The cycad coralloid root contains a diverse endophytic bacterial community including cyanobacteria encoding specific biosynthetic gene clusters

Francisco Barona-Gómez[1], Antonio Corona-Gómez[2], Karina Gutiérrez-García[1], Pablo Cruz-Morales[1,2], Pablo Suárez-Moo[2], Nelly Selem-Mójica[1], Miguel A. Perez-Farrera[3] & Angélica Cibrián-Jaramillo[2,*]

[1]Evolution of Metabolic Diversity and [2]Ecological and Evolutionary Genomics Laboratories, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Km 9.6 Libramiento Norte, Carretera Irapuato-León, CP 36821, Irapuato, Guanajuato, México

[3]Escuela de Biología, Universidad de Ciencias y Artes del Estado de Chiapas, Libramiento Norte Poniente s/n, Col. Lajas-Maciel, CP 29029, Tuxtla Gutiérrez, Chiapas, México.

*Author for correspondence: Angélica Cibrián-Jaramillo. Ecological and Evolutionary Genomics Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Km 9.6 Libramiento Norte, Carretera Irapuato-León, CP 36821, Irapuato, Guanajuato, México. Tel. +462 1663014, angelica.cibrian@cinvestav.mx.

Keywords

Cycad, coralloid root, microbiome, cyanobacteria, phylogenomics, EcoMining.

Abstract

Cycads are the only early seed plants that have evolved a specialized coralloid root to host endophytic bacteria that fix nitrogen for the plant. To provide evolutionary insights into this million-year old symbiosis, we investigate the phylogenetic and functional diversity of its endophytic bacterial community, based on the (meta)genomic characterization of the coralloid root of *Dioon merolae* collected from natural populations. We employed a co-culture-based metagenomics experimental strategy, termed EcoMining, to reveal both predominant and rare bacteria that were analyzed through phylogenomics and detailed metabolic annotation. Most of the characterized bacteria were identified as diazotrophic plant endophytes belonging to at least 18 different bacterial families. The draft genomes of three Cyanobacteria strains were obtained, and after whole-genome inferences they were found to form a monophyletic group, suggesting a level of specialization characteristic of co-evolved symbiotic relationships. In agreement with their large size over 8 Mbp, the draft genomes of these organisms were found to encode for biosynthetic gene clusters predicted to direct the synthesis of specialized metabolites present only in these symbionts. Overall, we provide a new notion of the composition and evolution of the cycad coralloid root that contributes to studies on phylogenetic and functional patterns in plant-bacteria symbiotic systems.

Introduction

Cycads (Cycadales) are among the few plant lineages that developed a specialized root organ to host symbiotic bacteria. Cycad lateral roots can develop into coralloid roots, a dichotomous and coral-like small cluster of roots, typically growing above ground, that acquire symbiotic bacteria (Norstog et al. 1997). Its main function is nitrogen fixation for the plant (Bergensen et al. 1986), akin to adaptive functions in legume nodules, but having appeared millions of years before angiosperms. In natural habitats coralloid roots appear in early life stages of the plant (Halliday et al. 1976); and in adults, mainly in habitats with poor or inaccessible nutrients (Grove et al. 1980), such as sand dunes, sclerophyll forests, steep rock outcrops with high exposure to salt, and lowland forests with recurrent fires. It is possible that the cycad coralloid root and its bacteria are a key early trait that enabled cycads to thrive and adapt to changing environments for millions of years.

Coralloid root endophytes have been studied since the 19th century [(Caiola 1980) and references therein]. However, most studies have focused on resolving the taxonomy of cultured Cyanobacteria, which can be visually detected in the root (see photo in fig. 1). These studies have been carried out using samples collected from botanic garden collections or from plants grown in greenhouses, typically outside of the cycad host natural range (Zimmerman et al. 1992, Costa et al. 1999, Costa Jl et al. 2002, Costa et al. 2004, Gehringer et al. 2010, Thajuddin et al. 2010, Cuddy et al. 2012), with only a handful of field-based studies from wild cycad populations (Gehringer et al. 2010, Yamada et al. 2012). Studies testing for the specificity of Cyanobacteria report contrasting results regarding the specialization of coralloid root symbionts (Caiola 1980, Ow et al. 1999, Zvyagintsev et al. 2010), and there is also inconclusive information on the origin

and transmission of cycad bacterial endophytes (Cuddy et al. 2012). Interestingly, anatomical studies of these Cyanobacteria have shown the presence of mucilaginous or protein-rich material that hosts associated bacteria (Chaudhuri et al. 1931, Caiola 1980, Baulina et al. 2003), whose taxonomic associations have only been suggested on the basis of low resolution data (Bershova et al. 1968, Grobbelaar et al. 1987, Huan et al. 1989 , Thajuddin et al. 2010, Zvyagintsev et al. 2010). Thus, despite progress, there is much to be learned about the bacterial endophyte biodiversity.

In addition to nitrogen fixation there have been suggestions of other unknown roles for the coralloid root (Caiola 1980), but there is no clear evidence of its broader function to date. Likewise, various chemical, physical and physiological processes appear to regulate the cycad-bacteria interaction (Lobakova et al. 2003, Meeks 2009), but no genes involved in specialized mechanisms of the symbiosis have been identified. The presence of uniquely specialized metabolites in the cycad coralloid root bacterial endophytes is of particular interest to us because they may be a result of co-evolution between the cycad host and the endophyte bacterial community, as it has been shown in other symbioses (Adnani et al. 2017). Certain soil-dwelling and aquatic bacteria, including Cyanobacteria, have genomic plasticity and the capacity to synthesize specialized metabolites with overwhelming chemical diversity, which are produced to cope with biotic and abiotic pressures (Traxler et al. 2015). These bacterial traits are concomitant with large genomes that code for specialized metabolites in rapidly evolving genetic units called biosynthetic gene clusters (BGCs) of about 25-40 Kbp (Doroghazi et al. 2014). The ability to capture and retain bacteria in the coralloid root could provide a mechanism for cycads to adapt quickly to local conditions by increasing their specialized metabolite repertoire.

Our goal in this study is to investigate the endophyte bacterial community of the coralloid roots of *Dioon merolae* (De Luca et al. 1981), which is a long-lived, entomophilous, dioecious, and arborescent cycad native to Mexico (Lázaro-Zermeño et al. 2012). We collected coralloid root samples from wild populations in two different habitats from their natural range, currently distributed in moderate population sizes of a few hundreds of individuals throughout Chiapas and Oaxaca in the South of Mexico (Lázaro-Zermeño et al. 2012). To overcome technical difficulties in characterizing the breadth of microbial diversity in these environmental samples, we used an enrichment co-culture strategy of sub-communities obtained from the original sample. This approach, termed EcoMining, uses complementing microbiological and (meta)genomic sequencing to characterize the coralloid microbiome's taxonomic and functional diversity. Our study provides evidence of the existence of a diverse endophytic bacterial community inside the coralloid root beyond Cyanobacteria, and suggests functional specificity, i.e. unique BGCs, throughout the evolutionary history of the cycad-cyanobacteria coralloid root system.

Methods

Overall strategy. We used a combined sub-community co-culture, metagenomics and phylogenomics strategy to detect and measure taxonomic diversity and phylogenetic relationships in the endophytes of the cycad coralloid root, as previously described (Cibrián-Jaramillo et al. 2016) (fig. 1). We grew and isolated bacteria from environmental samples using a diverse set of semi-selective media that aim to capture as much of the cultured bacterial diversity (*t0*), which was compared to a reference of the existing diversity determined with 16S rRNA

single amplicon metagenomics. Simultaneously, we enriched the same samples for specific sub-communities in co-cultures grown under conditions for cyanobacteria using BG11 media. Importantly, in addition to this autotrophic bacterial group, this second approach aims to capture other bacterial groups that interact with and depend on cyanobacteria present in the original sample at low titers (rare taxa). We allowed the sub-community co-culture to grow over time and sampled it after one month (*t1*) and at the end of a year (*t2*). We isolated bacteria grown in selective media over time (*t0* and *t1*) and sequenced sub-communities in co-cultures using shotgun metagenomics (*t1* and *t2*). The phylogenetic relationships of the isolated bacteria were reconstructed using 16S rRNA sequences obtained after PCR amplification and Sanger sequencing, and OTUs obtained from metagenomes. To further assess the natural history of the cyanobacteria-cycad interaction, two cyanobacterial draft genomes from nearly axenic cultures (strains 106C and T09), and one draft genome reconstructed from a metagenomic sub-community (RF31YmG), were obtained and analyzed. Specific methods, materials and analytical tools are described in the following sub-sections.

Field collections. We sampled coralloid roots from two wild cycad populations previously reported (Lázaro-Zermeño et al. 2012). In March of 2014 we sampled from two sites in deciduous tropical forests, at Jiquipilas, Oaxaca Mexico (JP or dry) at 560m above mean sea level, with an average annual precipitation of 320 mm and average annual temperature of 18 °C; and Raymundo Flores, Oaxaca, Mexico (RF or humid) at 900m above mean sea level, with 2500 mm and 25°C annual average precipitation and temperature, respectively (exact coordinates upon request to preserve the integrity of these populations). In some cycad plants, coralloid roots were easily visible aboveground, while in others it was necessary to dig to about 30 cm around the

main root until coralloid roots were found. In a population of approximately 40 individuals, we generally found 10-12 plants with coralloid roots, in almost exclusively juvenile plants. A total of 10 coralloid apogeotropic roots were cut from 10 plants, cleaned with sterile distilled water to remove soil excess, placed in 15 ml sterile Falcon tubes (Beckton Dickinson), and transported immediately to the laboratory at room temperature.

Coralloid root processing. We focused our effort on two samples of two individuals with the largest coralloid roots, in each of the two sites, Jiquipilas (JP or dry) and Raymundo Flores (RF or humid) for a total of four coralloid root samples (JP2 and JP6 plus RF1 and RF3, respectively), and stored the remaining samples at -80 °C for subsequent studies, such as single amplicon metagenomics targeting the 16S region of the rRNA gene for selected samples as references. We treated the coralloid root in a laminar flow hood (Nuaire Model Nu-126-400) with a series of washes to remove exogenous bacteria from the rhizosphere or epiphytic contamination. Each root was introduced in 50 ml sterile Falcon tubes containing 10 ml of each of the following solutions, and gently stirred for: three minutes in hydrogen peroxide (HO), seven minutes in 70% ethanol, 30 seconds in sterile dd-MilliQ water, four minutes in 6% sodium hypochlorite (NaClO), and three one-minute washes in sterile dd-MilliQ water. After this procedure, we plated out water from the last wash in Petri dishes containing the six media described below. Lack of growth in the last wash was considered a negative control. Only samples complying with this criterion were used for endophyte isolation. We undertook two approaches to bacterial isolation (fig. 1): sampling without enrichment directly from field samples treated for removal of epiphytes (*t0*), and sampling from the enriched co-cultures (*t1*), as described in the following sections.

Bacterial growth and isolation in selective media. Loss of some bacterial groups that were present in the original samples collected from the environment which were treated for removal of epiphytes (*t0*) is expected. Thus, we first assessed this potential bias using single amplicon metagenomics targeting the 16S region of the rRNA gene of the JP6 sample. Coralloid roots were washed as previously described and 120 mg of plant tissue was used to extract DNA using the Dneasy Plant mini kit (Qiagen). The amplicon library was developed in the Center for Gastrointestinal Biology and Disease, North Carolina (CGIBD), and sequenced in a 2 x 250 bp paired-end run on the illumina Miseq plataform. Next, to isolate bacteria from the original epiphyte-free field samples (*t0*) and after enrichment (*t1*), macerated roots or co-culture broth were used as inoculant, respectively. After enrichment (*t1*) we recover bacteria expected to be initially present in low abundances and required time to grow, and that did so as a response to the community nutritional interactions (e.g. amino acids derived from the process of fixing nitrogen) (Cibrián-Jaramillo et al. 2016). In all cases, roots were macerated in 10 ml of sterile water using a pestle and mortar until plant material was completely disintegrated. We used 100 μl from the root macerate to directly isolate bacteria in Petri dishes containing a total of six different media, chosen to selectively (oligotrophic, four media) or non-selectively (eutrophic, two media) recover bacterial diversity as much as possible. The four selective media used were chosen to target bacterial groups that are known to be either plant endophytes or rhizosphere bacteria, and included: 1) *Caulobacter* medium (glucose: 1 g/L; peptone: 1g/L; yeast extract: 1.5 g/L; trace metals solution: 1 mL/L; and 10 g/L of agar for solid medium); 2) *Rhizobium* medium (mannitol: 10 g/L; dipotassium phosphate: 0.5 g/L; magnesium sulfate: 0.2 g/L; yeast extract: 1 g/L; sodium chloride: 0.1 g/L; final pH 6.8; and 20 g/L for solid medium; 3) ISP4, for isolation of actinomycetes (starch: 10.0 g/L; dipotassium phosphate: 1 g/L; magnesium sulfate: 1 g/L;

sodium chloride: 1 g/L; ammonium sulfate: 2 g/L; calcium carbonate: 2 g/L; ferrous sulfate: 1 mg/L; magnesium chloride: 1 mg/L; zinc sulfate: 1 mg/L; final pH 7.0; and 20 g/L for solid media); 4) BG-11, a Cyanobacteria medium (sodium nitrate: 1.5 g/L; dipotassium phosphate: 0.04 g/L; magnesium sulfate: 0.075 g/L; calcium chloride: 0.036 g/L; citric acid: 0.006 g/L; ferric ammonium citrate: 0.006 g/L; EDTA (disodium salt): 0.001 g/L; sodium carbonate: 0.02 g/L; final pH 7.1 and agar solid media 10.0 gr/L. The non-selective, rich media, included: 5) Nutrient Broth (BD Bioxon, Mexico); and 6) As in *Caulobacter* medium, but supplemented with mannitol (*Caulobacter* + mannitol medium): 1g/L, with aim of providing a carbon source closer to that hypothetically encountered inside the cycad root.

Bacterial growth and cultivation in sub-communities. 100 μl of the macerated roots free from epiphytes from samples JP2, JP6 and RF1 and RF3, which passed the negative growth control were used to inoculate 100 ml of media in 250 ml flasks. The remaining macerated roots not used for fresh cultures were kept as frozen stocks for future studies (-80 °C in 10% glycerol), although community viability after freezing was found to diminish over time. We used selective oligotrophic medium, i.e. BG11 (medium No. 4) for these experiments. This medium lacks a carbon source but contains an amount of inorganic nitrogen that would not last after long-term cultures, such as *t1* and *t2*. BG11 cyanobacteria-centric co-cultures were grown for up to one year with constant stirring, with cycles of 16/8 hours of light/darkness per day. Sampling of the oligotrophic sub-communities was done after one month (*t1*) and one year (*t2*), and treated independently. Bacterial isolates were only obtained from *t1* whereas shotgun metagenomics were obtained for both time points.

Genomics and shotgun metagenomics. The genomic sequence of the cyanobacterial isolate 106C was obtained from sample JP6, the same sample assessed for its *t0* epiphyte-free bacterial diversity using single amplicon 16S rRNA metagenomics (previous sections). To test for host differences, we also obtained the genome sequence of isolate T09, obtained from coralloid roots of *Dioon caputoi* from an unrelated environment (Xeric shrubland, Tehuacan valley Puebla). Both organisms were grown on BG11 plates before DNA extraction, which was done using a CTAB-phenol chloroform standard protocol. Metagenomic DNA of all enriched sub-community co-cultures (200ml), coming from RF and JP samples, were extracted using the same protocol, with the exception that biomass was obtained from liquid cultures by high-speed centrifugation during 15 minutes. The RF31YmG genome was obtained from one of this metagenomic DNA samples, i.e. RF3-1yr or *t2*, after bioinformatic treatment as described below. DNA samples were processed with truseq nano kit and sequenced at Langebio, Cinvestav (Irapuato, Mexico) using the MiSeq Illumina platform in the 2X250 paired end reads format (T09) and the NextSeq mid output 2X150 paired end read format (106C and RF3-1yr or *t2*). The reads for each library were filtered with fastQ and trimmed using Trimommatic version 0.32 (Bolger et al. 2014). Assembly of 106C and T09 was done *de novo*, whereas the RF31YmG genome was obtained from metagenomic reads of the sub-community co-culture RF3-*t2* of *D. merolae*, using as reference the newly obtained sequence of 106C. The metagenomic reads were filtered by mapping them against the assembly of *Nostoc* sp. 106C with BWA (Li et al. 2009). In all cases the resulting reads were assembled with Velvet (Zerbino et al. 2008) using different *k*-mers: the assemblies with the largest length and the smaller number of contigs were selected and annotated using RAST (Aziz et al. 2008).

Taxonomic diversity. Merged sequences from the amplicon 16S rRNA library obtained from the environmental sample were de-multiplexed with a Phred score of >Q30 using QIIME (Caporaso et al 2010). The resulting sequences were clustered into operational taxonomic units (OTUs) using UCLUST (Edgar, 2010) based on 97% pairwise identity against Greengenes (gg_13_8_otus). Non-assigned sequences were searched in the Silva database (Release 123) using Mothur (Schloss et al, 2009). OTUs assigned to chloroplast and mitochondria, and sequences that did not have at least five counts across all samples, were removed. Chimera sequences, defined as hybrids between multiple DNA sequences, were identified using Mothur (Schloss et al, 2009) and also removed. We visualized the taxonomic diversity of the JP6 environmental sample, from which strain 106C was isolated, using Metacoder (Foster et al. 2017).

We then estimated taxonomic diversity using the 16S rRNA gene sequences after PCR amplification and Sanger sequencing, for our entire bacterial endophyte collection (*t0* and *t1*). PCR products of 1.4 Kbp in length, obtained using the F27 and R1492 primers (Lane 1991), were cloned and sequenced using the Sanger method (ABI 3730xl). The taxonomic identification was made using Blastn with an initial cut-off e-value of 1e-5 against the SILVA database (Quast et al. 2013) and more than 98% identity. We used the phylogenetic position of the top 10 hits from each search without duplicated matches, to determine both taxonomic diversity and phylogenetic relationships for both amplicons and Sanger. We only report genus-level taxonomy as species determination can be biased by reference 16S rRNA databases.

We used Kraken (http://ccb.jhu.edu/software/kraken/) to assign taxonomic labels to metagenomic DNA sequences based on exact alignment of *k*-mers (Wood et al. 2014). Kraken was used to exclude sequence contaminants from the draft assembly, allowing us to generate a

symbiotic cyanobacteria marker database as reference for future classification. We made a standard Kraken database using NCBI taxonomic information for all bacteria, as well as the bacterial, archaeal and viral complete genomes in RefSeq (October 2016). This database contains a mapping of every *k*-mer in Kraken's genomic library to the lowest common ancestor in a taxonomic tree of all genomes that contain that *k*-mer. We summarized the results in genera-level tables for each metagenome and filtered taxonomy hits that had one or more reads assigned directly to a taxon.

We estimated the efficiency of our sequencing effort with respect to the total possible taxa per metagenome using a previously reported rarefaction method (Hurlbert 1971). To do this we calculated the proportional number of sequences for each metagenome, in which the richness of metagenome-derived OTUs (from herein termed mOTUs) is sub-sampled randomly from the total of the community. The Shannon-Weaver *H'* and Simpson *L* indices for mOTUs from different environments were calculated. In both cases we used the Vegan v2.3-5 package in R (Oksanen J 2015). To visualize shared taxa among metagenomes, and their abundance, we used Cytoscape v3.4.0 (Shannon 2003), where each node and its size represent the abundance of a mOTU, and lines represent shared taxa between metagenomes. The network was made using an interaction matrix, where each of the mOTUs that had more than 14 readings assigned directly by Kraken, was linked to its cognate metagenome.

Reconstruction of phylogenomic relationships. We aligned annotated 16S rRNA sequences trimmed to 1.1 Kbp, using MUSCLE v3.8.31 with default parameters (Edgar 2004). This matrix was used for phylogenetic reconstruction using MrBayes v3.2 (Ronquist et al. 2012) with a gamma distribution type range with 1,000,000 generations. ModelTest (Posada et al. 1998)

showed that Kimura 2 parameters was the best substitution model. To explore major clades in more detail, we estimated individual phylogenies for each of the genera in our main tree and represented them graphically. To do this we first recovered a tree by generating a consensus sequence from all genera within each clade, then a Bayesian phylogeny with a gamma distribution and a million generations (additional generations did not change our results) was reconstructed using MrBayes v3.2 for each individual genus dataset. The resulting trees were edited and sorted with Environment for Tree Exploration Toolkit v3.0.0b35 (Huerta-Cepas et al. 2010) in Python v2.7.6.

To construct a complete phylogeny of Cyanobacteria strains using their whole genomes, we adopted a phylogenomics approach that we termed OrthoCores based on our in-house scripts (https://github.com/nselem/orthocore/wiki, dockerhub: nselem/orthocores). Our method is based on the premise that single gene or even housekeeping multilocus phylogenies are often not enough to resolve the evolutionary history of certain taxonomic groups, and thus total evidence tree evidence approaches must be used (Narechania et al. 2012). We first obtained a set of orthologous groups that are shared among all high-quality genomes, or the 'core genome', excluding paralogs and signatures of expansions (i.e. not all genes from a gene family are part of the Best Bidirectional Hits, BBH set). We define orthologs from the core genome as hits that satisfy that all gene members are BBHs all vs all (fig. s1). To place our samples within the Cyanobacteria phylogeny we used OrthoCores to generate a tree using representative genera from different families sampled across all Cyanobacteria groups. We selected closed and well-annotated genomes, although for Nostocales we used all the publicly available genomes. We obtained representatives from Chroococcidiopsidales, Oscillatoriales, Synechococcales,

Chroococcales, Pleurocapsales, Gloeobacterales and Nostocales publicly available and included the three genome sequences obtained in this study in the phylogeny.

The OrthoCores cut-off e-value was 1e-6, although most genes had an 1e-50 score given our all vs all criterion. Set of sequences were concatenated and aligned using MUSCLE v3.8.31 with default parameters (Edgar 2004), and trimmed by Gblocks (Talavera et al. 2007) with five positions as minimum block length, 10 as a maximum number of contiguous non-conserved positions and only considering positions with a gap in less than 50% of the sequences in the final alignment. The final matrix included 76 taxa and 198 core protein sequences with 45475 amino acids that represent a set of 76 cyanobacterial genomes including our two isolates (T09 and 106C) (table s1, table s2), plus the RF31YmG metagenomic assembly, and *Gloeobacter kilaueensis* JS1 and *Gloeobacter violaceus* PCC 7421 as outgroups. This matrix was used to reconstruct a tree with MrBayes, using a mixed substitution model based on posterior probabilities (aamodel[Wag]1.000) for proteins for one million generations. The tree was visualized with jsPhyloSVG (Smits et al. 2010), and once we identified our samples as Nostocales, we highlighted key traits of these organisms, such as genome size, habitat, and GC content of them and their closest taxa. Metadata were added using PhyD3 (Kreft L et al. 2017). Finally, we examined the level of colinearity among the genomes closest to our samples, using R2cat with default parameters (Husemann et al. 2010).

Genome mining for BGCs. To identify BGCs potentially directing the synthesis of specialized metabolites among selected Cyanobacteria, we annotated the genomes of the closely related strains 106C / RF31YmG and T09 (this study), together with strains PCC73102 and KVJ20 from the database, using antiSMASH (Weber et al. 2015). The predicted BGCs in these strains (88

potential BGCs as suggested by antiSMASH outputs) were manually curated to confirm their completeness (BGCs 3, 6, 18, 25 and 30 did not pass this filter), leading to 45 non-redundant BGCs used as references for subsequent searches in all of the selected genomes. This process allowed us to identify both conserved (4) and *Dioon*-specific (9) BGCs throughout our Nostocales genomes database, generated from publicly available genomes on NCBI (table s3). We considered that a BGC hit was found whenever any given enzyme, plus any other enzyme gene from the cognate BGC, were found in a vicinity of 10 genes on a cyanobacterial genome, within a blast e-value cut-off of 0.001.

The prediction of the chemical structures of the putative specialized metabolites associated with some BGCs, namely, those *Dioon*-specific present in the strains 106C / RF31YmG and T09 (BGCs 5, 9, 12, 20 and 23), was done after domain identification and specificity prediction of adenylation and acyl transfer domains, using the non-ribosomal peptide synthetase (NRPS) and polyketide synthases (PKS) NRPS-PKS server (Bachmann et al. 2009), PRISM (Skinnider et al. 2015) and antiSMASH. The functional inferences of selected BGCs present in these *Dioon*-specific strains, i.e. BGC 12 and 5, in parallel with their structural chemical prediction, was done after in-depth functional annotation of all genes belonging to these BGCs, using standard databases and sequence similarity approaches, together with analysis of the literature.

In addition to analyzing single genomes, including RF31YmG, we also assembled, annotated and mined *de novo*, the complete metagenomes of *t1* and *t2* sub-community co-cultures in an iterative fashion. First, by identifying sequence signatures of biosynthetic enzymes using antiSMASH. And second, by extending the contigs with hits by iterative mapping and assembly. After this process, we only recovered five short signal sequences (less than 3.5 Kbp)

for all metagenomes obtained at *t1*, that are suggestive of enzyme genes that could be part of BGCs (data not shown). Thus, these datasets were not used as part of our comparative genome mining efforts, and only whole genome sequences were mined for BGCs.

Results

Our experimental strategy to characterize the taxonomic and functional diversity of the coralloid root endophytic bacterial community led to hundreds of bacterial isolates obtained directly from the samples that were free from epiphyte contamination, termed (*t0)*; these samples were used to sequence 16S rRNA amplicons directly, and to culture them in axenic media, followed by Sanger 16S rRNA sequencing. And also, samples from enriched sub-communities in oligotrophic (BG11) medium (*t1*) aimed at promoting interactions between members of the coralloid root community. Individual markers and genomic sequences obtained from these isolates captured the taxonomic diversity of endophytes living in the coralloid root (*t2*), including bacteria initially present in low titers (rare) in the original sample. In the following sections we describe the results obtained from this effort, including the overall bacterial taxonomic diversity, and the phylogenomic relationships of the cyanobacteria present in the *Dioon* coralloid roots; as well as insights into metabolic specialization of the cyanobacteria-cycad interaction, after a stepwise genome mining of natural products BGCs.

Dioon coralloid roots show ample endophyte diversity of taxa beyond and within cyanobacteria

Taxa assessment using 16S rRNA (next-generation and Sanger sequencing). As a reference to the bacterial diversity present in the *Dioon* coralloid roots characterized using the abovementioned culture-dependent approaches, we first analyzed the overall diversity by means of single amplicons (fig. 2; table s4). The sample JP6 was selected for this experiment as it represents the original sample from which the strain 106C was isolated (see following sections). Amplification and deep sequencing of the 16S subunit of the rRNA gene using as template total DNA extracted from this sample after treatment for endophyte isolation (*t0*) yielded a total of 5916 sequences and 551 OTUs after initial quality filters, of which only 5587 sequences and 522 OTUs were kept after removal or chimeras, chloroplasts and mitochondria. A total of 86 genera were found, including *Streptomyces, Actinomadura, Burkholderia, Amycolaptosis, Chitinophaga* as the most abundant. Of the Nostocales, only the genera *Nostoc* and *Calothrix* were found, with *Calothrix* as the most abundant. Of the 86 genera, only 13 and 19 overlapped with genera found after analysis of 16S rRNA sequences obtained from axenic samples and mOTUs determined by metagenomics of the sub-community co-cultures, respectively, as discussed next.

Although cultivable bacteria constitute only a subset of the total endophyte biodiversity, our 16S rRNA Sanger sequences yielded 470 isolates grouped into 242 OTUs, distributed in 18 families and 11 bacterial orders, with 27 genera in total, representing most of the known bacterial groups (table 1; table s5). As seen in our 16S rRNA phylogenetic reconstruction (fig. 3), all of our sequences grouped within monophyletic clades, and most trees within each clade show that there are novel isolates that remain to be described, in almost all of the genera found within the cycad coralloid root (see also fig. s2 for a detailed phylogeny). 87% of the taxa identified can be taxonomically classified as diazotrophic plant endophytes. Most OTUs grouped within the genera *Streptomyces, Bacillus, Rhizobium, Stenotrophomonas, Pseudomonas, Mitsuaria,*

Achromobacter and *Burkholderia* (table 1).

We found *Bacillus*, which was previously reported as associated with the outside of the coralloid root; *Streptomyces*, isolated as an epiphyte (Lobakova et al. 2003), which grew on one of our selective media (ISP4); and *Pseudomonas* (Bershova et al. 1968), growing indistinctly in our four semi selective media. We isolated two strains identified as *Nostoc* spp, which also confirms the presence of this genus in the coralloid root (Caiola 1980). We sequenced the genome of one of these *Nostoc* isolates, strain 106C, as discussed in the following sections.

Presumably, if the coralloid root is adapted to specific environments, then species composition, diversity and abundance should differ among collection sites (dry vs humid), and should also be independent of the media on which they were cultivated. Our phylogenetic results show that OTUs are shared with no specific distribution among the various isolation culture media (fig. 3). Some genera occur only in dry (JP; e.g. *Rhizobium*), or only in humid (RF; e.g. *Xanthomonas*) forest environments, with a few genera occurring in both (e.g. *Burkholderia*). Diversity and abundance measured as the Shannon-Weaver and Simpson biodiversity indices (H'=1.50, L=3.8 in the dry environment and H'=1.57, L= 3.76 in the humid environment) are similar (table s6).

Taxa assessment based in sub-community co-cultures characterized with metagenomics. We extracted and sequenced whole-community metagenomic DNA from *t1* and *t2* sub-communities with the aim of enriching for specific interactions in response to oligotrophic (BG11) growth conditions. Basic indicators of metagenome quality are shown in (table s7), and rarefaction curves suggest we captured 40-60% of the microbial community growing in the BG11 media (fig. s3). Metagenomic sequencing of endophytic sub-communities was successful for the

following individuals, grown under the same conditions: two from each of the two environments (JP2, JP6 and RF1, RF3) after 30 days (*t1*); and two after 365 days (*t2*) one from each environment (JP6 and RF3).

With respect to identity and abundance of taxa, in the 30 days sample (*t1*), we found the following (total taxa/genus/most abundant in percentage): JP2, 57/*Rhizobium* (7%); JP6, 69/*Xanthomonas* (2%); RF1, 63/*Nostoc* (3%); and RF3 61/*Xanthomonas* (7%). As expected, we observed different bacterial proportions for the 1-year samples (*t2*): JP6, 69/*Deinococcus* (4%), RF3, 63/*Nostoc* (3%), and in both *t1* and *t2* samples *Nostoc* is present. The most abundant genera from the four metagenomes from humid and dry environments mostly overlap (shown for *t1* in fig. 4 table s8). There are some exceptions that were unique to an environment, which include genera such as *Shewanella*, only found in JP2 from the dry environment, and *Cronobacter*, only found in RF3 in the humid environment. These results support the trends found after analysis of the 16S rRNA Sanger sequences in which the abundant taxa are shared within the analyzed coralloid root samples, with exceptions of a few genera.

There is an overall proportion of 12% Cyanobacteria, including *Anabaena* (Chaudhuri et al. 1931) and *Nostoc* (Thajuddin et al. 2010a), and *Calothrix*, previously reported in *Encephalartos* (Grobbelaar et al. 1987, Huan et al. 1989) and in *Cycas revoluta* (Thajuddin et al. 2010a). We also identified *Caulobacter*, which to our knowledge is the first report of this genus as a cycad endophyte. *Caulobacter* can be found associated to unicellular green algae (Kim Bh et al. 2014) and has been found to enhance growth of *Anabaena* cyanobacteria (Berg et al. 2008).

Of the 86 genera identified with 16S rRNA amplicons, only 13 and 19 overlapped with genera found after analysis of 16S rRNA sequences obtained from axenic samples and mOTUs determined by metagenomics of the sub-community co-cultures, respectively, as discussed next.

The rather low overlap can be explained because our sequencing efforts were not deep enough to recover cyanobacteria-associated OTUs; because taxa presence may fluctuate in the cultures; and/or because Cycanobacteria sequences are too divergent to be captured or there are errors of annotation in the 16S rRNA databases. Despite these issues, we confirmed the occurrence of several of the bacterial endophyte taxonomic groups in the metagenomes, which we also isolated and characterized with 16S rRNA. Overall, the seven genera that were found conserved after these three different strategies are *Bacillus*, *Burkholderia*, *Nostoc*, *Rhizobium*, *Sphingobium*, *Sphingomonas*, and *Streptomyces*.

Dioon **cyanobacteria are a monophyletic group with large genomes within the Nostocales**

Given that Cyanobacteria phylogenetic history is likely reticulated (Zhaxybayeva et al. 2006) we adopted a whole-genome inference approach to explore the phylogenetic position of our cyanobacterial isolates. Using OrthoCores we reconstructed a phylogeny from 198 proteins identified as the core proteome encoded in 76 selected genomes that provide a broad taxonomic coverage and represent good quality sequences (table s2, table s3). We used this phylogeny to map general genome features, including genome size, (G+C)-content, ecological niche and metabolic specialization, as discussed in the following sub-section.

We found lack of monophyly of several Cyanobacteria groups in our whole-genome phylogeny, with the exception of the Nostocales that were monophyletic (fig. s4). Our isolates T09 from *D. caputoi*, and 106C / RF31YmG from *D. merolae,* form a monophyletic clade, and group with *Tolypothrix* sp PCC 7601. The *Dioon*-specific clade is sister to *Calothrix* sp. PCC 7507 and *Fortiea contorta* PCC7126, and this groups together to *Nostoc punctiforme* PCC73102

from the cycad *Macrozamia* and *Nostoc* sp KVJ20 from the liverwort *Blasia pusilla*. Interestingly, the closest genome at the synteny level outside of our own samples is KVJ20, suggesting convergence between the divergent plant niches hosting these cyanobacteria (fig. s5). When examining the overall synteny of our isolates with their closest relatives, we observe that they are more similar to each other, and to *Tolypothrix* sp PCC 7601, and have less overall genome conservation when compared to *Calothrix* sp. PCC 7507, *Fortiea contorta* PCC7126, and surprisingly, to *Nostoc punctiforme* from the cycad *Macrozamia*. Thus, although the genus and species of our isolates remains to be determined, we can conclude that our cyanobacteria endophytes belong to the Nostocales, are closely related despite different host origin, and appear to be specific to *Dioon* (fig. 5, fig s4).

These results contrast with previous studies in which different species of cycads host multiple cyanobacteria and do not form cycad or host-specific clades (Lindblad et al. 1989, Zimmerman et al. 1992, Zheng W et al. 2002), but are congruent with a 16S rRNA phylogeny of Nostocaceae cyanobacteria which shows that hormogonia-producing species symbiotic to *Gunnera* ferns, *Anthoceros*, and cycads, tend to cluster together (Papaefthimiou et al. 2008). Our phylogeny also shows diverging evolutionary trajectories of *Nostoc* species associated with *Dioon*, from those of the free-living Nostocaceae and other *Nostoc* associated with plants. The *Nostoc* free-living PCC 7120 grouped distantly to strains of symbiotic or plant-associated origin, while the *Nostoc azollae* and other plant-associated *Nostoc* are closely related to some *Anabaena* and *Aphanizomenon,* congruently with previous reports (Papaefthimiou et al. 2008).

When we mapped genome size and G+C content traits onto the phylogeny, we found that although there are no differences in G+C content, our coralloid endosymbiont cyanobacteria and related taxa have larger genomes compared to other clades (fig. 5, fig s4). The exception to this

is *Hassallia byssoidea,* a soil-dwelling bacterium that has a genome of similar size to our cycad symbionts, but it is phylogenetically distant. Notably, the other taxa with large genomes in Nostocales include their sister taxa *Tolypothrix* sp PCC 7601, and two other plant symbionts: *Nostoc* sp KVJ20 which lives in special cavities located on the ventral surface of the gametophyte of the Norway liverwort *Blasia pusilla* (Liaimer 2016); and *Nostoc punctiforme* PCC73102, associated with the Australian cycad *Macrozamia* (Meeks et al. 2001).

Identification of BGCs in sub-community metagenomes suggests metabolic specialization of *Dioon* cyanobacteria

Given that large genomes correlate with the ability of bacteria to produce specialized metabolites (Doroghazi et al. 2014), we aimed at exploring the functional specialization of the coralloid root cyanobacteria symbionts by identifying examples of BGCs, putatively directing the synthesis of specialized metabolites. Genome mining of our strains (106C / RF31YmG and T09), together with two closely related plant symbionts previously reported (PCC73102 and KVJ20), revealed in total 88 antiSMASH hits (table s9). After manual curation of these outputs, as described in Methods, a total of 45 non-redundant and well-defined BGCs (table s10) were used to analyze the distribution of these loci among the selected Nostocaceae genomes. This analysis shows that most of these BGCs are uniquely present within plant-associated genomes, with those from the *Dioon*-specific cyanobacteria, accounting for approximately one half (22 BGCs are contributed by the *Dioon*-specific cyanobacteria) (fig. 5). Nodularin, previously associated with other cyanobacteria (Gehringer et al. 2012) was not found.

The analysis of the distribution of these 45 non-redundant BGCs throughout the selected Nostocaceae genomes revealed four well-conserved BGCs, whose distribution goes beyond this clade. First, the well-characterized heterocyst glycolipid (Soriente et al. 1992), termed here BGC 13. Second, the ladderane lipids biosynthetic system, termed here BGC 16, which provides an unusually dense membrane that protects the cell from toxic intermediaries of the anammox reaction (anaerobic ammonium oxidation) and might be a specific adaptation to this metabolic process (Sinninghe Damsté et al. 2002). Third and fourth, respectively, BGCs 8 and 14, putatively directing the synthesis of a terpene and a bacteriocin-lantipeptide of unknown structure. The remaining BGCs could be overall annotated as three systems coding for lantipeptides, specifically BGC 2, 4, 21; and five pathways coding for NRPS-containing systems, of which some may contain additional PKSs, namely, BGC 5, 12, 20, 22 and 23 (tables s10). It should be noted that all BGCs, except BGC 2 and 22, are absent from *Nostoc punctiforme* PCC73102, a strain isolated from an Australian *Macrozamia* that has been recently re-sequenced (Moraes et al. 2017), supporting metabolic specialization in *Dioon*.

After obtaining this general overview, our next step was to translate, when possible, these BGCs into chemical structure predictions, as described in Methods. We focused on those BGCs that given their NRPS or PKS components, but limited to the biosynthetic systems that are *Dioon*-specific, structural predictions are feasible. These include five BGCs out of the nine that are exclusively found in our strains, i.e. BGCs 5, 9, 12, 20 and 23 (fig. 6; fig. s6). Of these biosynthetic systems, BGCs 5 and 12 may provide examples of predictions involving specialized metabolites with functional implications within the coralloid root.

Our sequence-based predictions suggest that BGC 12 may code for an assembly line capable of synthesizing an N-terminal acylated hexapeptide with several modifications, such as

the epimerization of four of its residues, the N-acylation of its second amidic bond, and the reduction of its C-terminal end to yield an aldehyde group (fig. 6). The N and C terminal modifications on this peptide would resemble small peptide aldehyde protease inhibitors, which have been previously reported on cyanobacteria (Fewer et al. 2013). Alternatively, the product of this biosynthetic system may be a siderophore, as iron-related genes were found next to the NRPS coding-genes, which include a reductase domain. The latter observation is significant as previous reports implicate reductase-containing NRPS systems in the synthesis of iron chelators, e.g. myxochelin biosynthesis (Li et al. 2008).

BGC 5, the most interesting of all our predictions, seems to code for a NRPS system putatively directing the synthesis of a tripeptide consisting of valine, isoleucine and tyrosine residues. This prediction also includes informative modifications, such as N-terminal acylation, N-methylation at an amide bond of the isoleucine residue, and modification of the tyrosine residue by a domain of unknown function. A search for peptides containing such modifications, performed with the server PRISM with the feature for de-replication of known chemical structures (Skinnider et al. 2015), directed our attention to nostoginins, specialized metabolites originally isolated from *Nostoc* (Ploutno et al. 2002). Interestingly, the biosynthetic pathway of this natural product remains to be identified, but it is known that nostoginins A and B, as well as their congeners microginins FR1 / SD755 and oscillaginins A / B (Sano et al. 1998), have protease inhibitory activity. In agreement with this observation, BGC 5 includes a peptidase-coding gene (fig. 6).

Thus far, based in both overall and specific genome mining predictions, it is clear that the *Dioon*-specific cyanobionts have distinctive metabolic capabilities that may result from the evolution of these organisms within a diverse microbiome present in the coralloid root. In the

light of these genome-based analyses, strain 106c warrants further biochemical and functional characterization with regards to the occurrence and functional role(s) of the predicted specialized metabolites.

Discussion

Our evidence undoubtedly shows that within the cycad coralloid root there is a diverse bacterial community that includes all of the previously reported Nostocales, other taxa for which their endophytic origin and presence was unclear, and newly reported genera. We support previous morphological observations that showed that an individual cycad plant could hold diverse communities that differ in their taxonomic composition and life strategies (Lobakova et al. 2003). Morphological studies observing mucilaginous material inside the coralloid root (Ow et al. 1999, Baulina et al. 2003) are also congruent with the bacterial consortia we describe. Overall, many shared genera among samples obtained from plants growing in different environmental conditions, including mostly diazotrophic plant endophytes, point toward functional congruence in a taxonomically diverse coralloid root.

In this study we increased our ability to identify a diverse array of organisms using co-cultured bacterial sub-communities (*t1, t2*). The differences in genera identified following this approach in comparison with analysis of diversity using the 16S rRNA gene as a marker, either after Sanger or next-generation sequencing (*t0*) was expected given that environmental sampling and enriched inoculant complement each other, and aim to recover distinct aspects of the microbiome's composition (Cibrián-Jaramillo et al. 2016). These patterns can also be explained by various scenarios: i) rare groups present in low abundance can only be recovered in sub-

community co-cultures on which they increase in biomass; ii) some organisms are fast growers irrespective of media, and will dominate in OTUs, simply by chance, iii) some groups are more media-specific; and/or iv) groups in BG11 (*t1*) are recovered as a result of functional interactions in cyanobacteria-associated groups.

The long-term one-year co-culture (*t2*) allowed us to explore at least some of the aforementioned possibilities, as the initial amount of nitrogen available in these co-cultures became a limiting factor over time. Hence, the establishment of stable sub-communities after one year -with emerging and surviving taxa- suggests that nitrogen fixation is a driving force in the assembly of the coralloid root community. Exploration of the metabolic-driven hypotheses emerging from these observations in different nutrient conditions is required to understand the complexity of the cycad coralloid community. The use of specific nutrients favoring 'keystone' taxa that help recover otherwise rare and less abundant taxa, is a feasible framework to tackle the complicated challenge of characterizing microbial life strategies, and linking them with their phylogenetic lineages (Ho et al. 2017).

We recover a monophyletic placement of our cyanobacterial samples, which confirms a previous single morphological observation of possible specificity among cyanobacteria coralloid root endophytes and their hosts, including *Dioon* (Caiola 1980). This contrasts with other reports that assume lack of specificity based on the observation that different strains isolated from one plant species are capable of infecting phylogenetically distant hosts (Johansson et al. 1994, Papaefthimiou et al. 2008, Whitton 2012). Partial taxon identification and lack of samples from their native habitats could confound the identification of co-evolutionary signatures between plant host and the bacterial microbiome in the latter studies.

Interestingly, our endosymbiont cyanobacterial isolates have larger genome sizes than all other closely related taxa in their phylogeny. This observation contrasts with most bacterial endosymbionts in eukaryotes that show a reduction in genome size compared to free-living relatives (Mccutcheon 2010). Large genome sizes in endosymbionts are usually attributed to a facultative relationship that requires retaining free-living stages (Maclean et al. 2007). In angiosperms facultative symbionts that form nitrogen-fixing root nodules have large genome sizes adapted to shifting from the soil to the plant environment (Normand et al. 2007). Congruently, the only other cyanobacteria cycad symbiont whose genome has been sequenced, *Nostoc punctiforme* from an African cycad *Macrozamia*, has variation within its large genome required to sense and respond to the environment outside and inside the plant (Meeks et al. 2001). Thus, a feasible yet-to-be-tested hypothesis is that the Nostocales taxa associated to the cycad coralloid root have experienced genome expansion driven by selection to initially survive the complexity of the environment from which they are recruited, but retain the complex life cycles involving cell differentiation needed to establish an advantageous relationship with the cycad plant (Meeks et al. 2002), and with other members of the microbiome inside the coralloid root.

The genome synteny observed within the cycad isolates, together with the nine BGCs unique to the *Dioon*-specific cyanobionts reported here, is in agreement with the hypothesis of a specialized community driven by metabolic adaptation. Of the nine unique BGCs found in our isolates, BGC 12 and BGC 5 putatively coding for two small peptide protease inhibitors, or a siderophore, could have a functional bearing on the evolution and biology of the *Dioon*-bacteria symbiosis. Peptide production may imply that proteolysis is involved in the cyanobacteria-cycad interaction, linked to the reconfiguration of the root architecture or the filtering of the

microbiome, as it has been observed in arbuscular mycorrhiza and legumes (Takeda et al. 2007). An obvious starting point to explore this hypothesis would be to confirm the predicted gene-metabolite relationship proposed here, by sequencing the genomes of nostoginin-producing organisms (Ploutno et al. 2002) together with functional characterization. If a siderophore, BGC12 could be mediating bacterial community interactions (Traxler et al. 2015) and/or interactions with the host, adding to the growing notion that symbiotic relations occur under heavy influence of metabolic exchange.

In sum, our study demonstrates the broad taxonomic bacterial diversity of the coralloid root, which combined with monophyletic Nostocales large genomes with at least nine unique BGCs, could be a sign of the constraints imposed by the transition from a facultative symbiotic lifestyle into a specialized organ. As data is gathered from more genomes of bacterial cycad symbionts isolated from different micro-niches of the rhizosphere-coralloid root ecological system, it will be possible to test for co-evolutionary relationships, including horizontal gene transfer between bacteria and the eukaryote host, and other patterns indicative of close evolutionary histories. These patterns will inform the evolutionary and ecological nature of cycad-bacterial interactions, which in turn could direct detailed functional characterization of these plant specialized bacterial communities.

1 **Table 1.** Taxonomic composition of endophytes cultivated from *Dioon* coralloid roots

Phylum	Class	Order	Family	Genus	OTUs
Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	*Mucilaginibacter*	3
	Cytophagia	Cytophagales	Cytophagaceae	*Dyadobacter*	1
Cyanobacteria	No rank	Nostocales	Tolypothrichaceae	***Tolypothrix***	6
			Nostocaceae	***Nostoc***	2
Firmicutes	Bacilli	Bacillales	Bacillaceae	***Bacillus***	16
			Paenibacillaceae	***Paenibacillus***	2
			Staphylococcaceae	*Staphylococcus*	1
Proteobacteria	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	***Rhizobium***	32
				Shinella	2
		Sphingomonadales	Sphingomonadaceae	*Sphingobium*	1
				Sphingomonas	2
			Brucellaceae	*Ochrobactrum*	1
	Betaproteobacteria	Burkholderiales	Alcaligenaceae	***Achromobacter***	33
			Burkholderiaceae	***Burkholderia***	39
			No rank	***Ralstonia***	2
				Mitsuaria	8
			Comamonadaceae	*Variovorax*	1
	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	***Enterobacter***	3
			Erwiniaceae	***Pantoea***	1
		Pseudomonadales	Pseudomonadaceae	***Pseudomonas***	21
		Xanthomonadales	Rhodanobacteraceae	*Luteibacter*	2

			Xanthomonadaceae	***Stenotrophomonas***	35
				Xanthomonas	2
Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	*Microbacterium*	5
		Streptomycetales	Streptomycetaceae	***Streptomyces***	19

[a]Taxa identified in the literature as diazotroph (bold) are shown.

5 **Legends to Figures**

6

7 **Figure 1. Diagram of experimental strategy to capture bacterial diversity (EcoMining).**
8 Coralloid roots from cycads growing naturally in deciduous tropical forests were sampled, and at
9 least for one individual (JP6) the bacterial diversity as a reference was analyzed by means of 16S
10 rRNA single amplicons. Endophytes from the macerated root were isolated, following two
11 strategies: directly from the sample (*t0*) and after enrichment of sub-communities co-cultures in
12 BG11 (*t1*). In both cases, cultivable bacteria were obtained using an array of six different media.
13 Co-cultures were characterized using shotgun metagenomics, and the resulting data were used to
14 select representative genomes from the endophyte culture collection that we mined for functional
15 and evolutionary information.

16

17 **Figure 2. 16S rRNA-amplicon heat tree of JP6 environmental sample.** 16S rRNA-amplicon
18 heat tree of the coralloid roots endosphere as a reference of bacterial diversity (at the genus level)
19 present in a *Dioon merolae* sample from Jiquipilas locality (JP6). Size and color of nodes
20 represents the abundance (number of reads) of genera in the microbiome. The color of an edge
21 represents the number of supertaxa (taxonomic levels). The **#** symbol indicates taxa that are
22 shared with Sanger 16SrRNA OTUs; ***** indicates taxa shared with mOTUs from co-cultures; and
23 the **&** symbol taxa shared between all approaches.

24

25 **Figure 3. Sanger-based 16S rRNA Bayesian phylogeny of cultivated endophytes from**
26 **coralloid roots of *Dioon merolae*.** The external ring refers to two different environments
27 sampled: dry or JP (D - orange) and humid or RF (H - blue) deciduous tropical forests. The inner

28 ring refers to the isolation strategy: directly from the epiphyte sample (*t0* - white) or after
29 enrichment of sub-communities co-cultures (*t1* - gray). Major bacterial groups are highlighted in
30 different colors across the tree. See also fig. s2 for detailed trees, and Table 1 for a summary of
31 these results.
32
33 **Figure 4. Network of taxa co-occurrence from different coralloid root sub-community**
34 **metagenomic samples.** The most abundant genera present in the four metagenomes
35 representative of dry (JP) and humid (RF), coming from *t1*, are represented by circles. Circle
36 diameters are scaled in accordance with the number of reads associated to each genus. The lines
37 connecting the circles represent shared taxa, orange lines correspond to dry environments and
38 blue to wet environments. Similar results were obtained when comparing metagenomes from *t2*
39 (data not shown).
40
41 **Figure 5. Phylogenomic tree of selected Nostocales species and their genomic features and**
42 **BGCs distribution.** Tree reconstructed with 198 conserved proteins to place 106C, T09 and
43 RF31YmG. Habitats for each species are indicated with colored bullets. Bars show a relatively
44 homogeneous (G+C)-content among cyanobacteria, and a trend for larger genomes in plant
45 symbionts including genomes obtained in this study. The closely related organisms showing a
46 sign of metabolic specialization, in accordance with similar ecological niche, are highlighted
47 with a grey box. BGCs that are *Dioon*-specific (light boxes), with chemical structure prediction
48 (*) and functional annotation (black square) are shown.
49

Figure 6. *Dioon*-specific BGCs with structure chemical prediction and functional annotation. Detailed analysis of BGCs 12 and 5 present in the *Dioon*-specific strains 106C and T09 is shown. Genes are shown as colored boxes, the tips of the boxes indicate the direction of their translation. Domain organization, biosynthetic logic and predicted products are indicated below each BGC. Gene functional annotation is provided within the figure.

Data deposition

The draft genomes generated during the current study are available in the GenBank public repository as follows:

SUBID	BioProject	BioSample	Accession	Organism
SUB2297132	PRJNA360300	SAMN06208854	MTAV00000000	*Nostoc* sp. T09
SUB2299096	PRJNA360305	SAMN06208961	MTAW00000000	*Nostoc* sp. 106C
SUB2299173	PRJNA360315	SAMN06209042	MTAX00000000	*Nostoc* sp. RF31YmG

Metagenomes are available at: https://www.ncbi.nlm.nih.gov/bioproject/369422

Supplementary files:

SupplementaryTables1-9.xls

Table S1. Cyanobacteria genomes used in this study

Table S2. Core proteome of cyanobacteria (198 proteins)

Table S3. Nostocales selected genomes

Table S4. Single amplicon 16S rRNA analysis (JP6 sample)

Table S5. List of 470 isolated bacteria characterized with Sanger 16S rRNA

Table S6. Biodiversity indices of Sanger 16S rRNA and OTUs

Table S7. Statistics and details of co-culture metagenomes sequenced

Table S8. Distribution and abundance of mOTUs from the co-culture metagenomes

Table S9. Prediction of BGCs obtained by antiSMASH in closely related organisms

Supplementary Table S10.xls . *Dioon*-specific BGCs and their distribution in closely related organisms

Supplementary Figures:

Figure S1. OrthoCores diagram

Figure S2. Detailed 16S rRNA phylogenetic tree (panels that can be zoomed in)

Figure S3. Rarefaction of metagenomic mOTUs defined by Kraken

Figure S4. Phylogenomic tree of selected Nostocales

Figure S5. Syntenic comparisons among selected genomes

Figure S6. Chemical structure prediction of *Dioon*-specific BGCs without functional annotation (BGCs 9, 20 and 23)

Acknowledgements

This work was supported by CONACyT #169701 and FON.INST./265/2016 (SA-FC2015-2-901) to ACJ. CONACyT #179290 and #177568 to FBG. We acknowledge Rafael Rincón and and José Daniel Hernández-Ríos for their help with preparation of supplementary material. We also thank Juan Palacios and Antonio Hernández for help during field collections, as well as Flor Zamudio for technical lab support.

References

Adnani, N, et al. (2017). Symbiosis-inspired approaches to antibiotic discovery. *Nat Prod Rep.* 34: 784-814.

Aziz, RK, et al. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9: 75.

Bachmann, B, et al. (2009). In silico prediction of microbial secondary metabolic pathways from DNA sequence data. *Methods in Enzymology* 458: 181-217.

Baulina, O, et al. (2003). Atypical cell forms overproducing extracellular substances in populations of cycad cyanobionts. *Microbiology* 72: 701-712.

Berg, KA, et al. (2008). High diversity of cultivable heterotrophic bacteria in association with cyanobacterial water blooms. *ISME J* 3: 314-325.

Bergensen, F, et al. (1986). Nitrogen fixation in coralloid roots of *Macrozamia communis*. L. Johnson. *Aus J Bio Sc* 18: 1135-1142.

Bershova, O, et al. (1968). The interrelations between the blue-green algae -the causative agents of the water 'bloom' - and bacteria. . *'Tsvetenie' Vody*. A. Topanchevsky. Kiev, USSR, Naukova Dumka**:** 159-171.

Bolger, AM, et al. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30: 2114-2120.

Caiola, M (1980). On the phycobionts of the cycad coralloid roots. *New Phytologist* 85: 537-544

Caporaso JG, et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7: 335–336.

Chaudhuri, H, et al. (1931). The coral-like roots of *Cycas revoluta, Cycas circinalis* and *Zamia floridana* and the alga inhabiting them. *J Indian Bot Soc* 10: 43-59.

Cibrián-Jaramillo, A, et al. (2016). Increasing metagenomic resolution of microbiome interactions through functional phylogenomics and bacterial sub-communities. *Frontiers in Genetics* 7: 4.

Costa, J, et al. (1999). Cyanobiont diversity within coralloid roots of selected cycad species. *FEMS Microbiol Ecol* 28: 85-91.

Costa, J, et al. (2004). Sequence based data supports a single *Nostoc* strain in individual coralloid roots of cycads. *FEMS Microbiol Ecol* 49: 481-487.

Costa JL, et al. (2002). Cyanobacteria in Symbiosis with Cycads. *Cyanobacteria in Symbiosis*. Dordrecht, Kluwer Academic Publishers**:** 195–205.

Cuddy, W, et al. (2012). Comparative analysis of cyanobacteria in the rhizosphere and as endosymbionts of cycads in drought-affected soils. *FEMS Microbiol Ecol* 80: 204-215.

De Luca, P, et al. (1981). *Dioon meroale* (Zamiaceae), a new species from Mexico *Brittonia* 33: 179-185.

Doroghazi, J, et al. (2014). A Roadmap for Natural Product Discovery Based on Large-Scale Genomics and Metabolomics. *Nature Chemical Biology.* 10: 963-968.

Edgar, RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792-1797.

Edgar, RC (2010). Search and clustering orders of magnitude faster than BLAST. 187 *Bioinformatics* 26(19):2460-2461.

Fewer, DP, et al. (2013). New Structural variants of aeruginosin produced by the toxic bloom forming cyanobacterium *Nodularia spumigena*. *PLoS ONE* 8: e73618.

Foster Z, Sharpton T and Grünwald N (2017). "Metacoder: An R package for visualization and manipulation of community taxonomic diversity data." *PLOS Computational Biology*, 13: 1-15.

Gehringer, M, et al. (2012). Nodularin, a cyanobacterial toxin, is synthesized in planta by symbiotic *Nostoc* sp. *The ISME Journal* 6: 1834–1847.

Gehringer, M, et al. (2010). Host selection of symbiotic cyanobacteria in 31 species of the Australian cycad genus: *Macrozamia* (Zamiaceae). *Molecular Plant-Microbe Interactions* 23: 811-822.

Grobbelaar, N, et al. (1987). The identification of the coralloid root endophytes of the southern African cycads and the ability of the isolates to fix dinitrogen. *South African Journal of Botany* 53: 111-118.

Grove, T, et al. (1980). Effects of fire on the growth, nutrient content and rate of nitrogen fixation of the cycad *Macrozamia riedlei*. *Australian Journal of Botany* 28: 271-281.

Halliday, J, et al. (1976). Symbiotic nitrogen fixation by blue algae in the cycad *Marozamia riedlei*: Physiological characteristics and ecological significance. *Aus J Plant Phys* 3: 349-358.

Ho, A, et al. (2017). Revisiting life strategy concepts in environmental microbial ecology. *FEMS Microbiology Ecology* 93: 1-14.

Huan, T, et al. (1989). Isolation and characterization of endosymbiotic *Calothrix* (Cyanophyceae) in *Encephalartos hildenbrandii* (Cycadales). *Phyocologia* 28: 464-468.

Huerta-Cepas, J, et al. (2010). ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* 11: 24.

Hurlbert, S (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52: 577-586.

Husemann, P, et al. (2010). r2cat: synteny plots and comparative assembly. *Bioinformatics* 26: 570–571.

Johansson, C, et al. (1994). Reconstitution of the symbiosis of *Gunnera manicata* Linden: cyanobacterial specificity. *New Phytologist* 126.

Kim BH, et al. (2014). Role of *Rhizobium*, a plant growth promoting bacterium, in enhancing algal biomass through mutualistic interaction. *Biomass and Bioenergy* 69: 95-105.

Kreft L, et al. (2017). PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics*: doi: 10.1093/bioinformatics/btx1324.

Lane, DJ (1991). 16S/23S rRNA sequencing. *Nucleic acid techniques in bacterial systematics*. E. S. a. M. Goodfellow. New York John Wiley & Sons**:** 115-175.

Lázaro-Zermeño, JM, et al. (2012). Historia natural de *Dioon merolae* (Zamiaceae) en Chiapas, México. *Botanical Sciences* 90: 73-87.

Li, H, et al. (2009). Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics* 25: 1754-1760.

Li, Y, et al. (2008). Myxochelin biosynthesis: direct evidence for two- and four-electron reduction of a carrier protein-bound thioester. *J. Am. Chem. Soc* 130: 7554–7555.

Liaimer, A, Jensen, JB, & Dittmann, E. (2016). A genetic and ghemical perspective on symbiotic recruitment of Cyanobacteria of the genus *Nostoc* into the Host Plant *Blasia pusilla* L. *Frontiers in Microbiology* 7.

Lindblad, P, et al. (1989). Microbiology. *Symbiosis*: 20- 24.

Lobakova, E, et al. (2003). Microbial complexes occurring on the apogeotropic roots and in the rhizosphere of cycad plants. *Microbiology* 72: 628.

MacLean, A, et al. (2007). Genomes of the Symbiotic Nitrogen-Fixing Bacteria of Legumes. *Plant Physiol.* 144: 615-622.

McCutcheon, J (2010). The bacterial essence of tiny symbiont genomes. *Curr Opin Microbiol* 13: 73-78.

Meeks, J (2009). Physiological adaptations in nitrogen-fixing *Nostoc*-plant symbiotic associations. *Prokaryotic Symbionts in Plants*. K. Pawlowski. Berlin, Springer-Verlag**:** 181–205.

Meeks, J, et al. (2002). Cellular differentiation in the cyanobacterium *Nostoc punctiforme*. *Arch Microbiol* 178: 395–403.

Meeks, J, et al. (2001). An overview of the genome of *Nostoc punctiforme,* a multicellular, symbiotic cyanobacterium. *Photosynthesis Research* 70: 85-106.

Moraes, L, et al. (2017). Resequencing and annotation of the Nostoc punctiforme ATTC 29133 genome: facilitating biofuel and high-value chemical production. *AMB Express* 7: doi:10.1186/s13568-13017-10338-13569.

196 Narechania, A, et al. (2012). Random Addition Concatenation Analysis: a novel approach to the
197 exploration of phylogenomic signal reveals strong agreement between core and shell genomic
198 partitions in the cyanobacteria. *Genome Biology and Evolution* 4: 30-43.

199 Normand, P, et al. (2007). Genome characteristics of facultatively symbiotic *Frankia* sp. strains
200 reflect host range and host plant biogeography. *Genome Res* 17: 7-15.

201 Norstog, KJ, et al. (1997). *The Biology of the Cycads*. New York, Cornell University Press.

202 Oksanen J, et al. (2015). Vegan: community ecology package. R Packag. version *2*: 264.

203 Ow, M, et al. (1999). Reconstitution of a cycad-cyanobacterial association. *Symbiosis* 27: 125-
204 134.

205 Papaefthimiou, D, et al. (2008). Diversity and host specificity of *Azolla* cyanobionts. *J Phycol*
206 44: 60-70.

207 Ploutno, A, et al. (2002). Modified peptides from a water bloom of the cyanobacterium *Nostoc*
208 sp. *Tetrahedron* 58: 9949-9957.

209 Posada, D, et al. (1998). ModelTest: testing the model of DNA substitution. *Bioinformatics.*
210 *1998* 14: 817-818.

211 Quast, C, et al. (2013). The SILVA ribosomal RNA gene database project: improved data
212 processing and web-based tools. *Nucl Acids Res* 41: D590--D596.

213 Ronquist, F, et al. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model
214 Choice Across a Large Model Space. *Systematic Biology* 61: 539–542.

215 Sano, T, et al. (1998). A 3-amino-10-chloro-2-hydroxydecanoic acid-containing tetrapeptide
216 from Oscillatoria agardhii. *Phytochemistry* 44: 1503-1505.

217 Schloss PD, et al. (2009). Introducing mothur: Open-Source, Platform-Independent, Community-
218 Supported Software for Describing and Comparing Microbial Communities. *Applied and*
219 *Environmental Microbiology*. 75:7537-7541.
220
221 Shannon, P, et al. (2003). Cytoscape: a software environment for integrated models of
222 biomolecular interaction networks. *Genome Research* 13: 2498–2504.

223 Sinninghe-Damsté, JS, et al. (2002) Linearly concatenated cyclobutane lipids form a dense
224 bacterial membrane. *Nature* 419: 708-7.

225 Skinnider, M, et al. (2015). Genomes to natural products PRediction Informatics for Secondary
226 Metabolomes (PRISM). *Nucleic Acids Res* 43: 9645–9662.

227 Smits SA, et al. (2010). jsPhyloSVG: A Javascript Library for Visualizing Interactive and
228 Vector-Based Phylogenetic Trees on the Web. *PLoS ONE* 5: e12267.
229 doi:12210.11371/journal.pone.0012267.

230 Soriente, A, et al. (1992). Structure of the "heterocyst glycolipids" of the marine cyanobacterium
231 *Nodularia harveyana*. *Tetrahedron* 48: 5375–5384.

232 Takeda, N, et al. (2007). Proteases in plant root symbiosis. *Phytochemistry* 68: 111-121.

233 Talavera, G, et al. (2007). Improvement of phylogenies after removing divergent and
234 ambiguously aligned blocks from protein sequence alignments. *Systematic Biology* 56: 564-577.

235 Thajuddin, N, et al. (2010a). Morphological and genetic diversity of symbiotic cyanobacteria
236 from cycads. *J Basic Microbiol* 50: 254-265.

237 Traxler, M, et al. (2015). Natural products in soil microbe interactions and evolution. *Nat Prod*
238 *Rep.* 32: 956-970.

239 Weber, T, et al. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of
240 biosynthetic gene clusters. *Nucl. Acids Res.* 43: W237-W243

241 Whitton, BA (2012). *Ecology of Cyanobacteria II: Their Diversity in Space and Time*, Springer
242 Science & Business Media.

243 Wood, D, et al. (2014). Kraken: ultrafast metagenomic sequence classification using exact
244 alignments. *Genome Biology* 15: R46.

245 Yamada, S, et al. (2012). Genetic diversity of symbiotic cyanobacteria in *Cycas revoluta*
246 (Cycadaceae). *FEMS Microbiol Ecol* 81: 696-706.

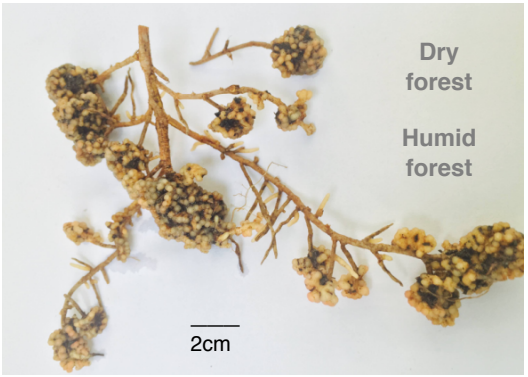
247 Zerbino, DR, et al. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn
248 graphs. *Genome Research* 18: 821-829.

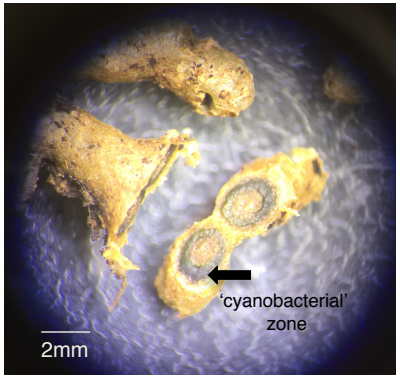
249 Zhaxybayeva, O, et al. (2006). Phylogenetic analyses of cyanobacterial genomes: Quantification
250 of horizontal gene transfer events. *Genome Research* 16: 1099–1108.

251 Zheng W, et al. (2002). High cyanobacterial diversity in coralloid roots of cycads revealed by
252 PCR fingerprinting. *FEMS Microbiol Ecol* 40: 215-222.

253 Zimmerman, WJ, et al. (1992). Cyanobiont diversity within and among cycads of one field site.
254 *Canadian J Microbiol* 38: 1324-1328.

255 Zvyagintsev, D, et al. (2010). Morphological and physiological modifications of cyanobacteria in
256 experimental cyanobacterium-actinomycete associations. *Microbiology* 79: 314-320.
257



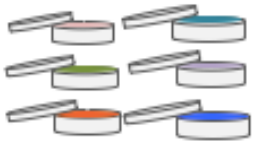


Epiphyte-free DNA sample

Amplicon 16S rRNA sequencing

t0

Strain isolation



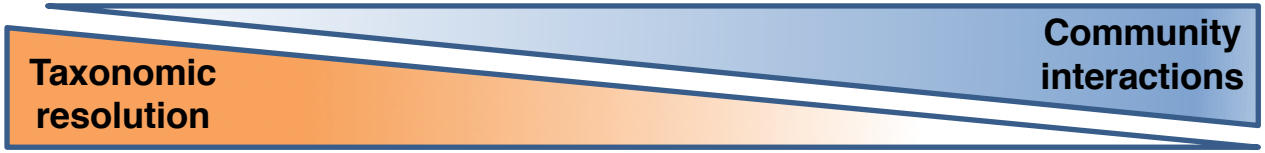
Six selective media

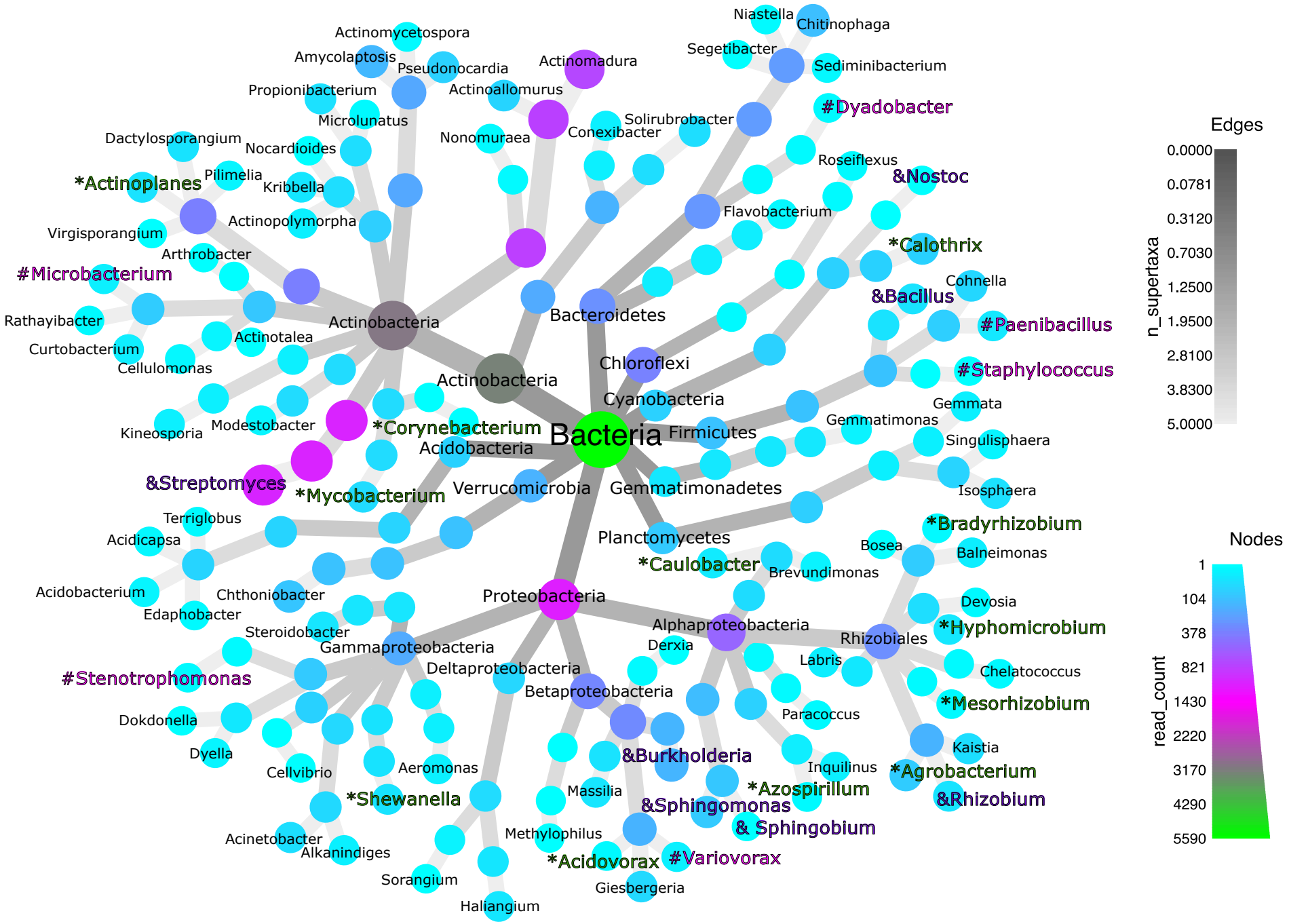
16S rRNA Sanger
Whole genome sequencing

Sub-community co-cultures

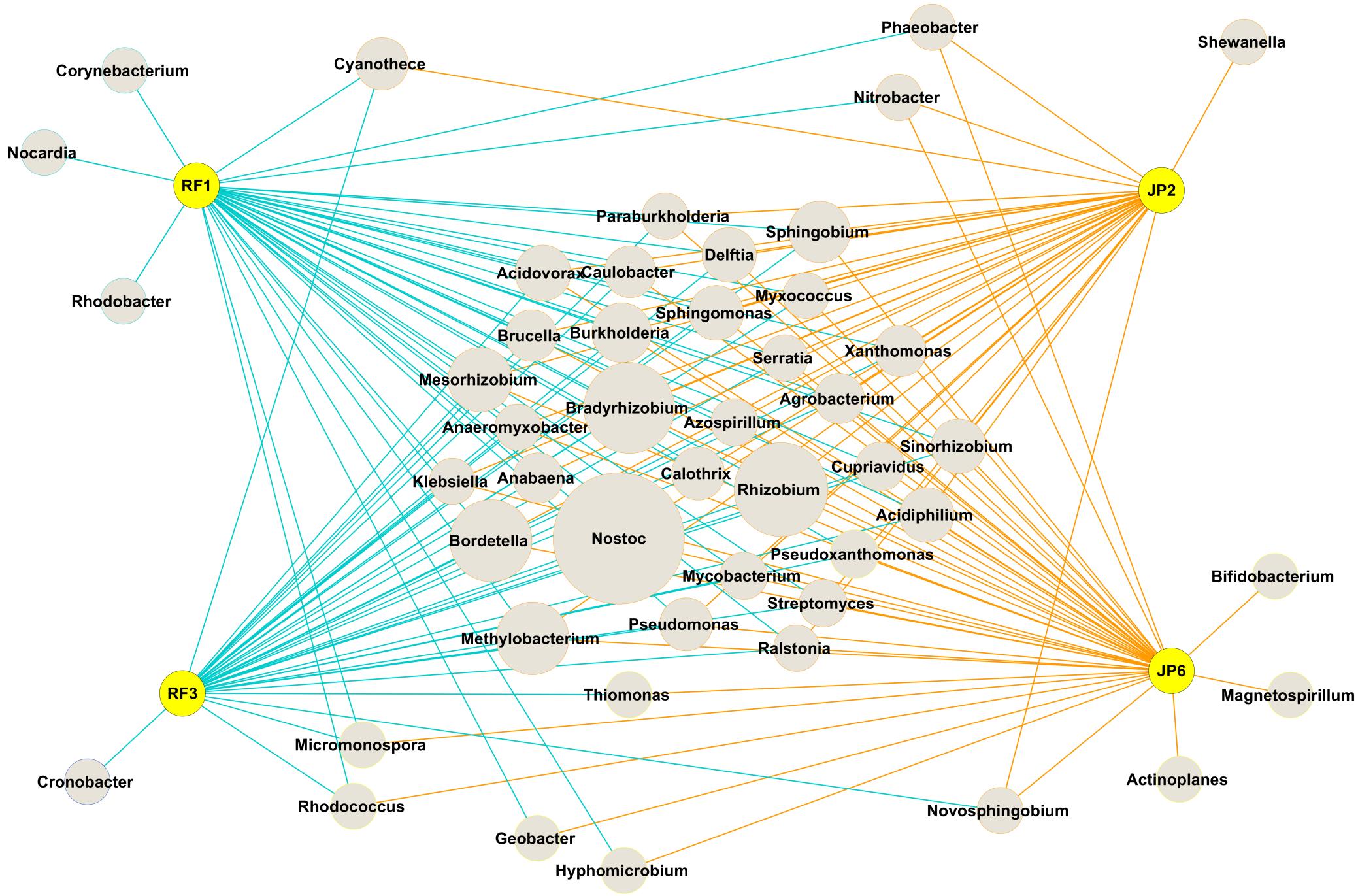
t1, t2..tn

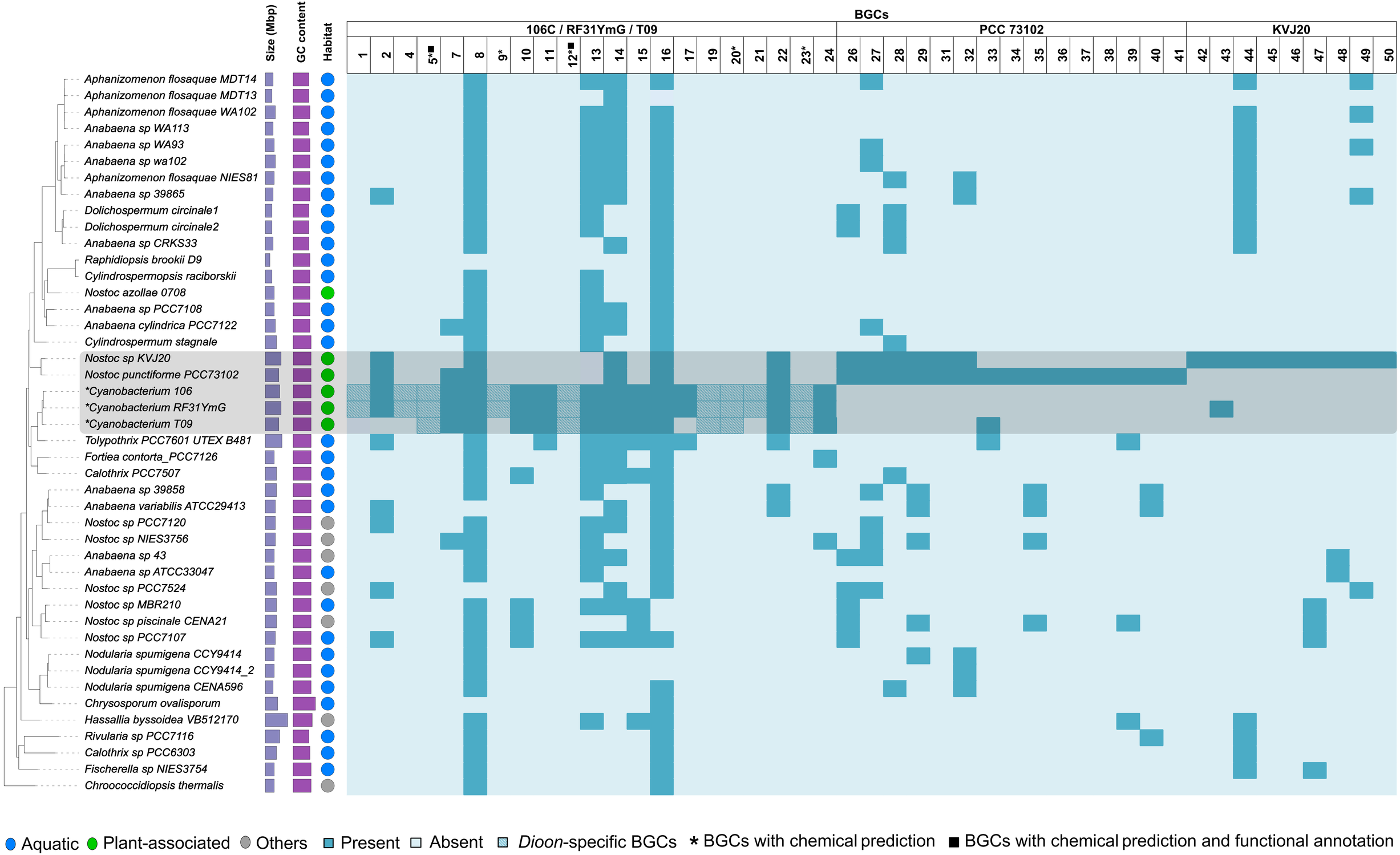
Shotgun metagenomic Sequencing (mOTUs)



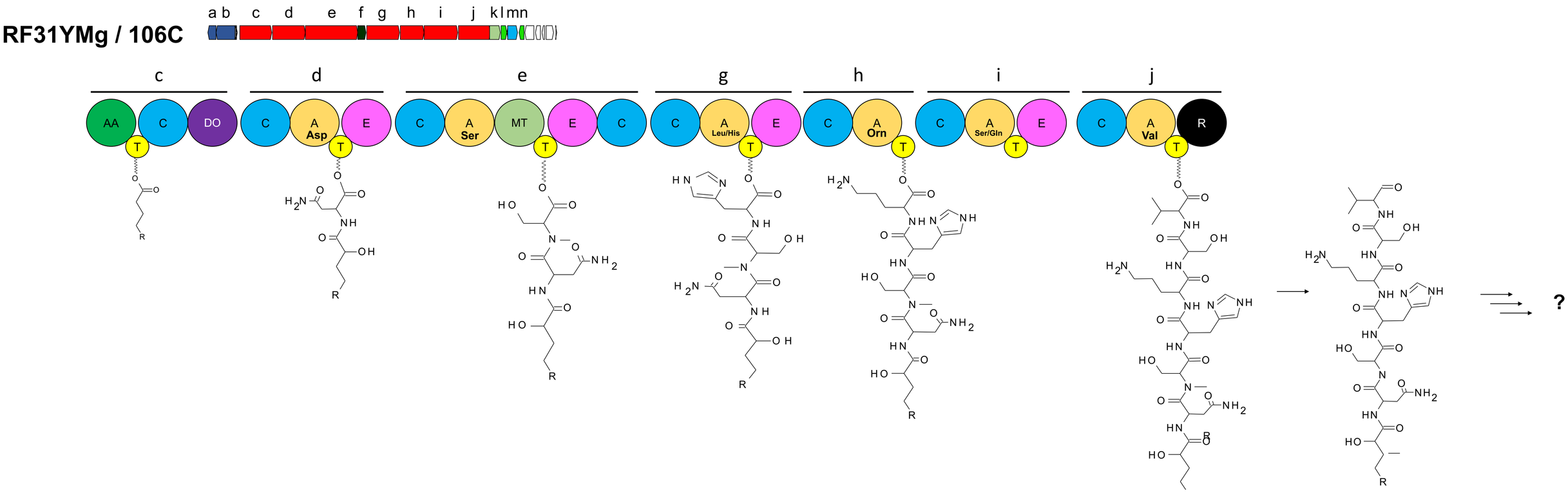
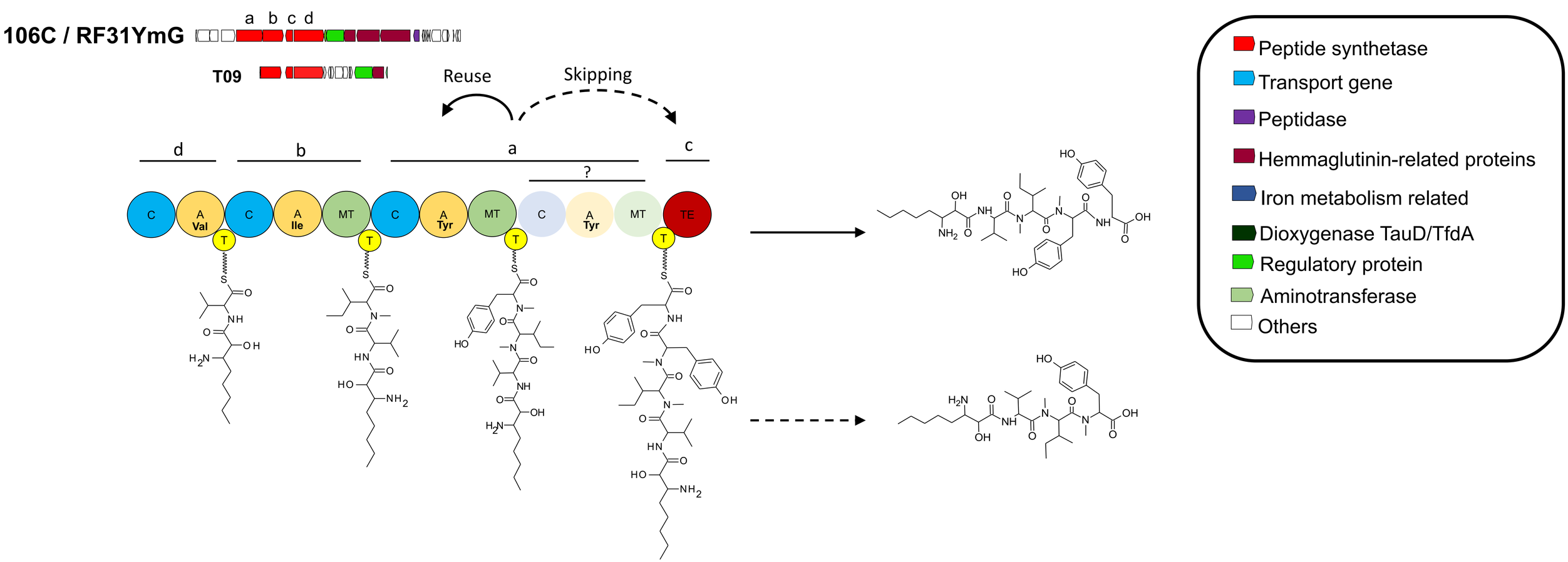

Bacteria
Actinobacteria
Actinobacteria
Bacteroidetes
Chloroflexi
Cyanobacteria
Firmicutes
Acidobacteria
Verrucomicrobia
Gemmatimonadetes
Planctomycetes
Proteobacteria
Alphaproteobacteria
Gammaproteobacteria
Deltaproteobacteria
Betaproteobacteria
Rhizobiales
Actinomycetospora
Amycolaptosis
Pseudonocardia
Propionibacterium
Actinoallomurus
Actinomadura
Microlunatus
Nonomuraea
Conexibacter
Solirubrobacter
Niastella
Chitinophaga
Segetibacter
Sediminibacterium
#Dyadobacter
Roseiflexus
&Nostoc
Flavobacterium
*Calothrix
Cohnella
&Bacillus
#Paenibacillus
#Staphylococcus
Gemmata
Gemmatimonas
Singulisphaera
Isosphaera
*Bradyrhizobium
Bosea
Balneimonas
Brevundimonas
Devosia
*Hyphomicrobium
Chelatococcus
*Mesorhizobium
Kaistia
*Agrobacterium
&Rhizobium
Inquilinus
Paracoccus
Derxia
Labris
*Azospirillum
& Sphingobium
&Sphingomonas
&Burkholderia
Massilia
Methylophilus
*Acidovorax
#Variovorax
Giesbergeria
Haliangium
Sorangium
Alkanindiges
Acinetobacter
*Shewanella
Aeromonas
Cellvibrio
Dyella
Dokdonella
#Stenotrophomonas
Steroidobacter
Edaphobacter
Chthoniobacter
Acidobacterium
Acidicapsa
Terriglobus
&Streptomyces
*Mycobacterium
*Corynebacterium
Kineosporia
Modestobacter
Cellulomonas
Curtobacterium
Rathayibacter
Actinotalea
#Microbacterium
Arthrobacter
Virgisporangium
Actinopolymorpha
*Actinoplanes
Pilimelia
Kribbella
Dactylosporangium
Nocardioides
Edges
0.0000
0.0781
0.3120
0.7030
1.2500
1.9500
2.8100
3.8300
5.0000
n_supertaxa
Nodes
1
104
378
821
1430
2220
3170
4290
5590
read_count

Achromobacter
Ralstonia
Burkholderia
Mitsuaria
Variovorax
Pseudomonas
Pantoea
Enterobacter
Stenotrophomonas
Xanthomonas
Lutelibacter
Rhizobium
Shinella
Ochrobactrum
Sphingomonas
Mucilaginibacter
Dyadobacter
Tolypothrix
Nostoc
Bacillus
Paenibacillus
Streptomyces
Microbacterium
Aquifex
0.25


Phaeobacter
Shewanella
Corynebacterium
Cyanothece
Nitrobacter
Nocardia
RF1
JP2
Paraburkholderia
Sphingobium
Delftia
Acidovorax
Caulobacter
Myxococcus
Rhodobacter
Sphingomonas
Brucella
Burkholderia
Serratia
Xanthomonas
Mesorhizobium
Bradyrhizobium
Agrobacterium
Anaeromyxobacter
Azospirillum
Sinorhizobium
Cupriavidus
Klebsiella
Anabaena
Calothrix
Rhizobium
Acidiphilium
Bordetella
Nostoc
Pseudoxanthomonas
Mycobacterium
Bifidobacterium
Streptomyces
Pseudomonas
Methylobacterium
Ralstonia
JP6
RF3
Thiomonas
Magnetospirillum
Micromonospora
Cronobacter
Actinoplanes
Rhodococcus
Novosphingobium
Geobacter
Hyphomicrobium


Size (Mbp)
GC content
Habitat
BGCs
106C / RF31YmG / T09
PCC 73102
KVJ20
1
2
4
5*■
7
8
9*
10
11
12*■
13
14
15
16
17
19
20*
21
22
23*
24
26
27
28
29
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
Aphanizomenon flosaquae MDT14
Aphanizomenon flosaquae MDT13
Aphanizomenon flosaquae WA102
Anabaena sp WA113
Anabaena sp WA93
Anabaena sp wa102
Aphanizomenon flosaquae NIES81
Anabaena sp 39865
Dolichospermum circinale1
Dolichospermum circinale2
Anabaena sp CRKS33
Raphidiopsis brookii D9
Cylindrospermopsis raciborskii
Nostoc azollae 0708
Anabaena sp PCC7108
Anabaena cylindrica PCC7122
Cylindrospermum stagnale
Nostoc sp KVJ20
Nostoc punctiforme PCC73102
*Cyanobacterium 106
*Cyanobacterium RF31YmG
*Cyanobacterium T09
Tolypothrix PCC7601 UTEX B481
Fortiea contorta_PCC7126
Calothrix PCC7507
Anabaena sp 39858
Anabaena variabilis ATCC29413
Nostoc sp PCC7120
Nostoc sp NIES3756
Anabaena sp 43
Anabaena sp ATCC33047
Nostoc sp PCC7524
Nostoc sp MBR210
Nostoc sp piscinale CENA21
Nostoc sp PCC7107
Nodularia spumigena CCY9414
Nodularia spumigena CCY9414_2
Nodularia spumigena CENA596
Chrysosporum ovalisporum
Hassallia byssoidea VB512170
Rivularia sp PCC7116
Calothrix sp PCC6303
Fischerella sp NIES3754
Chroococcidiopsis thermalis
Aquatic
Plant-associated
Others
Present
Absent
Dioon-specific BGCs
* BGCs with chemical prediction
■ BGCs with chemical prediction and functional annotation

A. BGC 12: Putative siderophore or protease inhibitor



B. BGC 5: Nostoginin prediction