flavin-dependent halogenases. For nonheme iron/2OG-dependent halogenases, we excised the halogenase domain from multi-domain enzyme sequences. After removing repeated sequences and trimming the alignments to their core shared positions, maximum-likelihood phylogenetic trees were constructed for each halogenase class and BGCs were annotated manually (Figs. S12, S13 and S14). Flavin-dependent halogenases were commonly associated with cyanopeptolin, 2,4-dibromophenol and pyrrolnitrin BGCs and with orphan BGCs of distinct architectures (Fig. S12). Regarding nonheme iron/2OG-dependent halogenases, we identified barbamide, curacin, hectochlorin and terpene/indole [65] BGCs and several distinct orphan BGCs (Fig. S13). For dimetal-carboxylate halogenases, columbamide, microginin, chlorosphaerolactylate, bartoloside and cyclophane BGCs were identified (Fig. S14). However, while some of the CylC homolog-encoding orphan BGCs previously identified by antiSMASH and manual searches were detected by CORASON, the Rieske- and the PriA-containing BGCs were not. Hence, several CylC homologs were not accounted for in this analysis. For the same reasons, the other two halogenase types could also be missing some of its members in the CORASON-derived datasets. To circumvent this limitation and obtain a more comprehensive picture of the abundance of the three types of halogenase in cyanobacterial genomes, we used BLASTp searches against available cyanobacterial genomes in the NCBI database (including non-reference genomes). Several representatives of each halogenase class were used as query in each search (CylC, BrtJ, "Mic" – the halogenase in the putative microginin gene cluster – ColD, ColE, NocO and NocN for dimetal-carboxylate halogenases; PrnA, Bmp5 and McnD for flavin dependent halogenases; the halogenase domain from CurA and the halogenases BarB1, HctB, WelO5 and AmbO5 for nonheme iron-dependent halogenases). Non-redundant sequences obtained for these searches using a 1×10^{-20} e-value cutoff (corresponding to > 30 % sequence identity) were considered to share the same function as the query. It is worth mentioning that, for nonheme iron/2OG-dependent enzymes, a single amino acid difference can convert hydroxylation activity into halogenation [66], so

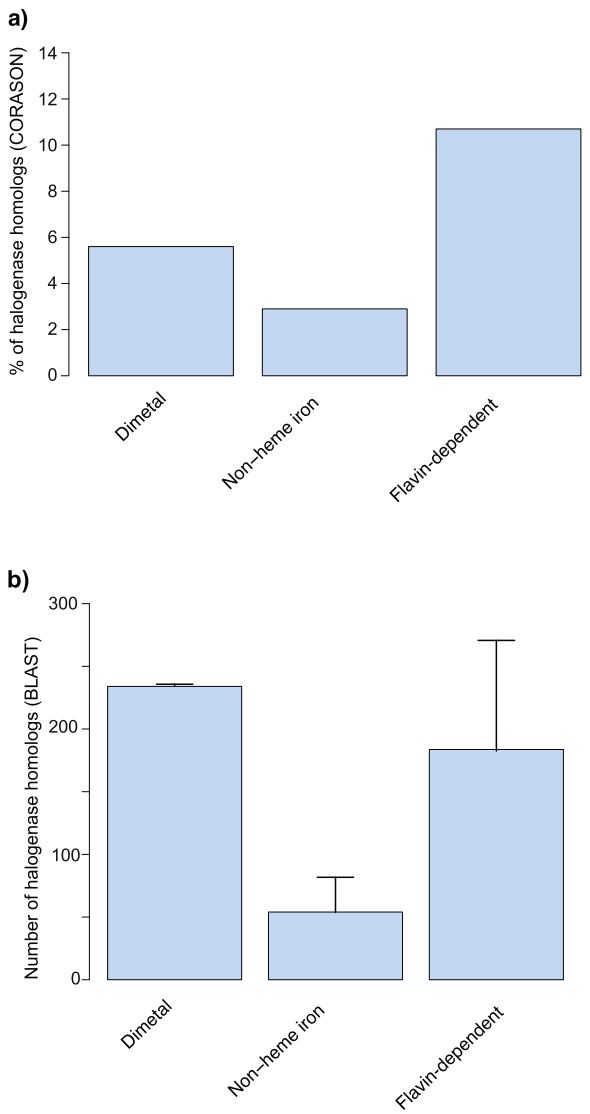

Fig. 5 Prevalence of putative cyanobacterial halogenases. Frequency of halogenase homologs in Cyanobacteria from CORASON analysis (**a**) and NCBI BLASTp analysis (**a**). Dimetal-carboxylate halogenases: CylC - NCBI reference genomes, $n = 2054$ and LEGEcc genomes, $n = 41$ CylC-containing BGCs and 56 genomes; Flavin-dependent halogenases: PrnA - NCBI reference genomes, $n = 2051$ and LEGEcc genomes, $n = 56$ genomes; Bmp5- NCBI reference genomes, $n = 2050$ and LEGEcc genomes, $n = 56$ genomes; McnD: NCBI reference genomes, $n = 2052$ and LEGEcc genomes, $n = 54$ genomes; Nonheme iron/2OG-dependent halogenases: halogenase domain from CurA - NCBI reference genomes, $n = 2052$ and LEGEcc genomes, $n = 56$ genomes. **B** Average of the total number of homologs per dimetal-carboxylate halogenases (CylC, BrtJ, "Mic", ColD, ColE, NocO, NocN), flavin-dependent halogenases (PrnA, Bmp5 and McnD) and nonheme iron/2OG-dependent halogenases (Barb1, HctB, WelO5, AmbO5 and the halogenase domain from CurA)

it is possible that – at least for this class – the sequence space considered does not correspond exclusively to halogenation activity. Dimetal-carboxylate and flavin-dependent halogenase homologs were found to be the most abundant in cyanobacteria, each with roughly 0.2 homologs per genome, while nonheme iron/2OG-dependent halogenase homologs are less common (~ 0.05 per genome) (Fig. 5b).

Discussion

CylC is the single characterized member of the dimetal-carboxylate halogenases. A handful of homologs are encoded in BGCs whose corresponding NPs are known, and their halogenase function can be deduced to some extent from the NP structures. In this study, we show that the remaining homologs, which are mostly found in cyanobacteria, can be used to guide the discovery of new chemistry. In particular, SSN analyses of CylC homologs attests to the uniqueness of these dimetal-carboxylate enzymes in the current protein-sequence landscape, as no homologs with additional functions could be retrieved. CylC homologs therefore represent a region of protein sequence space that is vastly unexplored. Their activity might not be limited to halogenation – like in the case of iron/2OG-dependent enzymes, it is possible that some CylC homologs perform other types of oxidative transformations.

To obtain an expanded representation of CylC homolog sequences, apart from retrieving these from publicly available genomes, we used a PCR-based strategy to screen our in-house culture collection and retrieve additional homologs. To increase effectiveness, we used more than one primer pair. In general, and for each primer pair, the PCR screen retrieved mostly sequences that were closely related and clustered as a single or, at a maximum, two phylogenetic clades. This can likely be explained by the geographical bias that might exist in the LEGEcc culture collection [54] and/or with primer design and PCR efficiency issues, which might have favored certain phylogenetic clades. Overall, the two approaches showed a wide but punctuated presence of CylC homologs among the cyanobacterial diversity considered in this study. In light of this, it is unclear how much of the current CylC homolog distribution reflects vertical inheritance or horizontal gene transfer events.

The phylogenetic analysis and the genomic context analyses that we have carried out for dimetal-carboxylate halogenases show that they have evolved to interact with different partner enzymes to generate chemical diversity, but that their phylogeny is, in some cases, not entirely consistent with BGC architecture. These observations suggest that functionally convergent associations between CylC homologs and other proteins have emerged multiple times during evolution. Examples include the CylC/CylK and BrtJ/BrtB associations, which use cryptic halogenation to achieve C-C and C-O bond formation, respectively [32, 67]. However, the role of the CylC homolog-mediated halogenation of fatty acyl moieties observed for other cyanobacterial metabolites is not currently understood. Interestingly, while a number of CylC homologs, including those that are part of characterized BGCs, likely act on ACP-tethered fatty acyl substrates [32, 67], those from the PriA- Rieske- and cytochrome P450/sulfotransferase categories do not have a neighboring carrier protein and therefore might not require a tethered substrate. This would be an important property for a CylC-like biocatalyst [16].

When comparing dimetal-carboxylate halogenases with the nonheme iron/2OG-dependent and flavin-dependent halogenases, we found that the former are clearly a major group of halogenases in cyanobacteria, despite having been the latest to be discovered [32]. Notwithstanding, homologs of each of the three halogenase classes are associated with a large number of orphan BGCs and all classes represent opportunities for NP discovery.

Conclusions

The discovery of a new biosynthetic enzyme class brings with it tremendous possibilities for biochemistry and catalysis research, both fundamental and applied. Their functional characterization can also be used as a handle to identify and deorphanize BGCs that encode their homologs. CylC typifies an unprecedented halogenase class, which is almost exclusively found in cyanobacteria. By searching CylC homologs in both public databases and our in-house culture collection, we report here more than 100 new cyanobacterial CylC homologs. We found that dimetal-carboxylate halogenases are widely distributed throughout the phylum. The genomic neighborhoods of these halogenases are diverse and we identify a number of different BGC architectures associated with either one or two CylC homologs that can serve as starting points for the discovery of new NP scaffolds. In addition, the herein reported diversity and biosynthetic contexts of these enzymes will serve as a roadmap to further explore their biocatalysis-relevant activities. Despite their prevalence and distribution, there is no strong evidence for a role of CylC-like halogenases in primary metabolism and their diverse genomic contexts suggests otherwise. Finally, bartoloside-like BGCs and another CylC-associated BGC architecture (nitronate monooxygenase-containing) were found only in the LEGEcc, reinforcing the importance of geographically focused strain isolation and maintenance efforts for the Cyanobacteria phylum. However, to fully realize the potential of this new halogenase class, biochemical and structural characterization of additional homologs is

warranted. This will not only provide mechanistic insight into catalysis but also enable sequence-based predictions of carrier-protein requirement and regioselectivity, all of which are open questions for these new enzymes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07939-x.

Additional file 1: Table S1. Accession numbers of *cylC* homologs and *aurF*genes used for primer design. **Figure S1.** (a) Phylogenetic tree (FastTree GTR with a rate of 100) of *cylC* homologs highlighted according to the groups selected for degenerate primer design. (b) Schematic representation of the different pairs of degenerate primers. **Figure S2.** PCR-based detection of *cylC* homologs in the LEGEcc. Five pairs of primers were designed based on conserved regions identified in the *cylC* gene. Each primer pair was used in a PCR screen of the gDNA obtained from diverse strains (n = 326) of the LEGEcc. The resulting amplicons were cloned and sequenced. Sequences for each primer pair were aligned with the corresponding regions of *cylC* genes found in the NCBI reference genomes (cyanobacteria only) and those from LEGEcc strains' genomes. Shown are the resulting cladograms (RaxML, 1000 replicates) for each primer pair used in the screening. Blue squares indicate sequences obtained from the PCR screen. **Figure S3.** RaxML cladogram (1000 replicates) of the 16S rRNA gene of LEGEcc strains (grey squares) and from cyanobacterial strains with NCBI-deposited reference genomes, screened in this study. Taxonomy is presented at the order level (colored ranges). Strains whose genomes encode CylC homologs are denoted by black squares. Green squares indicate that at least one CylC homolog was detected by PCR-screening and verified by retrieving the sequence of the corresponding amplicon through cloning followed by Sanger sequencing. The cladogram topology is the same as shown in Fig. 3 of the main manuscript, but here bootstrap values (equal or above 0.7) are shown. **Table S2.** GenBank or RefSeq assembly acession number and LEGEcc genome used for CORASON analysis. **Table S3.** BLASTp search of CylC homologs against *Aliterella sp.*, *Chroococcidiopsis sp.* and *Gloeobacter sp.* **Figure S4.** Rieske-containing biosynthetic gene clusters encoding CylC homolog(s). **Figure S5.** PriA-containing biosynthetic gene clusters encoding CylC homolog(s). **Figure S6.** Cytochrome P450/sulfotransferase-containing biosynthetic gene cluster encoding a CylC homolog. **Figure S7.** Type I PKS (chlorosphaerolactylate/columbamide/microginin/puwainaphycin-like) biosynthetic gene clusters encoding CylC homolog(s). **Figure S8.** Dialkylresorcinol biosynthetic gene clusters encoding CylC homolog(s). **Figure S9.** Type III PKS biosynthetic gene clusters encoding CylC homolog(s). **Figure S10.** Nitronate monooxygenase-containing biosynthetic gene clusters encoding a CylC homolog. **Figure S11.** Unclassified (likely incomplete) biosynthetic gene clusters encoding a CylC homolog. **Table S4.** BLAST search of Rieske-containing BGCs genes from *Calothrix brevissima* NIES 22 against *Synechocystis* sp. PCC 6803. **Figure S12.** Phylogenetic tree of FAD-dependent halogenases based on CORASON outputs with illustrative BGC architectures. **Figure S13.** Phylogenetic tree of nonheme iron-dependent halogenases based on CORASON outputs with illustrative BGC architectures. **Figure S14.** Phylogenetic tree of dimetal-carboxylate halogenases based on CORASON outputs with illustrative BGC architectures.

Acknowledgements

We thank Hitomi Nakamura, Samantha Cassell, Diana Sousa and João Reis for technical assistance during this study, and the Blue Biotechnology and Ecotoxicology Culture Collection (LEGEcc) for the genomic DNA used for the PCR screening.

Authors' contributions

N.E., E.P.B and P.N.L. designed research, N.E., A.R., N.R.G. and R.C-B. performed experiments and bioinformatics analysis, N.E., N.R.G., E.P.B and P.N.L. wrote the manuscript with contributions from all authors. All authors reviewed the manuscript. The author(s) read and approved the final manuscript.

Funding

This work was funded by Fundação para a Ciência e a Tecnologia (FCT) through grant PTDC/BIA-BQM/29710/2017 to PNL and through strategic funding UIDB/04423/2020, UIDP/04423/2020 and by the National Science Foundation (NSF) through grants CHE-1454007 and CHE-2003436 to EPB. AR and RCB are supported by doctoral grants from FCT (SFRH/BD/140567/2018 and SFRH/BD/136367/2018, respectively). This material is based upon work supported by an NSF Postdoctoral Research Fellowship in Biology (Grant No 1907240 to NRG). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available at the NCBI database ID number PRJNA667061. It can be accessed using the following link: https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA667061.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 18 May 2021 Accepted: 14 August 2021
Published online: 31 August 2021

References

1. Pham JV, Yilma MA, Feliz A, Majid MT, Maffetone N, Walker JR, Kim E, Cho HJ, Reynolds JM, Song MC, et al. A review of the microbial production of bioactive natural products and biologics. Front Microbiol. 2019;10:1404.
2. Noda-Garcia L, Tawfik DS. Enzyme evolution in natural products biosynthesis: target- or diversity-oriented? Curr Opin Chem Biol. 2020;59:147–54.
3. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: history and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J. 2020;18:9–19.
4. Zhang MM, Qiao Y, Ang EL, Zhao H. Using natural products for drug discovery: the impact of the genomics era. Expert Opin Drug Discov. 2017;12(5):475–87.
5. Gkotsi DS, Dhaliwal J, McLachlan MMW, Mulholand KR, Goss RJM. Halogenases: powerful tools for biocatalysis (mechanisms applications and scope). Curr Opin Chem Biol. 2018;43:119–26.
6. Agarwal V, Miles ZD, Winter JM, Eustáquio AS, El Gamal AA, Moore BS. Enzymatic halogenation and dehalogenation reactions: pervasive and mechanistically diverse. Chem Rev. 2017;117(8):5619–74.
7. Weichold V, Milbredt D, van Pée K-H. Specific enzymatic halogenation—from the discovery of halogenated enzymes to their applications in vitro and in vivo. Angew Chem Int Ed. 2016;55(22):6374–89.
8. Schnepel C, Sewald N: Enzymatic halogenation: a timely strategy for regioselective C – H Activation. Chem Eur J. 2017;23(50):12064–86.
9. Petrone DA, Ye J, Lautens M. Modern transition-metal-catalyzed carbon–halogen bond formation. Chem Rev. 2016;116(14):8003–104.
10. Jeschke P. The unique role of halogen substituents in the design of modern agrochemicals. Pest Manage Sci. 2010;66(1):10–27.
11. Xu Z, Yang Z, Liu Y, Lu Y, Chen K, Zhu W. Halogen bond: its role beyond drug–target binding affinity for drug discovery and development. J Chem Inform Model. 2014;54(1):69–78.
12. Wu L, Maglangit F, Deng H. Fluorine biocatalysis. Curr Opin Chem Biol. 2020;55:119–26.
13. Hillwig ML, Zhu Q, Ittiamornkul K, Liu X. Discovery of a promiscuous non-heme iron halogenase in ambiguine alkaloid biogenesis: implication for an evolvable enzyme family for late-stage halogenation of aliphatic carbons in small molecules. Angew Chem Int Ed Engl. 2016;55(19):5780–4.
14. Liu X. In Vitro Analysis of cyanobacterial nonheme iron-dependent aliphatic halogenases WelO5 and AmbO5. Methods Enzymol. 2018;604:389–404.

15. Pratter SM, Ivkovic J, Birner-Gruenberger R, Breinbauer R, Zangger K, Straganz GD. More than just a halogenase: modification of fatty acyl moieties by a trifunctional metal enzyme. Chembiochem. 2014;15(4):567–74.
16. Hillwig ML, Liu X. A new family of iron-dependent halogenases acts on freestanding substrates. Nat Chem Biol. 2014;10(11):921–3.
17. Chang Z, Flatt P, Gerwick WH, Nguyen VA, Willis CL, Sherman DH. The barbamide biosynthetic gene cluster: a novel marine cyanobacterial system of mixed polyketide synthase (PKS)-non-ribosomal peptide synthetase (NRPS) origin involving an unusual trichloroleucyl starter unit. Gene. 2002; 296(1–2):235–47.
18. Flatt PM, O'Connell SJ, McPhail KL, Zeller G, Willis CL, Sherman DH, Gerwick WH. Characterization of the initial enzymatic steps of barbamide biosynthesis. J Nat Prod. 2006;69(6):938–44.
19. Galonić DP, Vaillancourt FH, Walsh CT. Halogenation of unactivated carbon centers in natural product biosynthesis: trichlorination of leucine during barbamide biosynthesis. J Am Chem Soc. 2006;128(12):3900–1.
20. Chang Z, Sitachitta N, Rossi JV, Roberts MA, Flatt PM, Jia J, Sherman DH, Gerwick WH. Biosynthetic pathway and gene cluster analysis of curacin A, an antitubulin natural product from the tropical marine cyanobacterium Lyngbya majuscula. J Nat Prod. 2004;67(8):1356–67.
21. Edwards DJ, Marquez BL, Nogle LM, McPhail K, Goeger DE, Roberts MA, Gerwick WH. Structure and Biosynthesis of the Jamaicamides, New Mixed Polyketide-Peptide Neurotoxins from the Marine Cyanobacterium Lyngbya majuscula. Chem Biol. 2004;11(6):817–33.
22. Ramaswamy AV, Sorrels CM, Gerwick WH. Cloning and biochemical characterization of the hectochlorin biosynthetic gene cluster from the marine cyanobacterium Lyngbya majuscula. J Nat Prod. 2007;70(12):1977–86.
23. Keller S, Wage T, Hohaus K, Hölzer M, Eichhorn E, van Pée K-H. Purification and Partial Characterization of Tryptophan 7-Halogenase (PrnA) from *Pseudomonas fluorescens*. Angew Chem Int Ed. 2000;39(13):2300–2.
24. Agarwal V, El Gamal AA, Yamanaka K, Poth D, Kersten RD, Schorn M, Allen EE, Moore BS. Biosynthesis of polybrominated aromatic organic compounds by marine bacteria. Nat Chem Biol. 2014;10(8):640–7.
25. Agarwal V, Moore BS. Enzymatic synthesis of polybrominated dioxins from the marine environment. ACS Chem Biol. 2014;9(9):1980–4.
26. Liu M, Ohashi M, Hung Y-S, Scherlach K, Watanabe K, Hertweck C, Tang Y. AoiQ Catalyzes Geminal Dichlorination of 1,3-Diketone Natural Products. J Am Chem Soc. 2021;143(19):7267–71.
27. Kocher S, Resch S, Kessenbrock T, Schrapp L, Ehrmann M, Kaiser M. From dolastatin 13 to cyanopeptolins, micropeptins, and lyngbyastatins: the chemical biology of Ahp-cyclodepsipeptides. Nat Prod Rep. 2020;37(2):163–74.
28. Rouhiainen L, Paulin L, Suomalainen S, Hyytiainen H, Buikema W, Haselkorn R, Sivonen K. Genes encoding synthetases of cyclic depsipeptides, anabaenopeptilides, in Anabaena strain 90. Mol Microbiol. 2000;37(1):156–67.
29. Cadel-Six S, Dauga C, Castets AM, Rippka R, Bouchier C, Tandeau de Marsac N, Welker M. Halogenase genes in nonribosomal peptide synthetase gene clusters of Microcystis (cyanobacteria): sporadic distribution and evolution. Mol Biol Evol. 2008;25(9):2031–41.
30. Nishizawa T, Ueda A, Nakano T, Nishizawa A, Miura T, Asayama M, Fujii K, Harada K, Shirai M. Characterization of the locus of genes encoding enzymes producing heptadepsipeptide micropeptin in the unicellular cyanobacterium Microcystis. J Biochem. 2011;149(4):475–85.
31. Nakamura H, Hamer HA, Sirasani G, Balskus EP. Cylindrocyclophane Biosynthesis Involves Functionalization of an Unactivated Carbon Center. J Am Chem Soc. 2012;134(45):18518–21.
32. Nakamura H, Schultz EE, Balskus EP. A new strategy for aromatic ring alkylation in cylindrocyclophane biosynthesis. Nat Chem Biol. 2017;13(8): 916–21.
33. Vaillancourt FH, Yeh E, Vosburg DA, O'Connor SE, Walsh CT. Cryptic chlorination by a non-haem iron enzyme during cyclopropyl amino acid biosynthesis. Nature. 2005;436(7054):1191–4.
34. Kleigrewe K, Almaliti J, Tian IY, Kinnel RB, Korobeynikov A, Monroe EA, Duggan BM, Di Marzo V, Sherman DH, Dorrestein PC, et al. Combining mass spectrometric metabolic profiling with genomic analysis: a powerful approach for discovering natural products from cyanobacteria. J Nat Prod. 2015;78(7):1671–82.
35. Leão PN, Nakamura H, Costa M, Pereira AR, Martins R, Vasconcelos V, Gerwick WH, Balskus EP. Biosynthesis-assisted structural elucidation of the bartolosides, chlorinated aromatic glycolipids from cyanobacteria. Angew Chem Int Ed Engl. 2015;54(38):11063–7.
36. Mareš J, Hájek J, Urajová P, Kust A, Jokela J, Saurav K, Galica T, Čapková K, Mattila A, Haapaniemi E, et al. Alternative biosynthetic starter units enhance the structural diversity of cyanobacterial lipopeptides. Appl Environ Microbiol. 2019;85(4):e02675-02618.
37. Abt K, Castelo-Branco R, Leao PNC: Biosynthesis of Chlorinated Lactylates in Sphaerospermopsis sp. LEGE 00249. 2020.
38. Latham J, Brandenburger E, Shepherd SA, Menon BRK, Micklefield J. Development of Halogenase Enzymes for Use in Synthesis. Chem Rev. 2018; 118(1):232–69.
39. Zallot R, Oberg N, Gerlt JA. The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry. 2019;58(41): 4169–82.
40. Kotai J. Instructions for preparation of modified nutrient solution Z8 for algae. Norwegian Inst Water Res. 1972;11:5.
41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
42. Rippka R, Waterbury JB, Stanier RY. Isolation and purification of cyanobacteria: some general principles. In: Starr MP, Stolp H, Trüper HG, Balows A, Schlegel HG, editors. The prokaryotes: a handbook on habitats, isolation, and identification of bacteria. Berlin, Heidelberg: Springer; 1981. p. 212–20.
43. Singh SP, Rastogi RP, Häder D-P, Sinha RP. An improved method for genomic DNA extraction from cyanobacteria. World J Microbiol Biotechnol. 2011;27(5):1225–30.
44. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15(3):R46.
45. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
46. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
47. Wu YW, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32(4):605–7.
48. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44(14):6614–24.
49. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res. 2019;47(W1):W81–7.
50. Posada D. jModelTest: phylogenetic model averaging. Mol Biol Evol. 2008; 25(7):1253–6.
51. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 Gateway Computing Environments Workshop (GCE). 2010. p. 1–8.
52. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH, Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S, et al. A computational framework to explore large-scale biosynthetic diversity. Nat Chem Bio. 2020;16(1):60–8.
53. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2016;45(D1):D158–69.
54. Ramos V, Morais J, Castelo-Branco R, Pinheiro Â, Martins J, Regueiras A, Pereira AL, Lopes VR, Frazão B, Gomes D, 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. 2018; 30(3):1437–51.
55. Dittmann E, Gugger M, Sivonen K, Fewer DP. Natural product biosynthetic diversity and comparative genomics of the cyanobacteria. Trends Microbiol. 2015;23(10):642–52.
56. D'Agostino PM, Woodhouse JN, Makower AK, Yeung AC, Ongley SE, Micallef ML, Moffitt MC, Neilan BA. Advances in genomics, transcriptomics and proteomics of toxin-producing cyanobacteria. Environ Microbiol Rep. 2016; 8(1):3–13.
57. Calteau A, Fewer DP, Latifi A, Coursin T, Laurent T, Jokela J, Kerfeld CA, Sivonen K, Piel J, Gugger M. Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. BMC Genomics. 2014;15(1):977.

58. Baran R, Ivanova NN, Jose N, Garcia-Pichel F, Kyrpides NC, Gugger M, Northen TR. Functional genomics of novel secondary metabolites from diverse cyanobacteria using untargeted metabolomics. Mar Drugs. 2013; 11(10):3617–31.
59. Alvarenga DO, Fiore MF, Varani AM. A metagenomic approach to Cyanobacterial genomics. Front Microbiol. 2017;8:809–809.
60. Beck C, Knoop H, Axmann IM, Steuer R. The diversity of cyanobacterial metabolism: genome analysis of multiple phototrophic microorganisms. BMC Genomics. 2012;13(1):56.
61. Okino T, Matsuda H, Murakami M, Yamaguchi K. Microginin, an angiotensin-converting enzyme inhibitor from the blue-green alga Microcystis aeruginosa. Tetrahedron Lett. 1993;34(3):501–4.
62. Voráčová K, Hájek J, Mareš J, Urajová P, Kuzma M, Cheel J, Villunger A, Kapuscik A, Bally M, Novák P, et al. The cyanobacterial metabolite nocuolin a is a natural oxadiazine that triggers apoptosis in human cancer cells. PLOS ONE. 2017;12(3):e0172850.
63. Fuchs SW, Bozhüyük KAJ, Kresovic D, Grundmann F, Dill V, Brachmann AO, Waterfield NR, Bode HB. Formation of 1,3-Cyclohexanediones and Resorcinols Catalyzed by a Widely Occuring Ketosynthase. Angew Chem Int Ed. 2013;52(15):4108–12.
64. Zallot R, Oberg NO, Gerlt JA. 'Democratized' genomic enzymology web tools for functional assignment. Curr Opin Chem Biol. 2018;47:77–85.
65. Liu Y, Klet RC, Hupp JT, Farha O. Probing the correlations between the defects in metal-organic frameworks and their catalytic activity by an epoxide ring-opening reaction. Chem Commun (Camb). 2016;52(50):7806–9.
66. Mitchell AJ, Dunham NP, Bergman JA, Wang B, Zhu Q, Chang W-C, Liu X, Boal AK. Structure-guided reprogramming of a hydroxylase to halogenate its small molecule substrate. Biochemistry. 2017;56(3):441–4.
67. Reis JPA, Figueiredo SAC, Sousa ML, Leão PN. BrtB is an O-alkylating enzyme that generates fatty acid-bartoloside esters. Nat Commun. 2020;11(1):1458–1458.

Publisher's Note