bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Recovering genomic clusters of secondary metabolites from lakes: a Metagenomics 2.0 approach

Rafael R. C. Cuadrat[1,2], Danny Ionescu[1], Alberto M. R. Davila[3], Hans-Peter Grossart[1,4] *

[1] Leibniz-Institute of Freshwater Ecology and Inland Fisheries, Alte Fischerhuette 2, OT Neuglobsow, 16775, Stechlin, Germany
[2] Berlin Center for Genomics in Biodiversity Research, Königin-Luise-Straße 6-8, 14195, Berlin, Germany
[3] Computational and Systems Biology Laboratory, Oswaldo Cruz Institute, Fiocruz, Avenida Brasil 4365, Rio de Janeiro CEP 21040-360, Brazil
[4] Potsdam University, Institute for Biochemistry and Biology, Potsdam, Germany

* Corresponding author

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Background

Metagenomic approaches became increasingly popular in the past decades due to decreasing costs of DNA sequencing and bioinformatics development. So far, however, the recovery of long genes coding for secondary metabolism still represents a big challenge. Often, the quality of metagenome assemblies is poor, especially in environments with a high microbial diversity where sequence coverage is low and complexity of natural communities high. Recently, new and improved algorithms for binning environmental reads and contigs have been developed to overcome such limitations. Some of these algorithms use a similarity detection approach to classify the obtained reads into taxonomical units and to assemble draft genomes. This approach, however, is quite limited since it can classify exclusively sequences similar to those available (and well classified) in the databases.

In this work, we used draft genomes from Lake Stechlin, north-eastern Germany, recovered by MetaBat, an efficient binning tool that integrates empirical probabilistic distances of genome abundance, and tetranucleotide frequency for accurate metagenome binning. These genomes were screened for secondary metabolism genes, such as polyketide synthases (PKS) and non-ribosomal peptide synthases (NRPS), using the Anti-SMASH and NAPDOS workflows.

Results

With this approach we were able to identify 243 secondary metabolite clusters from 121 genomes recovered from the lake samples. A total of 18 NRPS, 19 PKS and 3 hybrid

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PKS/NRPS clusters were found. In addition, it was possible to predict the partial structure of several secondary metabolite clusters allowing for taxonomical classifications and phylogenetic inferences.

Conclusions

Our approach revealed a great potential to recover and study secondary metabolites genes from any aquatic ecosystem.

Keywords: Metagenomics 2.0, secondary metabolites, nonribosomal peptide synthetase, polyketide synthase

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Background

Metagenomics, also known as environmental genomics, describes the study of a microbial community without the need of *a priori* cultivation in the laboratory. It has the potential to explore uncultivable microorganisms by accessing and sequencing their nucleic acid [1]. In recent years, due to decreasing costs of DNA sequencing - metagenomic databases [2] (e.g., MG-RAST) have rapidly grown and archive billons of short read sequences [3]. Many metagenomic tools and pipelines were proposed to better analyse these enormous datasets [4]. Additionally, these tools allow to (i) infer ecological patterns, alfa- and beta-diversity and richness [5]; (ii) assemble environmental contigs from the reads [6] and more recently, (iii) recover draft genomes from metagenomic bins [7][8][9]. By recovering a high number of draft genomes from these so far uncultivable organisms, it is now possible to screen for new genes and clusters, unlocking a previously underestimated metabolic potential such as secondary metabolite gene clusters by using a metagenomic approach called Metagenomics 2.0 [10].

Polyketide synthases (PKS) and non-ribosomal peptide synthases (NRPS) are two families of modular mega-synthases, both are very important for the biotechnological and pharmaceutical industry due to their broad spectrum of products, spanning from antibiotics and antitumor drugs to food pigments. They act in an analogous way, producing polyketides (using acyil-coA monomers) and peptides (using aminoacyl monomers), respectively. Both families are broadly distributed in many taxonomical groups, ranging from bacteria (alphaproteobacteria, cyanobacteria, actinobacteria) to fungi [11][12].

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

PKS enzymes can be classified in types (I, II and III), where type I can be further classified into modular or iterative classes. The iterative PKS use the same domain many times, iteratively, to synthetize the polyketide. The modular PKS are large multi-domain enzymes in which each domain is used only once in the synthesis process [13][14]. The production of the polyketide follows the co-linearity rule, each module being responsible for the addition of one monomer to the growing chain [15].

Type I PKS are characterized by multiple domains in the same open reading frame (ORF) while in type II each domain is encoded in a separate ORF, acting interactively [16]. Type III is also known as Chalcone synthase and has different evolutionary origin from type I and II [17]. Type III PKSs are self-contained enzymes that form homodimers. Their single active site in each monomer catalyzes the priming, extension, and cyclization reactions iteratively to form polyketide products [17]. Hybrid PKS/NRPS and NRPS/PKS are also modular enzymes, encoding lipopeptides (hybrid between polyketides and peptides) and occur in bacterial as well as fungal genomes [18][19][20].

PKS and NRPS are very well explored in genomes from cultivable organisms, mainly *Actinomycetes* [21] and *Cyanobacteria* [22]. Recently, by using a metagenomic approach, studies have demonstrated the presence of these metabolite-genes in aquatic environments, as for example, Brazilian coast waters (in free living and particle-associated bacteria) and from the microbiomes of Australian marine sponges [23]. However, there are few metagenomic studies whose scope is to find these gene families in freshwater environments where most studies are based on isolation approaches [24][25].

In addition, due to the rather large size of genes involved in these pathways, yet, it is not possible to recover the full genes by using traditional read-based metagenomics or the

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

single sample assembly approach. Most of the studies aim to solely find specific domains, like Keto-synthase (KS) in PKS and Condensation domain (C) in NRPS, due to the high conservation of these domains [26].

We used a metagenomics 2.0 approach to overcome these limitations and improve the screening for secondary metabolism genes and clusters while evaluating the potential of microbial communities for future research on potential drugs. This study aims to (i) generate draft genomes from Lake Stechlin; (ii) to screen these genomes for new complete multi-modular enzymes from PKS and NRPS families, exploring their diversity and phylogeny.

Methods

Sampling and sequencing

A total of 26 metagenomic samples from Lake Stechilin, north-eastern Germany were used. Water was collected as metagenomic samples on several occasions (April, June 2013, July 2014, Aug 2015) in sterile 2 L Schott bottles from Lake Stechlin (53°9’5.59N, 13°1’34.22E). All samples, except those from Aug 2015, were filtered through 5 µm and subsequently 0.2 µm pore-size filters. The samples collected in Aug 2015 were not size-fractionated and directly filtered on a 0.2 µm pore size filter due to specific research

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

demands. Genomic DNA was extracted using a phenol/chloroform protocol as described in [27] and was sent for sequencing.

Sequencing was conducted at MrDNA (Shallowater, Texas) on an Illumina Hiseq 2500, using the V3 chemistry, following, fragmentation, adaptor ligation and amplification of 50 ng genomic DNA from each sample, using the Nextera DNA Sample Preparation Kit.

Table S1 shows the general information about the 26 samples used in this study.

Environmental draft genomes

Briefly, all samples were pre-processed by Nesoni (https://github.com/Victorian-Bioinformatics-Consortium/nesoni) to remove low quality sequences and to trim adaptors, and afterwards assembled together using MegaHIT (default parameters) [6]. The reads from each sample were mapped back to these assembled contigs using BBMAP (https://sourceforge.net/projects/bbmap/) and then all data was binned using MetaBAT [7] to generate the draft genomes. The completeness and taxonomical classification were checked using CheckM [28].

Screening secondary metabolism genes and phylogenetic analysis of NRPS and PKS domains

DNA fasta files of the generated bins (288) were submitted to a locally installed version of Anti-SMASH (--clusterblast –smcogs –limit 1500) [29]. Using in-house ruby scripts, the domains from PKS and NRPS were parsed. The PKS KS domains and NRPS C domains

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

were submitted to NAPDOS for classification [30]. In addition, all the KS and C domains (trimmed by NAPDOS) were submitted to BLASTP against RefSeq database [31], using the default parameters. The 3 best hits of each domain were extracted and added to the original multi-fasta file with the environmental domains. The full set of KS and C domains (from bins and references obtained by the blast on RefSeq database) was submitted for NAPDOS for the phylogenetic analysis. The resulting alignment and tree were exported and the trees were manually checked and annotated.

Relative abundance of bins in each sample

The reads from each sample were mapped (using BBMAP) against each bin fasta file and an in-house ruby parser script was used to calculate the relative abundance of each bin in each sample, normalizing the read counts by the number of reads of each sample. The table with the results was loaded into STAMP [32] in order to analyse the significant differences of bin abundance over the samples.

Results

Environmental draft genomes obtained (bins)

Metagenomic binning resulted in 288 draft environmental genomes (called bins in this study). Of these, 45 had a predicted completion level of more than 75% according to CheckM.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S2 shows the general information about each bin, including completeness, genome size, number of open reading frames (ORFs) and taxonomical classifications (from CheckM).

Screening secondary metabolism genes and phylogenetic analysis

By using Anti-SMASH, at least one secondary metabolite gene cluster was found in 121 of the bins, totalling 243 clusters and 2200 ORFs. From these 243 clusters, 125 (51.4%) were classified in the Terpene and 35 (14.40%) in the Bacteriocin pathway. In addition, a total of 18 NRPS, 6 type I PKS and 3 hybrid PKS/NRPS clusters were found in 15 different bins (Figure 1a). The latest 3 obtained pathway clusters are the main focus of our study.

Figure 1b shows the taxonomical classification at phylum level for the bins showing NRPS, type I PKS and hybrid clusters. Supplementary table S3 shows the distribution of all clusters in all bins.

<FIGURE 1>

Figure 1A: Abundance of secondary metabolite cluster types obtained with Anti-SMASH in the recovered 288 bins (environmental genomes). B: Taxonomical classification of bins (Phyla) in

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

which NRPS, PKS and Hybrid PKS/NRPS clusters were found. Red bar and pie: NRPS; blue bar and pie: Type I PKS; green bars and pie: Hybrid clusters (NRPS-PKS and PKS-NRPS).

A total of 43 condensation (C) domains were obtained from NRPS clusters. All these sequences were submitted to NAPDOS analysis. Figure 2a shows the classification of C domains into classes. Most of the sequences were classified as LCL domains (58%). This kind of domain catalyzes the formation of a peptide bond between two L-amino acids.

<FIGURE 2>

Figure 2A: NAPDOS classification of the NRPS KS domain. Modular: possess a multidomain architecture consisting of multiple sets of modules; hybridKS: are biosynthetic assembly lines that include both PKS and NRPS components; PUFA: Polyunsaturated fatty acids (PUFAs) are long chain fatty acids containing more than one double bond, including omega-3-and omega-6- fatty acids; Enediyne: a family of biologically active natural products. The Enediyne core consists of two acetylenic groups conjugated to a double bond or an incipient double bond within a nine- or ten-membered ring. **2B: NAPDOS classification of NRPS C domain.** Cyc: cyclization domains catalyze both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues; DCL: link an L-amino acid to a growing peptide ending with a D-amino acid; Epim: epimerization domains change the chirality of the last amino acid in the chain from L- to D- amino acid; LCL: catalyze formation of a peptide bond between two L-amino acids; modAA: appear to be involved

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the modification of the incorporated amino acid; Start: first module of a Non-ribosomal peptide synthase (NRPS).

The screening for type I PKS resulted in 9 KS domain sequences. Most of them are classified as modular type I PKS (56%). All of them were submitted to NAPDOS and classified into 4 different classes (Figure 2b).

All the KS and C domains were also submitted to similarity analysis by using BLASTP against RefSeq database (table S4 and S5), and the best 3 hits of each sequence were extracted and used for phylogenetic analyses with NAPDOS. The trees for C and KS domains are shown in Figures 3 and 4, respectively.

<FIGURE 3>

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3: NAPDOS phylogenetic tree of C domains (environmental domains, the top 3 blast results on RefSeq and the NAPDOS reference sequences). The shadow colours represent the domain classifications (LCL, CYC, Start domains, EPIM, ModAA, Dual, and DCL). The sidebars represent phyla (Proteobacteria, Cyanobacteria, Firmicutes, Actinobacteria, Verrucomicrobia). All sequences from environmental bins are in red.

<FIGURE 4>

Figure 4: NAPDOS tree of KS domains (environmental domains, the top 3 blast results on RefSeq and the NAPDOS reference sequences). The shadow colours represent the domain classifications (Modular, KS1, Iterative, Trans-AT, Hybrid, PUFA, Enediyenes, Type II and Fabs – Fatty acid synthase). The sidebars represent phyla (Cyanobacteria and Actinobacteria). All the sequences from environmental bins are in red.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Relative abundance of bins on each sample

The relative abundance of all bins in each sample was estimated by mapping the reads from each sample against the assembled bins. Table S6 shows the normalized bin abundance for every sample.

Due to differences between the filtration methods, we decided to classify the samples in 3 groups: particle associated samples (PA) – filtered on 5 µm membranes (and also samples from aggregates), free-living samples (FL) – pre-filtered through 5.0 µm membranes and subsequently filtered on 0.22 µm membranes, and non-size fractionated samples (NSF) – filtered direct on 0.22 µm membranes (without previous filtering) retaining the whole bacterial community. In total, we obtained 7 samples in the PA group, 5 in the FL group, and 14 in the NSF group.

The table with the relative abundance of the bins in all samples was loaded on STAMP and an ANOVA test was conducted followed Games-Howell POST-HOC test and Benjamini-Hochberg FDR correction. Table S7 shows 158 bins for which the difference in relative abundance was statistically significant ($p < 0.05$) between the 3 groups (FL, PA and NSF).

From the 15 bins containing NRPS and/or type I PKS clusters, only 4 showed significant difference between the 3 groups. Bins 1 and 2 are more abundant in PA samples and bins 193 and 235 are more abundant in the FL samples.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Exploring NRPS, type I PKS and hybrid clusters from draft genome bins

We highlight 3 bins (with less than 35% contamination and more than 70% completeness) out of the 15 obtained type I PKS and/or NRPS and explore their clusters.

In bin 34 (*Pseudomonas*, 98.28% completeness) it was possible to retrieve 7 clusters, including 3 NRPS clusters (Figure 7) and 2 Bacteriocin clusters.

In cluster 2 (ctg181), multiple domains of NRPS (with the 3 minimal modules) and regulatory genes were identified, e.g., smCOG: SMCOG1057 (TetR family transcriptional regulator) (Figure 5, in green arrows).

<FIGURE 5>

Figure 5: Bin 34, NRPS clusters detailed annotation and synteny. The synteny of the clusters with a functional classification for each ORF is given. In addition, for the NPRS biosynthetic ORFS the

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

domain annotations are given. CAL: Co-enzyme A ligase domain, C: condensation, A: adenylation, E: epimerization, TE: Termination, KR: Ketoreductase domain, and ECH: Enoyl-CoA hydratase

All clusters show a high similarity with *Pseudomonas* proteins. Cluster 2 has a similarity of 92% with *Pseudomonas synxantha* bg33r, conserving also the gene synteny.

The C domain sequences were submitted to NAPDOS analysis and 2 were classified as belonging to the Syringomycin pathway and the LCL class, and one was classified as belonging to the Microcystin pathway or/and the DCL class (links an L-amino acid to a growing peptide ending with a D-amino acid).

In cluster 3 (ctg415) (Figure 5), in addition to the NRPS domains, the following transporter related genes are found: smCOG: SMCOG1288 (ABC transporter related protein) and SMCOG1051 (TonB-dependent siderophore receptor) (blue narrows). Nevertheless, this cluster is not complete and just one C domain was found (LCL class), which was also classified to the Syringomycin pathway.

Cluster 6 (ctg857 – Figure 5) shows many NRPS domains, regulatory factors and transporters genes, including drug resistance genes, e.g. SMCOG1005 (drug resistance transporter, EmrB/QacA), SMCOG1044 (ABC transporter, permease protein) and SMCOG1051 (TonB-dependent siderophore receptor) (blue arrows). Two C domains from this cluster were classified as belonging to the heterocyclization class. This class catalyzes both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues [33]. Both domains were classified in the Pyochelin pathway by NAPDOS. The

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

phylogenetic tree (Figure 3) confirms both the functional and taxonomical classification (confidence value 100).

<FIGURE 6>

Figure 6: Bin 193 and 131 type I PKS clusters detailed annotation and synteny. It is possible to see the synteny of the cluster with the functional classification for each ORF. In addition, for the PKS biosynthetic ORF the domain-specific annotations can be seen. KS: keto-synthase, AT: acyltransferase, KR: ketoreductase, E: epimerization, DH: eehydratase, and ER: enoylreductase.

In **bin 193** (Mycobacterium, 73.37 % completeness) a type I PKS cluster was identified (ctg514) (Figure 6). Five PKS domains were retrieved, including the minimal core from one of the ORFs on this contig. The KS domain BLASTP result shows 82% (and 99% coverage) similarity with *Mycobacterium kansasii*. The NAPDOS analysis from the KS domain suggests that it could be a modular (Epothilone pathway) or iterative type I PKS similar to the Calicheamicin pathway. However, by using the phylogenetic analysis it was clustered as the iterative clade (confidence value 98.4) together with the *Mycobacterium kansasii* sequence (confidence value 100) (Figure 4).

Two further clusters were recovered: one type III PKS and one unclassified one. All clusters show similarity with the *Mycobacterium* clusters.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bin 131 (unclassified bacteria by CheckM) has 84.09% of completeness reported by CheckM. In this bin it was possible to find one cluster and 3 domains of type I PKS (KS, AT and KR) (Figure 6). The KS domain was classified by NAPDOS as belonging to the Maduropeptin and Neocarzinostatin pathways. Using clusterblast inside Anti-Smash it was not possible to find any similar cluster, but using BLASTP it was possible to find similarity with the cyanobacteria *Microcystis aeruginosa* (64% identity and 99% coverage on BLASTP search). In the phylogenetic tree, it was clustered within the Enediynes clade (Figure 4) and also with *Microcystis aeruginosa* (confidence value 99).

In addition, there are 12 more bins with NRPS or type I PKS clusters, but with less than 70% of completeness or more than 35% contamination. The bins 1 (69.54% completeness) and 2 (16.52%) were classified as the genus *Anabaena*, showing NRPS and hybrid NRPS-type 1 PKS, respectively. Bins 6 (39.66% completeness), 7 and 8 were classified as the genus *Planktothrix* and show a high diversity of secondary metabolites: 3 NRPS clusters, one type I PKS and 2 NRPS-PKS hybrid (bin 6), 2 NRPS (bin 7) and 2 NRPS and one type I PKS (bin 8). The bin 8 also shows a Microviridin cluster.

Bins 73 and 217 are classified as *Acidobacteria* showing PKS and NRPS, respectively. Bin 235 (*Burkholderiaceae* family) shows 3 NRPS clusters and bin 13 (*Comamonadaceae* family, also from *Burkholderiales* order) shows one NRPS cluster.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Bin 78 is classified as *Verrucomicrobiaceae* shows 1 NRPS cluster. Additionally, there is an unclassified *Archaea* (NRPS cluster) and one bin without any classification (bin 136, type I PKS).

The Anti-Smash results for all bins are available in the Supplemental Information (SI 1).

Discussion

The field of metagenomics has generated a vast amount of data in the last decades [34]. Most of the data is poorly annotated and little quality controlled when loaded into the public databases, hence awaiting a more in-depth analysis [35]. There are many open challenges in this field, e.g., (i) the lack of representative genomic databases from uncultivable organisms to be used in a similarity-based annotation procedure; (ii) high confidence assembly of short reads from species-rich samples; (iii) obtaining high enough coverage for every organism in the sample, including those with a low abundance, etc. [36]. Recently, some new algorithms have been proposed to overcome these limitations and to obtain partial or near complete genomes from environmental samples, e.g., MetaBat [7] and MetaWatt [8]. Most of them require many samples and high coverage sequencing per sample as an input. Recently, studies have been done to recover genomes even from rare bacteria [37]. The term Metagenomics 2.0 was introduced to describe this new generation of metagenomic analysis by Katherine McMahon [10] and most of the studies using this approach have been conducted to reveal ecological interactions and networks [38][39].

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this study, we recovered 288 environmental draft genomes using 26 samples from Lake Stechlin, a temperate oligo-mesotrophic lake. One of the advantages of this approach is to enable the recovery of large genomic clusters, especially the Megasynthases clusters of the secondary metabolism, e.g., involved in biosynthesis of antibiotics, including its regulatory and transporter genes. Here, we have used the Anti-SMASH and NAPDOS pipelines to identify, annotate, classify and to carry out the phylogenetic analysis of a total of 243 clusters of known secondary metabolites. To our knowledge, this is the first study using the metagenomics 2.0 approach to recover Megasynthases clusters. A number of previous studies had been conducted using a traditional PCR based screening [40] and shotgun metagenomics approach [41][42] exploring the abundance and diversity of individual genes and domains, but these studies are missing the genomic context. By obtaining the entire genomic context it is possible, in future studies, to clone and to do heterologous expression for all the genes, including promoters and transporters.

Screening the bins for secondary metabolite clusters, we can see that the most abundant cluster belongs to the Terpene pathway (125 clusters) (Figure 1). This biosynthesis pathway is well known to be present in many plant and fungi genomes, but recently it was proposed to be also widely distributed in bacterial genomes. One study revealed 262 distinct terpene synthases in the bacterial domain of life [43]. Consequently, it can represent a fertile source of new natural products – yet, greatly underestimated. The second most abundant class of clusters belongs to the Bacteriocin pathway (35 clusters) (Figure 1). Bacteriocin is a group of ribosomal synthesized antimicrobial peptides which can kill or inhibit bacterial strains closely related or non-related to the Bacteriocin producing bacteria [44]. It has been

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

suggested as a viable alternative to traditional antibiotics and can be used as narrow-spectrum antibiotics [45]. Only a few studies have been conducted to screen for Bacteriocin genes by using a metatenomic approach, and solely for the host-associated microbiome [46] or fermented food microbiome [47][48][49]. None of these studies was conducted for natural environments and none have used the metagenomics 2.0 approach.

In this study, we focused on 2 families of large modular secondary metabolite genes, type I PKS and NRPS. With our approach, it was possible to find a total of 18 NRPS, 6 type I PKS and 3 hybrid PKS/NRPS clusters. For NRPS clusters, it was possible to recover 43 C domains, most of them (58%) from the LCL class. An LCL domain catalyzes a peptide bond between two L-amino acids [50]. A previous study also found that the LCL class was the most abundant in another aquatic environment, dominated by gram-negative bacteria [42]. Many studies have shown that the LCL class in aquatic environments is limited to gram-negative bacteria [51]. Our results further support this as we also found the LCL class only in bins of gram-negative bacteria (Figure 3) with the only exception of the unclassified Archaea (bin 233), which should be further investigated in order to confirm the phylogenetic classification. It was also possible to recover 9 KS domains from the type I PKS clusters, 56% from the modular class and 22% from the hybrid PKS/NRPS class. Those classes are larger (with many copies of each domain) than the iterative ones, increasing the chances to be recovered by metagenomic approaches. Accordingly, the NRPS and PKS clusters were more in depth analyzed, including syntheny, domain phylogeny, and partial metabolite protein structure predictions.

From the bins showing secondary metabolite genes, the most complete was from bin 34 (*Pseudomonas*). The 7 clusters on this genome vary from 8,675 base pairs (bp) to 52,516

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bp in size, been only possible to be recovered due to the high completeness of the assembled genome (98.28%). The presence of a great diversity of clusters in Pseudomonas is expected, as many active secondary metabolites (encoded by NRPS) have been previously described in the Pseudomonas genus, ranging from antibiotics and antifungal to siderophores [52][53][54].

From those 7 clusters in bin 34, the NRPS clusters 2 and 3 showed a high similarity with Syringomycin (three domains) and Microcystin (one domain) pathways. The first one is found, for example, in the *Pseudomonas syringae* (a plant pathogen) genome, as a virulence factor (Syringomycin E) [55] which also has antifungal activity against *Saccharomyces cerevisiae* [56]. On the other hand, Microcystin is a class of toxins produced by freshwater Cyanobacteria species [57] and it can be produced in large quantities during massive bloom events [58]. Due to the taxonomical classification of the bin and the higher number of domains similar to Syringomycin, however, it is more likely that the product encoded by this cluster is functionally close to the latter pathway.

In bin 34 - cluster 6, both C domains were classified as belonging to the Pyochelin pathway. This peptide is a siderophore of *Pseudomonas aeruginosa* [59]. The presence of a TonB-dependent siderophore receptor in cluster 6 provides additional evidence about its functional classification. Additionally, two Bacteriocin clusters and one Aryl polyene cluster were found in bin 34. Aryl polyenes are structurally similar to the well-known carotenoids with respect to their polyene systems and it was recently demonstrated that it can protect bacteria from reactive oxygen species, similarly to what is known for carotenoids [60]. These results suggest that a wide range of metabolites is encoded in this

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Pseudomonas genome, providing it an "arsenal" of secondary products, increasing the likelihood of the *Pseudomonas* species to succeed in aquatic systems.

Bin 131 (unclassified bacteria) shows a PKS cluster and 3 domains. It was classified as belonging to the Enediynes pathway. These compounds are toxic to DNA and are under investigation as anti-tumor agents, with several compounds under clinical trials [61]. All are encoded by type I iterative PKS [62] and it was possible to recover the minimal core (KS, AT and KR) as well as transporter genes from the environmental genome. The most similar PKS I present in the public databases stems from *Microcystis aeruginosa*, but only with an identity of 64%, suggesting that it is encoding for a new compound, which has not previously been described.

In bin 193 (*Mycobacterium* - sister linage *M. rhodesiae*), one of the 3 recovered clusters is a Type I PKS, similar to iterative PKS in the NAPDOS analysis (confirmed by the phylogenetic analysis). The most similar KS sequence belongs to *Mycobacterium kansasii*, with 82% similarity. *M. rhodesiae* and *M. kansasii* are both non-tuberculous mycobacteria (NTM) that can be found in different environments, but both can also be opportunistic pathogens and cause a chronic pulmonary infection in immunosuppressed patients [63]. The species *M. kansasii* comprises various subtypes and some are often recovered from tap water and occasionally from river or lake water [64][65]. There is still controversy about how the transmission from environment to human host occurs and also about the implications on public health. The presence of PKS in Mycobacterium genus was discovered more than a decade ago and most of the polyketides encoded by different species of this genus play role in virulence and/or components of the extraordinarily

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

complex mycobacterial cell envelope [66]. Further studies must be done in order to investigate better the potential of this bin to cause infections on humans, i.e. by screening virulence factors on the full genome.

Five bins of the phyla Cyanobacteria contained PKS and NRPS clusters. In bins 1 and 2 (Anabaena, now called "Dolichospermum"), it was possible to recover 2 NRPS and 1 hybrid NRPS-PKS, respectively. The genus *Anabaena* is known to encode several toxins, including the dangerous Anatoxin-a, and to produce toxic blooms in lakes and reservoirs [67][68][69][70]. However, the Anatoxin-a is encoded by a type I PKS cluster [71], unlike the NRPS and Hybrid clusters found in the *Anabaena* bins from this study.

On the other hand, the hepatotoxic heptapeptide of the class Mycrocystin is present in many genera of Cyanobacteria, including *Anabaena*, and they are encoded by NRPS and also Hybrid NRPS-PKS clusters [72][73][74][75]. The results of NAPDOS reveal one C domain from bin 1 classified in the pathway of Mycrocystin with e-value 6e-83.

In bins 6, 7 and 8 (*Planktothrix*), it was possible to find several type I PKS, NRPS and hybrid clusters. In bin 6 there are 3 NRPS, one type 1 PKS and one hybrid cluster. The bin 7 shows 2 NRPS and bin 8 shows one type I PKS and 2 NRPS clusters. The genus *Planktothrix* also can be producer of Anatoxin-a [75] and the presence of type I PKS cluster on these bins can be alarming. However, the 3 KS domains from type I PKS from bin 6 reveal great similarity with the Epothilone pathway. The NapDOS analysis and the KS domain from bin 8 suggest a high similarity to the neurotoxin Jamaicamides pathway. A previous study showed in 2010 [76] the presence of an Anatoxin-a-producing

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

cyanobacterium in northeastern Germany Lake Stolpsee, rising concerns about the presence of these toxins in the waters of these lakes.

The absence of Anatoxin-a genes in the studied lake is in agreement with previous screening for a toxin screening in the lake [77].

In the bin 73 (*Acidobacteriales*) 1 PKS sequence was found. By the phylogenetic classification, its KS domain is clustered with *trans*-AT KS domains. The AT domains of *trans*-AT PKSs are not integrated into the assembly lines but expressed as free-standing polypeptides, unlike the more familiar *cis*-AT PKSs [78]. However, the NAPDOS result shows the AT domain of this bin in the same ORF with KS and KR domains, showing a syntheny that suggests a *cis*-AT PKS. In addition, the classification by similarity from NAPDOS suggests a polyunsaturated fatty acid (PUFA) but only with 31% of identity

To assess the life style of the bins (free-living or particle-associated), we calculated the relative abundance of the bins in every sample. A total of 158 bins with significant difference between the 3 groups were found (Table S7), however from the 15 bins on which this study focused (showing NRPS and/or type I PKS clusters), only 4 bins (26.6%) were significantly differently present in the life-styles. Bins 1 and 2 (*Anabaena* genus) are more abundant in the PA group, especially on samples B7 and B9 (and also the replicates Old_b7 and Old_b9), accounting for 20-25% (bin 1) and 10-15% (bin 2) on these samples. The very high abundance of these bins on the samples can be explained, based on the long term monitoring program of IGB on Lake Stechlin, by the fact that these samples were collected during the occurrence of a massive cyanobacterial bloom.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

From the other bins containing PKS/NRPS clusters, we can see that Bins 6, 7, and 8 (*Planktothrix*), beside the lack of significant difference between FL and PA groups (p-value > p 0.05), they are clearly more abundant in NSF. The possible explanation for this notion is that the NSF samples were collected during a mesocosm experiment, whereas the other samples were directly derived from the lake.

Conclusions

Using the Metagenomics 2.0 approach, we were able to recover full megasynthases sequences and its genomic context from environmental draft genomes. However, there are limitations, e.g., the genomic coverage of less abundant organisms and the possibility of chimeras. Recently, it has been demonstrated that with an increasing number of samples, it will be possible to recover individual species genomes with a high confidence [7]. In the near future, with the advent of the 3rd generation sequencing, with longer reads, up to 100 kilobases, it will be possible to further improve the quality of the assemblies [79]. These new approaches unlock the possibility of studying these newly recovered environmental pathways and their evolution in detail. Thus, allowing cloning and expressing these clusters will provide new insights on natural products of great interest for biotechnological and pharmaceutical industry. Moreover, studies have demonstrated the possibility to synthesise large functional DNA [80], and together with additional screening techniques, it will be

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

possible to obtain such sequences and to synthesise the full cluster for heterologous expression, skipping the cloning and functional screening process, saving considerable time and money. In addition, the current work highlights the great potential for the discovery of new metabolically active compounds in freshwaters such as oligo-mesotrophic Lake Stechlin. Further, the study of complete or near complete genomes from uncultivated bacteria in the natural environment will enable us to better understand the multiple forms of interactions between species and how they compete for the limiting natural resources.

Declarations

Available of data and materials

The sequences generated for this study (metagenomic reads) were deposited in ENA (PRJEB22274 and PRJEB7963).

Competing interests

The authors declare no competing interests

Funding

This study was supported by the Science without Borders Program (Ciência Sem Fronteiras), CNPq. DI and HPG were funded by German science foundation (DFG) projects Aquameth (GR1540/21-1) and Aggregates (GR1540/28-1).

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Authors' contributions

Conceived and designed the experiments: RRCC, DI, AMRD, and HPG. Performed the experiments: RRCC, DI. Analyzed the data: RRCC, DI, and HPG. Contributed reagents/materials/analysis tools: DI and HPG. All authors wrote the manuscript and revised it for significant intellectual content.

Acknowledgements

We thank Dr. Camila Mazzoni and all the team of Berlin Center for Genomics in Biodiversity Research (BeGenDiv) for allowing us to use the facilities and computational resources for the bioinformatics analyses. Elke Mach and the MIBI group are thanked for their technical support and fruitful discussions.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Rodríguez-Valera F: **Environmental genomics, the big picture?** *FEMS Microbiology Letters* 2004, **231**:153–158.

2. Vincent AT, Derome N, Boyle B, Culley AI, Charette SJ: **Next-generation sequencing (NGS) in the microbiological world: How to make the most of your money**. *Journal of Microbiological Methods* 2016.

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

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

4. Huson DH, Auch AF, Qi J, Schuster SC: **MEGAN analysis of metagenomic data**. *Genome Research* 2007, **17**:377–386.

5. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R: **QIIME allows analysis of high-throughput community sequencing data**. *Nature Methods* 2010, **7**:335–336.

6. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W: **MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph**. *Bioinformatics* 2015, **31**:1674–1676.

7. Kang DD, Froula J, Egan R, Wang Z: **MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities**. *PeerJ* 2015, **3**:e1165.

8. Strous M, Kraft B, Bisdorf R, Tegetmeyer HE: **The Binning of Metagenomic Contigs for Microbial Physiology of Mixed Cultures**. *Frontiers in Microbiology* 2012, **3**.

9. Wu Y-W, Tang Y-H, Tringe SG, Simmons BA, Singer SW: **MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm**. *Microbiome* 2014, **2**:1.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

10. McMahon K: **"Metagenomics 2.0."** *Environmental Microbiology Reports* 2015, **7**:38–39.

11. Gokhale RS, Sankaranarayanan R, Mohanty D: **Versatility of polyketide synthases in generating metabolic diversity**. *Current Opinion in Structural Biology* 2007, **17**:736–743.

12. Koglin A, Walsh CT: **Structural insights into nonribosomal peptide enzymatic assembly lines**. *Natural Product Reports* 2009, **26**:987.

13. Lal R, Kumari R, Kaur H, Khanna R, Dhingra N, Tuteja D: **Regulation and manipulation of the gene clusters encoding type-I PKSs**. *Trends in biotechnology* 2000, **18**:264–274.

14. Cane DE, Walsh CT, Khosla C: **Harnessing the biosynthetic code: combinations, permutations, and mutations**. *Science* 1998, **282**:63–68.

15. **Minowa Y, Araki M, Kanehisa M: Comprehensive Analysis of Distinctive Polyke tide and Nonribosomal Peptide Structural Motifs Encoded in Microbial Genomes. *Journal of Molecular Biology* 2007, 368:1500–1517.**

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

16. Sun W, Peng C, Zhao Y, Li Z: **Functional Gene-Guided Discovery of Type II Polyketides from Culturable Actinomycetes Associated with Soft Coral Scleronephthya sp**. *PLoS ONE* 2012, **7**:e42847.

17. Austin MB, Noel JP: **The chalcone synthase superfamily of type III polyketide synthases.** *Nat Prod Rep* 2003, **20**:79–110.

18. Fisch KM: **Biosynthesis of natural products by microbial iterative hybrid PKS–NRPS**. *RSC Advances* 2013, **3**:18228.

19. Mizuno CM, Kimes NE, López-Pérez M, Ausó E, Rodriguez-Valera F, Ghai R: **A Hybrid NRPS-PKS Gene Cluster Related to the Bleomycin Family of Antitumor Antibiotics in Alteromonas macleodii Strains**. *PLoS ONE* 2013, **8**:e76021.

20. Masschelein J, Mattheus W, Gao L-J, Moons P, Van Houdt R, Uytterhoeven B, Lamberigts C, Lescrinier E, Rozenski J, Herdewijn P, Aertsen A, Michiels C, Lavigne R: **A PKS/NRPS/FAS Hybrid Gene Cluster from Serratia plymuthica RVH1 Encoding the Biosynthesis of Three Broad Spectrum, Zeamine-Related Antibiotics**. *PLoS ONE* 2013, **8**:e54143.

21. Komaki H, Ichikawa N, Hosoyama A, Takahashi-Nakaguchi A, Matsuzawa T, Suzuki K, Fujita N, Gonoi T: **Genome based analysis of type-I polyketide synthase and**

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

nonribosomal peptide synthetase gene clusters in seven strains of five representative Nocardia species. *BMC genomics* 2014, **15**:1.

22. Micallef ML, D'Agostino PM, Sharma D, Viswanathan R, Moffitt MC: **Genome mining for natural product biosynthetic gene clusters in the Subsection V cyanobacteria**. *BMC Genomics* 2015, **16**.

23. Woodhouse JN, Fan L, Brown MV, Thomas T, Neilan BA: **Deep sequencing of non-ribosomal peptide synthetases and polyketide synthases from the microbiomes of Australian marine sponges**. *The ISME journal* 2013, **7**:1842–1851.

24. Zothanpuia, Passari AK, Gupta VK, Singh BP: **Detection of antibiotic-resistant bacteria endowed with antimicrobial activity from a freshwater lake and their phylogenetic affiliation**. *PeerJ* 2016, **4**:e2103.

25. Silva-Stenico ME, Silva CSP, Lorenzi AS, Shishido TK, Etchegaray A, Lira SP, Moraes LAB, Fiore MF: **Non-ribosomal peptides produced by Brazilian cyanobacterial isolates with antimicrobial activity**. *Microbiological Research* 2011, **166**:161–175.

26. Selvin J, Sathiyanarayanan G, Lipton AN, Al-Dhabi NA, Valan Arasu M, Kiran GS: **Ketide Synthase (KS) Domain Prediction and Analysis of Iterative Type II PKS Gene in Marine Sponge-Associated Actinobacteria Producing Biosurfactants and Antimicrobial Agents**. *Front Microbiol* 2016, **7**:2727–12.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

27. Ionescu D, Siebert C, Polerecky L, Munwes YY, Lott C, Häusler S, Bižić-Ionescu M, Quast C, Peplies J, Glöckner FO, Ramette A, Rödiger T, Dittmar T, Oren A, Geyer S, Stärk H-J, Sauter M, Licha T, Laronne JB, de Beer D: **Microbial and Chemical Characterization of Underwater Fresh Water Springs in the Dead Sea**. *PLoS ONE* 2012, **7**:e38319–21.

28. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW: **CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes**. *Genome research* 2015, **25**:1043–1055.

29. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH: **antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters**. *Nucleic Acids Research* 2015, **43**:W237–W243.

30. Ziemert N, Podell S, Penn K, Badger JH, Allen E, Jensen PR: **The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity**. *PLoS ONE* 2012, **7**:e34064.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

31. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, et al.: **Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation**. *Nucleic Acids Res* 2016, **44**:D733–D745.

32. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. **STAMP: statistical analysis of taxonomic and functional profiles**. *Bioinformatics.* 2014;30(21):3123-3124. doi:10.1093/bioinformatics/btu494.

33. Di Lorenzo M, Stork M, Naka H, Tolmasky ME, Crosa JH: **Tandem heterocyclization domains in a nonribosomal peptide synthetase essential for siderophore biosynthesis in Vibrio anguillarum**. *BioMetals* 2008, **21**:635–648.

34. **National Research Council (US) Committee on Metagenomics: Challenges and Functional Applications. The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet**. Washington (DC): National Academies Press (US); 2007. 5, Data Management and Bioinformatics Challenges of Metagenomics.

35. Gilbert JA, Meyer F, Bailey MJ: **The future of microbial metagenomics (or is ignorance bliss?)**. *ISME Journal-International Society for Microbial Ecology* 2011, **5**:777.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

36. Teeling H, Glockner FO: **Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective**. *Briefings in Bioinformatics* 2012, **13**:728–742.

37. Albertsen M, Hugenholtz P, Skarshewski A, Nielsen KARL, Tyson GW, Nielsen PH: **Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes**. *Nature Biotechnology* 2013, **31**:533–538.

38. Sangwan N, Xia F, Gilbert JA: **Recovering complete and draft population genomes from metagenome datasets**. *Microbiome* 2016, **4**.

39. Vanwonterghem I, Jensen PD, Rabaey K, Tyson GW: **Genome-centric resolution of microbial diversity, metabolism and interactions in anaerobic digestion: Genome-centric resolution through deep metagenomics**. *Environmental Microbiology* 2016, **18**:3144–3158.

41. Amos GCA, Borsetto C, Laskaris P, Krsek M, Berry AE, Newsham KK, Calvo-Bado L, Pearce DA, Vallin C, Wellington EMH: **Designing and Implementing an Assay for the Detection of Rare and Divergent NRPS and PKS Clones in European, Antarctic and Cuban Soils**. *PLOS ONE* 2015, **10**:e0138327.

41. Foerstner KU, Doerks T, Creevey CJ, Doerks A, Bork P: **A Computational Screen for Type I Polyketide Synthases in Metagenomics Shotgun Data**. *PLoS ONE* 2008, **3**:e3515.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

42. Cuadrat R, Cury J, Dávila A: **Metagenomic Analysis of Upwelling-Affected Brazilian Coastal Seawater Reveals Sequence Domains of Type I PKS and Modular NRPS**. *International Journal of Molecular Sciences* 2015, **16**:28285–28295.

43. Yamada Y, Kuzuyama T, Komatsu M, Shin-ya K, Omura S, Cane DE, Ikeda H: **Terpene synthases are widely distributed in bacteria**. *Proceedings of the National Academy of Sciences* 2015, **112**:857–862.

44. Yang S-C, Lin C-H, Sung CT, Fang J-Y: **Antibacterial activities of bacteriocins: application in foods and pharmaceuticals**. *Frontiers in Microbiology* 2014, **5**.

45. Cotter PD, Ross RP, Hill C: **Bacteriocins — a viable alternative to antibiotics?** *Nature Reviews Microbiology* 2012, **11**:95–105.

46. Zheng J, Gänzle MG, Lin XB, Ruan L, Sun M: **Diversity and dynamics of bacteriocins from human microbiome: Bacteriocins of human microbiome**. *Environmental Microbiology* 2015, **17**:2133–2143.

47. Illeghems K, Weckx S, De Vuyst L: **Applying meta-pathway analyses through metagenomics to identify the functional properties of the major bacterial communities of a single spontaneous cocoa bean fermentation process sample**. *Food Microbiology* 2015, **50**:54–63.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

48. Więckowicz M, Schmidt M, Sip A, Grajek W: **Development of a PCR-based assay for rapid detection of class IIa bacteriocin genes: Detection of class IIa bacteriocins**. *Letters in Applied Microbiology* 2011, **52**:281–289.

49. Escobar-Zepeda A, Sanchez-Flores A, Quirasco Baruch M: **Metagenomic analysis of a Mexican ripened cheese reveals a unique complex microbiota**. *Food Microbiology* 2016, **57**:116–127.

50. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH: **Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution**. *BMC Evolutionary Biology* 2007, **7**:78.

51. Woodhouse JN, Fan L, Brown MV, Thomas T, Neilan BA: **Deep sequencing of non-ribosomal peptide synthetases and polyketide synthases from the microbiomes of Australian marine sponges**. *The ISME Journal* 2013, **7**:1842–1851.

52. Esmaeel Q, Pupin M, Kieu NP, Chataigné G, Béchet M, Deravel J, Krier F, Höfte M, Jacques P, Leclère V: ***Burkholderia* genome mining for nonribosomal peptide synthetases reveals a great potential for novel siderophores and lipopeptides synthesis**. *MicrobiologyOpen* 2016, **5**:512–526.

53. Van Der Voort M, Meijer HJG, Schmidt Y, Watrous J, Dekkers E, Mendes R, Dorrestein PC, Gross H, Raaijmakers JM: **Genome mining and metabolic profiling of the**

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

rhizosphere bacterium Pseudomonas sp. SH-C52 for antimicrobial compounds. *Frontiers in Microbiology* 2015, **6**.

54. Pan H-Q, Hu J-C: **Draft genome sequence of the novel strain Pseudomonas sp. 10B238 with potential ability to produce antibiotics from deep-sea sediment.** *Mar Genomics* 2015, **23**:55–57.

55. Scholz-Schroeder BK, Soule JD, Gross DC: **The sypA, sypB, and sypC synthetase genes encode twenty-two modules involved in the nonribosomal peptide synthesis of syringopeptin by Pseudomonas syringae pv. syringae B301D**. *Molecular Plant-Microbe Interactions* 2003, **16**:271–280.

56. Stock SD, Hama H, Radding JA, Young DA, Takemoto JY: **Syringomycin E inhibition of Saccharomyces cerevisiae: requirement for biosynthesis of sphingolipids with very-long-chain fatty acids and mannose-and phosphoinositol-containing head groups**. *Antimicrobial agents and chemotherapy* 2000, **44**:1174–1180.

57. Dawson RM: **The toxicology of microcystins.** *Toxicon* 1998, **36**:953–962.

58. Bouhaddada R, Nelieu S, Nasri H, Delarue G, Bouaicha N: **High diversity of microcystins in a Microcystis bloom from an Algerian lake.** *Environ Pollut* 2016, **216**:836–844.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

59. Brandel J, Humbert N, Elhabiri M, Schalk IJ, Mislin GLA, Albrecht-Gary A-M: **Pyochelin, a siderophore of Pseudomonas aeruginosa: Physicochemical characterization of the iron(iii), copper(ii) and zinc(ii) complexes**. *Dalton Transactions* 2012, **41**:2820.

60. Schoner TA, Gassel S, Osawa A, Tobias NJ, Okuno Y, Sakakibara Y, Shindo K, Sandmann G, Bode HB: **Aryl Polyenes, a Highly Abundant Class of Bacterial Natural Products, Are Functionally Related to Antioxidative Carotenoids.** *Chembiochem* 2016, **17**:247–253.

61. Jones GB, Fouad FS: **Designed enediyne antitumor agents.** *Curr Pharm Des* 2002, **8**:2415–2440.

62. Ahlert J, Shepard E, Lomovskaya N, Zazopoulos E, Staffa A, Bachmann BO, Huang K, Fonstein L, Czisny A, Whitwam RE, Farnet CM, Thorson JS: **The calicheamicin gene cluster and its iterative type I enediyne PKS.** *Science* 2002, **297**:1173–1176.

63 - Fedrizzi T, Meehan CJ, Grottola A, Giacobazzi E, Serpini GF, Tagliazucchi S, Fabio A, Bettua C, Bertorelli R, De Sanctis V, Rumpianesi F, Pecorari M, Jousson O, Tortoli E, Segata N: **Genomic characterization of Nontuberculous Mycobacteria**. *Nature Publishing Group* 2017:1–14.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

64. van der Wielen PWJJ, Heijnen L, van der Kooij D: **Pyrosequence Analysis of the hsp65 Genes of Nontuberculous Mycobacterium Communities in Unchlorinated Drinking Water in the Netherlands**. *Applied and Environmental Microbiology* 2013, **79**:6160–6166.

65. Bakuła Z, Safianowska A, Nowacka-Mazurek M, Bielecki J, Jagielski T: **Short Communication: Subtyping of Mycobacterium kansasiiby PCR-Restriction Enzyme Analysis of the hsp65Gene**. *BioMed Research International* 2013, **2013**:1–4.

66. Quadri LEN: **Biosynthesis of mycobacterial lipids by polyketide synthases and beyond**. *Critical Reviews in Biochemistry and Molecular Biology* 2014, **49**:179–211.

67. Brown NM, Mueller RS, Shepardson JW, Landry ZC, Morré JT, Maier CS, Hardy FJ, Dreher TW: **Structural and functional analysis of the finished genome of the recently isolated toxic Anabaena sp. WA102**. *BMC Genomics* 2016:1–18.

68. Li X, Dreher TW, Li R: **An overview of diversity, occurrence, genetics and toxin production of bloom-forming Dolichospermum (Anabaena) species**. *Harmful Algae* 2016, **54**:54–68.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

69. 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**:977–14.

70. Carmichael WW, Biggs DF, Gorham PR. **Toxicology and pharmacological action of *Anabaena flos-aquae* toxin**. Science 1975, ;187:542-544

71. Méjean A, Paci G, Gautier V, Ploux O: **Biosynthesis of anatoxin-a and analogues (anatoxins) in cyanobacteria**. *Toxicon* 2014, **91**(C):15–22.

72. **Structural organization of microcystin biosynthesis in Microcystis aeruginosa PCC7806: an integrated peptide^polyketide synthetase system**. 2000:1–12.

73. Rouhiainen L, Vakkilainen T, Siemer BL, Buikema W, Haselkorn R, Sivonen K: **Genes Coding for Hepatotoxic Heptapeptides (Microcystins) in the Cyanobacterium Anabaena Strain 90**. *Applied and Environmental Microbiology* 2004, **70**:686–692.

74. Rastogi RP, Madamwar D, Incharoensakdi A: **Bloom Dynamics of Cyanobacteria and Their Toxins: Environmental Health Impacts and Mitigation Strategies**. *Front Microbiol* 2015, **6**:223–22.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

75. Viaggiu E, Melchiorre S, Volpi F, Di Corcia A, Mancini R, Garibaldi L, Crichigno G, Bruno M: **Anatoxin-a toxin in the cyanobacterium Planktothrix rubescens from a fishing pond in northern Italy**. *Environ Toxicol* 2004, **19**:191–197.

76. Ballot A, Fastner J, Lentz M, Wiedner C: **First report of anatoxin-a-producing cyanobacterium Aphanizomenon issatschenkoi in northeastern Germany**. *Toxicon* 2010, **56**:964–971.

77. Dadheech P, Selmeczy G, Vasas G, Padisák J, Arp W, Tapolczai K, Casper P, Krienitz L: **Presence of Potential Toxin-Producing Cyanobacteria in an Oligo-Mesotrophic Lake in Baltic Lake District, Germany: An Ecological, Genetic and Toxicological Survey**. *Toxins* 2014, **6**:2912–2931.

78 . Weissman KJ: **The structural biology of biosynthetic megaenzymes**. *Nat Chem Biol* 2015, **11**:660–670.

79. Frank JA, Pan Y, Tooming-Klunderud A, Eijsink VGH, McHardy AC, Nederbragt AJ, Pope PB: **Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data**. *Scientific Reports* 2016, **6**:25373.

80. Hutchison CA, Chuang R-Y, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, others: **Design and synthesis of a minimal bacterial genome**. *Science* 2016, **351**:aad6253.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

917

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

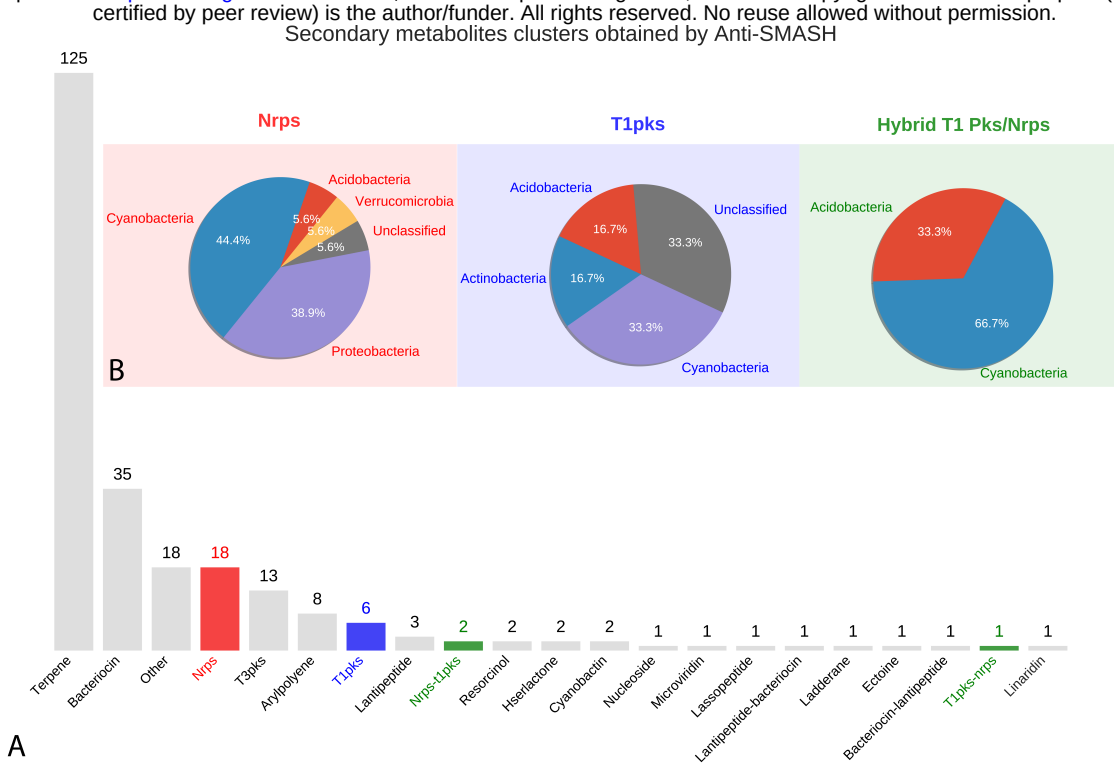


Figure 1A: Abundance of secondary metabolite cluster types obtained with Anti-SMASH in the recovered 288 bins (environmental genomes). B: Taxonomical classification of bins (Phyla) in which NRPS, PKS and Hybrid PKS/NRPS clusters were found. Red bar and pie: NRPS; blue bar and pie: Type I PKS; green bars and pie: Hybrid clusters (NRPS-PKS and PKS-NRPS).

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

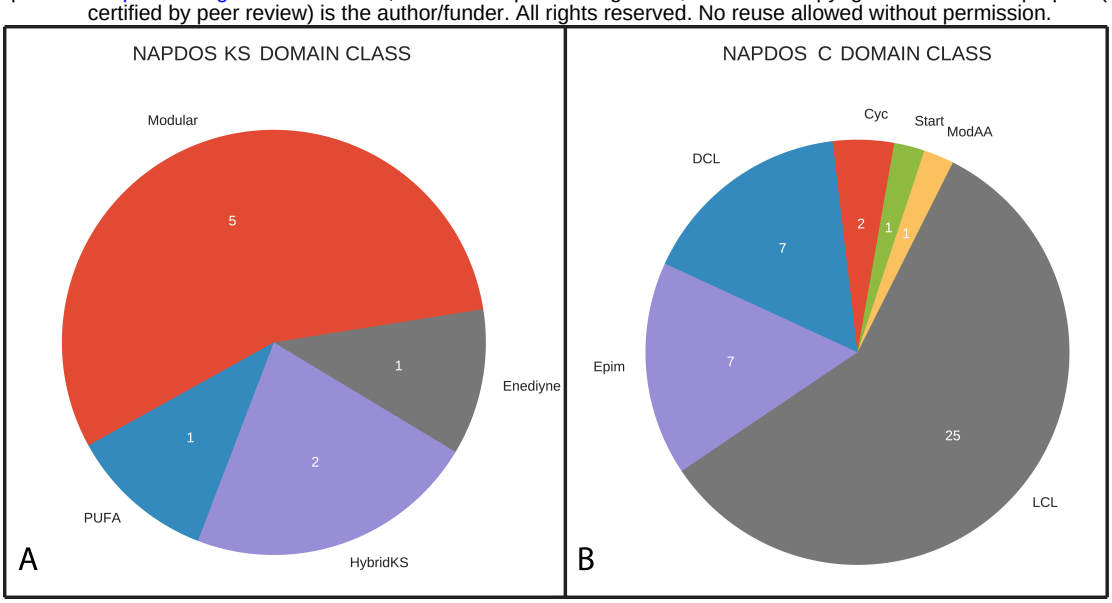


Figure 2A: NAPDOS classification of the NRPS KS domain. Modular: possess a multidomain architecture consisting of multiple sets of modules; hybridKS: are biosynthetic assembly lines that include both PKS and NRPS components; PUFA: Polyunsaturated fatty acids (PUFAs) are long chain fatty acids containing more than one double bond, including omega-3-and omega-6- fatty acids; Enediyne: a family of biologically active natural products. The Enediyne core consists of two acetylenic groups conjugated to a double bond or an incipient double bond within a nine- or ten-membered ring. **2B: NAPDOS classification of NRPS C domain.** Cyc: cyclization domains catalyze both peptide bond formation and subsequent cyclization of cysteine, serine or threonine residues; DCL: link an L-amino acid to a growing peptide ending with a D-amino acid; Epim: epimerization domains change the chirality of the last amino acid in the chain from L- to D- amino acid; LCL: catalyze formation of a peptide bond between two L-amino acids; modAA: appear to be involved in the modification of the incorporated amino acid; Start: first module of a Non-ribosomal peptide synthase (NRPS).

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

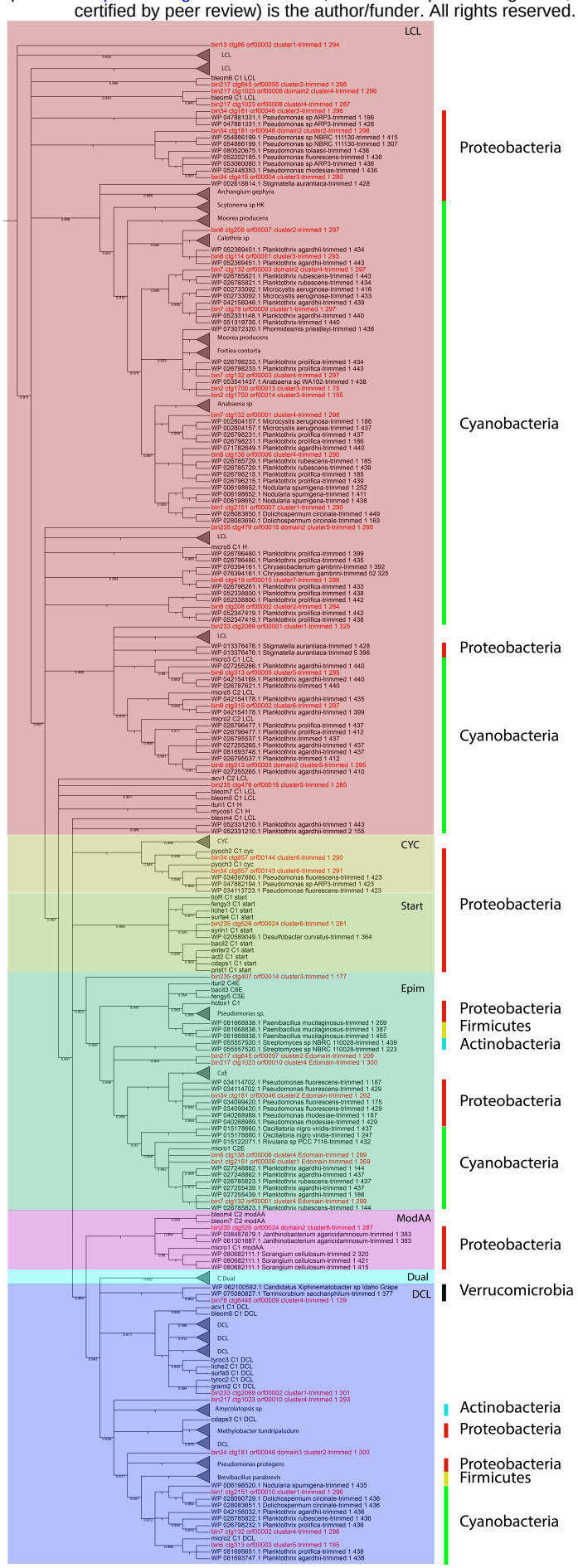
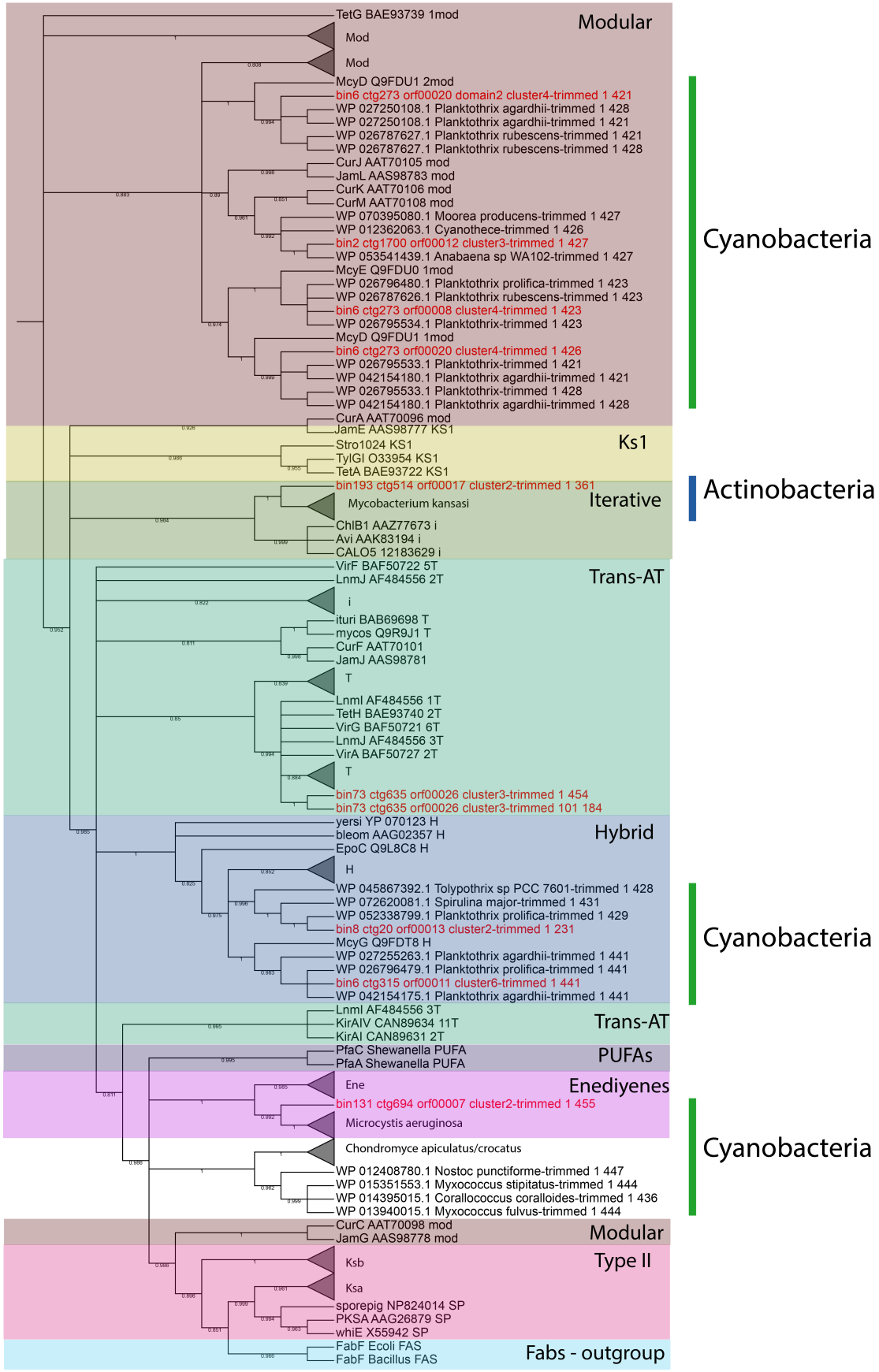


Figure 3: NAPDOS

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

phylogenetic tree of C domains (environmental domains, the top 3 blast results on RefSeq and the NAPDOS reference sequences). The shadow colours represent the domain classifications (LCL, CYC, Start domains, EPIM, ModAA, Dual, and DCL). The sidebars represent phyla (Proteobacteria, Cyanobacteria, Firmicutes, Actinobacteria, Verrucomicrobia). All sequences from environmental bins are in red.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.



bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 4: NAPDOS tree of KS domains (environmental domains, the top 3 blast results on RefSeq and the NAPDOS reference sequences). The shadow colours represent the domain classifications (Modular, KS1, Iterative, Trans-AT, Hybrid, PUFA, Enediyenes, Type II and Fabs – Fatty acid synthase). The sidebars represent phyla (Cyanobacteria and Actinobacteria). All the sequences from environmental bins are in red.

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

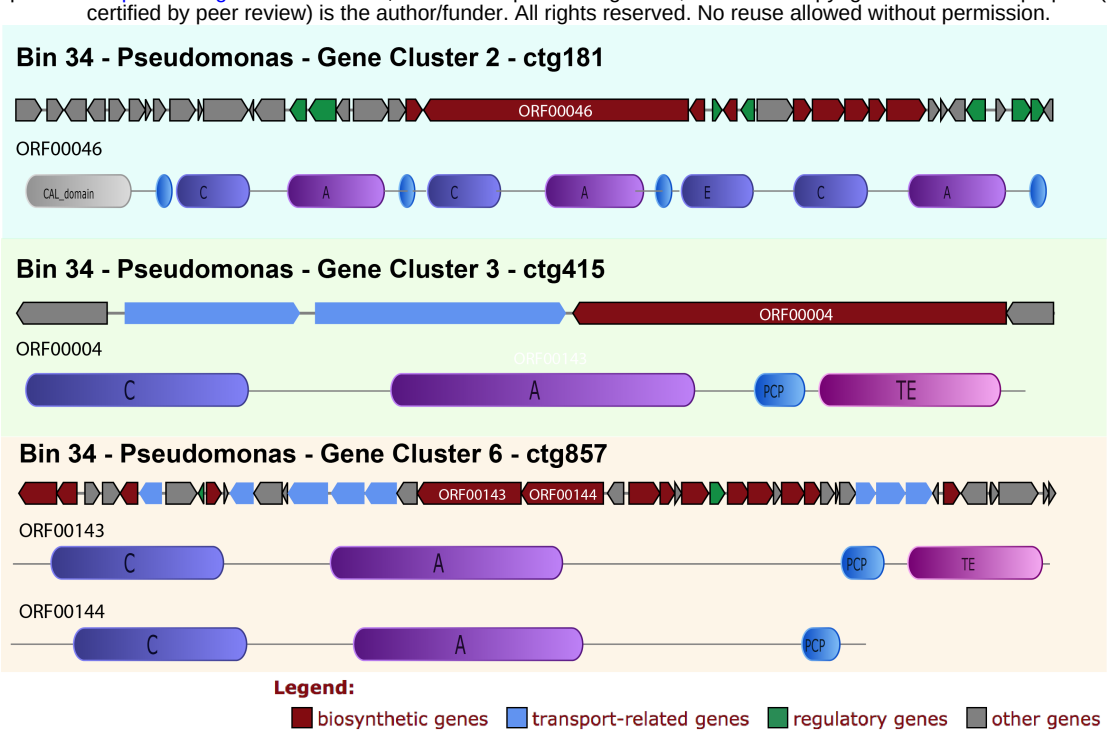


Figure 5: Bin 34, NRPS clusters detailed annotation and synteny. The synteny of the clusters with a functional classification for each ORF is given. In addition, for the NPRS biosynthetic ORFS the domain annotations are given. CAL: Co-enzyme A ligase domain, C: condensation, A: adenylation, E: epimerization, TE: Termination, KR: Ketoreductase domain, and ECH: Enoyl-CoA hydratase

bioRxiv preprint doi: https://doi.org/10.1101/183061; this version posted August 31, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

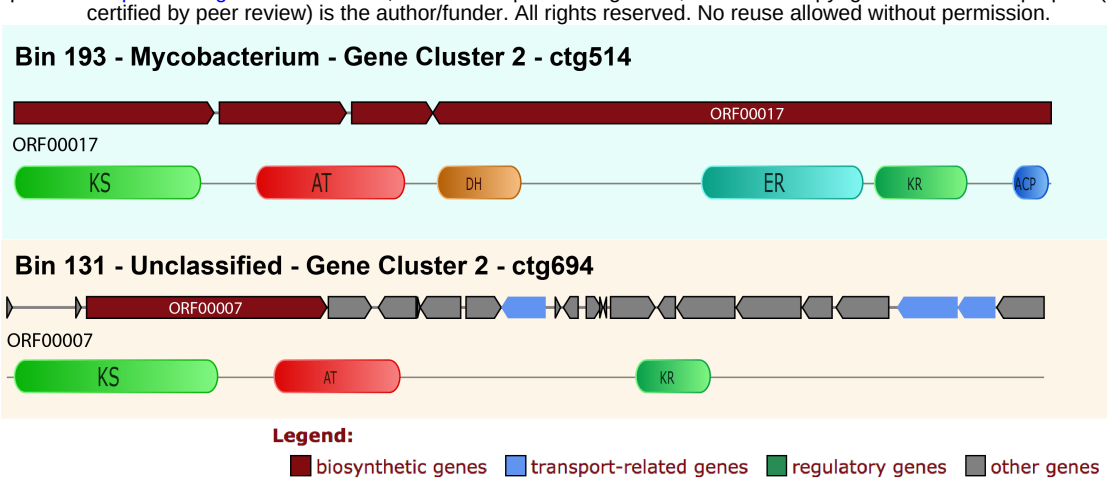


Figure 6: Bin 193 and 131 type I PKS clusters detailed annotation and synteny. It is possible to see the synteny of the cluster with the functional classification for each ORF. In addition, for the PKS biosynthetic ORF the domain-specific annotations can be seen. KS: keto-synthase, AT: acyltransferase, KR: ketoreductase, E: epimerization, DH: eehydratase, and ER: enoylreductase.