Title page

The cycad coralloid root contains a diverse endophytic bacterial community with novel biosynthetic gene clusters unique to its microbiome

Pablo Cruz-Morales[1,2], Antonio Corona-Gómez[2], Nelly Selem-Mójica[1], Miguel A. Perez-Farrera[3], Francisco Barona-Gómez[1], 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.

Pablo Cruz Morales: cruzmoralesp@gmail.com

José A. Corona-Gómez: jose.corona@cinvestav.mx

Nelly Selem-Mojica: nselem84@gmail.com

Miguel Perez-Farrera: miguel.perez@unicach.mx

Francisco Barona-Gómez: francisco.barona@cinvestav.mx

*Angelica Cibrian-Jaramillo: angelica.cibrian@cinvestav.mx

Corresponding author

Keywords

Cycad, *Dioon*, coralloid root, microbiome, sub-community co-culture, cyanobacteria, *Nostoc*, specialized metabolites.

Abstract

Cycads are the only gymnosperms and ancient seed plants that have evolved a specialized coralloid root to host endophytic bacteria. There are no studies exploring the taxonomic, phylogenetic and functional diversity of the bacterial endophyte microbiome of this 300 million-year old symbiosis. We provide a genomic characterization of the cycad coralloid root microbiome of the Mexican cycad *Dioon merolae* collected from their natural environment. We employed a co-culture-based metagenomics experimental strategy jointly with phylogenomic analyses to reveal both predominant and rare bacteria, to capture biological diversity, and also the presence of biosynthetic gene clusters associated with specialized metabolites. Most taxa were identified as diazotroph plant endophytes that include undescribed taxa and at least 27 genera belonging to 17 bacterial families in addition to Cyanobacteria. Three cyanobacteria genomes obtained from our samples formed a monophyletic group, suggesting a level of specialization characteristic of co-evolved symbiotic relationships. This contrasted with our finding of their large genome sizes and their broad biosynthetic potential, distinctive of facultative endosymbionts of complex alternative lifestyles. Nine out of 23 novel biosynthetic gene clusters identified after detailed genome mining are specific to these coralloid root endophytes, including a NRPS system predicted to direct the synthesis of nostoginins, protease inhibitors whose biosynthetic pathway remains to be discovered. Combined, our results show that the highly diverse taxonomic composition of the coralloid root and

its biosynthetic repertoire, correlate more with a degree of specificity to the cycad plant host than to other closely related plant endosymbionts or to the environment. We support the growing notion that plant-bacteria relations occur under heavy influence of chemical and genomic interactions, and we add to the understanding of the evolution of cycad-bacteria microbiome, with a bearing on bioprospecting of natural products for drug discovery and other applications.

Background

Cycads (Cycadales) are the only early seed plants and the only gymnosperms that develop coralloid roots, a specialized root dichotomous and coral-like in appearance typically growing above ground, which acquires and maintains bacteria [1] (**Fig. 1**). The coralloid root is present in all cycad lineages, likely due to its adaptive value as a significant source of fixed nitrogen for the plant [2]. In natural habitats coralloid roots appear in the most vulnerable early life stages [3], or as adults in habitats with poor or inaccessible nutrients [4] such as sand dunes, sclerophyll forests, steep rock outcrops with high exposure to salt, and lowland forests with recurrent fires. The cycad coralloid root is probably a key trait that enabled cycads to thrive and adapt to novel environments for millions of years.

Coralloid root endophytes have been studied since the 19th century ([5] and references therein). However, most studies have focused on resolving the biology or taxonomy of the Cyanobacteria, and most samples have been collected from botanic garden collections or grown in greenhouses, typically outside of the cycad host natural range [6-12]. Anatomical studies have shown the presence of mucilaginous or protein-rich material that hosts other unidentified bacterial groups [5, 13, 14], with only a few specific bacterial taxa suggested [15-19]. Studies testing for the specificity of cyanobacteria and the cycad host have been conducted in plants collected outside of their native distribution, with contrasting results regarding the specialization of coralloid root symbionts [5, 15, 20]. Moreover, the handful of field-based studies from wild cycad populations, focused only on cyanobacteria identified with molecular markers [11, 21], and show that diversity ranges from a single cyanobacteria strain inside an individual root, to diverse species complexes among roots, and within and among various cycad genera. Studies on the origin and transmission of bacterial endophytes are also inconclusive [12], thus the degree of cycad-bacteria co-evolution in this symbiotic system remains a mystery.

In addition to nitrogen fixation there have been suggestions of additional -unknown- roles for the coralloid root, but there is no clear evidence of its broader function to date [5]. Likewise, various chemical, physical and physiological mechanisms appear to regulate the cycad-bacteria interaction [22, 23], but no genes involved in novel specialized metabolite production in the light of the symbiosis have been identified. In all, the taxonomic composition and the function of the cycad coralloid root microbiome, defined as the bacteria living inside this specialized organ plus their genes and products, remains undescribed almost entirely. What is more, the evolutionary history of the microbiome within a *ca.*300 million-year-old symbiotic plant-bacteria relationship is still incipiently explored.

Our goal in this study is to investigate the microbiome of the coralloid roots of *Dioon merolae* [24]. *Dioon merolae* is a long-lived, entomophilous, dioecious, and arborescent cycad native to Mexico [25]. We collected coralloid root samples from wild populations in two different habitats from its natural range, currently distributed in moderate population sizes of a few hundreds of individuals throughout Chiapas and Oaxaca in the south of Mexico [25]. The availability of whole-genome and metagenomic sequencing enabled us to provide insights on the diversity and phylogenetic distribution of its endophytes and their cycad-related specialized functions.

The presence of uniquely specialized metabolites in the cycad coralloid root microbiome was of particular interest to us because they may be a result of co-evolution between the cycad host and the endophyte bacterial community. Bacteria have dynamic genomic diversity and the capacity to synthesize specialized metabolites with overwhelming chemical diversity that are produced to cope with biotic and abiotic pressures [26]. Bacteria codify specialized metabolites in rapidly evolving genetic units called biosynthetic gene clusters (BGCs) of about 25-40 Kbp. 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, in a known host and environment. From a more anthropocentric view, conserved BGCs of the coralloid root bacterial endophytes may also be of interest as a source of novel natural products for drug discovery.

To overcome technical difficulties in characterizing the breadth of microbial diversity in environmental samples, we used an enrichment co-culture strategy of sub-communities obtained from the original sample [27]. We employed complementing microbiological, genomic and metagenomic sequencing, and phylogenomic approaches to characterize the coralloid microbiome’s taxonomic diversity and gain insights into its function. Our study is the first to characterize the taxonomy and function of the coralloid root beyond cyanobacteria, providing a glimpse into the evolutionary history of the cycad-bacteria coralloid root system.

Methods

Overall strategy. We used a combined co-culture, metagenomics and phylogenomic strategy to detect and measure taxonomic diversity, phylogenetic relationships and biosynthetic potential in the endophytes of the cycad coralloid root, as previously described under the term of EcoMining [27] (**Fig. 1**). In this approach, we grew and isolated bacteria from environmental samples using a diverse set of media that aim to capture all possible cultivable bacterial diversity (*t0*). Simultaneously, we enriched the same samples in co-cultures grown under specific conditions for cyanobacteria using BG11 media. In addition to this autotrophic bacterial group, this approach captures other bacterial groups that have interactions with cyanobacteria, present in the original sample at low titers. We allowed the co-culture to grow over time and sampled it after one month (*t1*) and at the end of a year (*t2*) to capture organisms that depend on other bacteria of the community to grow. We isolated axenic bacteria (*t0* and *t1*) and sub-communities in co-cultures (*t1* and *t2*), and reconstructed phylogenetic relationships and assessed taxonomic diversity, using 16S rRNA and metagenomic OTUs (mOTUs) data, respectively. Furthermore, genomes of isolated endophytes were obtained and thoroughly mined together with metagenomes for BGCs potentially directing the synthesis of specialized metabolites.

Field collections. We sampled coralloid roots from two wild cycad populations previously reported [25]. In March of 2014 we sampled from two sites in deciduous tropical forests, at Jiquipilas, Mexico (JP or dry; Lat 16° 37' 26.87''N, Long 93° 34' 34.64" O) at 560m above mean sea level, with an average annual precipitation of 320 mm and average annual temperature of 18 °C; and Raymundo Flores Mexico (RF or humid; Lat 16° 3' 26.75''N, Long 93° 35' 55.26" O) at 900m above mean sea level, with 2500 mm and 25°C annual average precipitation and temperature, respectively. In some cycad plants, coralloid roots were easily visible aboveground, while in others we dug 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 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 three samples of three 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 six coralloid root samples (JP1, JP2, JP6 and RF1, RF3 and RF9), and stored the remaining samples at -80 °C for subsequent studies. When DNA samples from these individuals were pooled for sequencing purposes they are referred to as JPPOOL or RFPOOL, respectively. 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 other contamination sources. 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 (H_2O_2), 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 five media described below. Lack of growth in the last wash was considered a negative control, and 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 (*t0*), and sampling from the enriched co-cultures (*t1*), as described in the following sections.

Bacterial isolation. To isolate bacteria from field samples before (*t0*) and after (*t1*) enrichment, macerated roots or co-culture broth were used as inoculant, respectively. Loss of some bacterial groups that were present in the sample collected from the environment (*t0*) is expected. However, after enrichment (*t1*) we recover bacteria that were 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) [27]. 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 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) [28]; 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 [29, 30]; 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) [31]; 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 [32]. 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 sub-communities cultivation. We took 100 µl of the macerated roots that passed the negative growth controls after the final washing step (i.e. samples JP1, JP2, JP6 and RF1, RF3 and RF9, which also lead to JPPOOL and RFPOOL samples as described next), and inoculated 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 is expected to diminish over time. We used one non-selective eutrophic medium, i.e. enriched *Caulobacter* + mannitol medium (medium No. 6), which we expected to favor growth of the majority of the generalist taxa in the root bacterial community; and one selective oligotrophic medium, i.e. BG11 (medium No. 4). This medium lacks a carbon source but contains a limited amount of inorganic nitrogen. 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. Eutrophic cultures were sampled after 72 hours, and their DNA extracts pooled (JPPOOL and RFPOOL), whereas sampling of the oligotrophic co-cultures was done after 1 month (*t1*) and 1 year (*t2*), and treated independently. Moreover, bacterial isolates were only obtained for the former, whereas for both time points shotgun metagenomics were obtained, allowing for genome mining of specialized metabolites.

Genomics and shotgun metagenomics. To sequence metagenomes from enriched sub-community co-cultures, we collected their biomass by centrifugation (6000 RPM during 15 minutes) and used for DNA extraction using a CTAB-phenol chloroform standard protocol. Isolate 106C, obtained from sample JP6, and isolate T09, obtained from coralloid roots of *Dioon caputoi* from an unrelated environment (Xeric shrubland, Tehuacan valley, Mexico), were both grown on BG11 plates. Genomic DNA from these cultures was obtained with exactly the same CTAB-phenol chloroform protocol. Genomic and metagenomic DNA samples were processed with truseq nano kit Q28 and were 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 y RF3-1yr). The reads for each library were filtered with fastQ and trimmed using Trimommatic version 0.32 [33], and assembled using Velvet 1.2.10 [34] with different *k*-mers: the assemblies with the largest length and the smaller number of contigs were selected and annotated using RAST [35]. The assembly of "*Nostoc* sp. 1031Ymg" was obtained from metagenomic reads of co-culture RF3- t2. These reads were filtered by mapping them against the assembly of

Nostoc sp. 106C with BWA [36]. The resulting reads were assembled with Velvet using different *k*-mers: the assemblies with the largest length and the smaller number of contigs were selected and annotated using RAST [35]. JPPOOL and RFPOOL metagenomes from eutrophic conditions were obtained after pooling DNA samples from JP and RF, respectively, and treated as individual samples.

Taxonomic diversity. We first estimated taxonomic diversity using the 16S rRNA gene as a marker for our entire bacterial endophyte collection. PCR products of 1.4 Kbp in length, obtained using the F27 and R1492 primers [37], were obtained 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 [38]. We used the phylogenetic position of the top 10 hits from each search without duplicated matches, to determine both taxonomic diversity and phylogenetic relationships.

To measure the taxonomic composition of the sub-community co-cultures from metagenomes, we contrasted different methods of OTU identification and abundance that we presumed would be able to capture the breadth of taxa in our samples. We were particularly concerned with capturing cyanobacteria diversity. First, we used mOTUS, a method based on single-copy marker genes obtained from metagenomes and reference genomes [39]. We trimmed and filtered the Illumina reads and kept those with a minimum cutoff identity of 93%, and all other parameters as default. Taxa abundance from mOTUs, defined as the percentage of the genera present in each sample, was calculated with the Vegan v2.3-5 package in R [40]. We estimated the efficiency of our sequencing effort with respect to the total possible taxa per metagenome using the rarefaction method based on [41]. To do this we calculated the proportional number of sequences for each metagenome, in which the richness of mOTUs is sub-sampled randomly from the entire community.

Second, we used Kraken, a taxonomic analyzer to assign taxonomic labels to metagenomic DNA sequences based on exact alignment of *k*-mers [42]. Kraken is a taxonomic analyzer based on assigned taxonomy to short DNA reads, using a reference data base to identify alignments and the lowest common ancestor [42]. We implemented Kraken using the pipeline available at http://ccb.jhu.edu/software/kraken/ in our cluster Mazorka with five nodes each with 2 Intel Xeon E5-2650 @ 2.30GHz CPUs (“Haswell”, 10 cores/socket, 20 cores/node) and 768 GB of RAM memory. We used Kraken-build to make 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.

Our third method to estimate metagenomic taxonomic diversity was MG-RAST [43], which we used to annotate each metagenome at the level of genera using the default parameters, and selected only taxa that had at least 10,000 number of reliable hits. Each taxonomic annotation indicates the percentage of reads with predicted proteins and ribosomal RNA genes annotated to the indicated taxonomic level.

To visualize shared taxa among metagenomes, and their abundance, we used Cytoscape v3.4.0 [44], where each node and its size represent the abundance of an OTU, and lines represent shared taxa between metagenomes. The network was made by an interaction matrix, where each of the OTUs that had more than 14 readings assigned directly by Kraken identification, was linked to the metagenome from which it came. Identified nodes were manually ordered to prevent visual overlap. We also calculated the Shannon-Weaver *H'* and Simpson *L* indices for OTUs from all three methods using the Vegan v2.3-5 package in R [40].

Reconstruction of phylogenetic relationships. We aligned annotated 16S rRNA sequences trimmed to 1.1 Kbp, using MUSCLE v3.8.31 with default parameters [45]. This matrix was used for phylogenetic reconstruction using MrBayes v3.2 [46] with a gamma distribution type range with 1,000,000 generations. ModelTest [47] 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 in MUSCLE v3.8.31 with default parameters [45]. 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 [48] in Python v2.7.6.

To construct a complete phylogeny of cyanobacteria strains we used the amino acid sequences of GyrB and RpoB as markers [49]. However, their corresponding phylogenies lacked support and resolution even after concatenation, thus we included into the matrix orthologs of the Carbamoyl-phosphate synthase large subunit (CPS), Phenylalanine-tRNA ligase beta subunit (PheT) and the Trigger factor (Tig). Sequences of RpoB, GyrB, CPS, PheT and Tig were extracted from an in-house database of cyanobacterial genomes obtained from GenBank, and annotated using RAST [35]. The sequences were obtained using Blast with a cut-off e-value of 1e-50 and a bitscore of 200. Each set of sequences were aligned using MUSCLE v3.8.31 with default parameters [45], and trimmed manually. Independent phylogenies were performed for each marker to filter out redundant and divergent sequences. The sequences that passed this filter were included in the final array, which included the organisms for which all five markers could be retrieved. The final matrix included 289 taxa, with 3617 aminoacids, and it was used to reconstruct a tree with MrBayes, using a mixed substitution model based on posterior probabilities (aamodel[Wag]1.000) for proteins for a 10 million generations. Convergence of runs was reached after 1 million generations.

Finally, a high resolution cyanobacteria phylogenetic tree was constructed using a set of 198 conserved proteins (Additional file 1: **Table S1**), which represent the core of a set of 77 cyanobacterial genomes (Additional file 2: **Table S2**) including our two isolates (T09 and 106C) and the RF31YmG assembly; and *Fischerella* sp. NIES 3754 and *Hassallia byssoidea* VB512170 as outgroups. We extracted and assembled the cyanobacterial genomes from the metagenome RF3-T2. To obtain the RF31YmG genome, contigs from the 106C assembly were used as reference to match and extract reads from the RF3-t2 metagenome using BWA [36]. The obtained reads were assembled using Velvet with the extension columbus with different *k*-mers. The best assembly, considered as the largest assembly with the lower number of contigs, was selected and annotated with RAST as previously. The core genome was obtained using an in-house script available at https://github.com/nselem/EvoDivMet/wiki, which will be reported elsewhere in due course. Then, a set of 198 core proteins was selected from only 33 Nostocales genomes in our database to construct the final concatenated matrix, which included 45477 amino acids. We used this matrix to reconstruct a phylogeny using MrBayes v3.2 with a mixed model (not partitioned), for a million generations.

Genome mining for BGCs. To identify BGCs potentially directing the synthesis of specialized metabolites among selected cyanobacteria, we annotated the genome of the isolate 106C with antiSMASH [50]. The predicted BGCs were used as a reference for further searches among the selected genomes. For this purpose we used our in-house pipeline, called CORASON (available at https://github.com/nselem/EvoDivMet/wiki), which will be reported elsewhere in due course. CORASON allows for the identification of conserved and unique BGCs among the selected genomes. Prediction of the chemical structures of the putative specialized metabolites associated with these BGCs was done after domain identification and specificity prediction, mainly of adenylation and acyl transfer domains, with NRPS-PKS server [51], PRISM [52] and antiSMASH [50].

Results

Our experimental strategy (**Fig. 1**) to characterize the taxonomic diversity of the coralloid root endophytic microbiome led to hundreds of bacterial isolates obtained directly from the original sample (*t0)*; and 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 root, including bacteria present in low titers in the original sample (*t2*). It also provided a mean to obtain insights into the biosynthetic potential specific to the cyanobacteria inhabiting the coralloid root, which could be driving community interactions. In the following sections we describe the results obtained from this effort in three sub-sections, overall taxonomic diversity, cyanobacteria phylogenetic relationships and specificity of BGCs present in the *Dioon* coralloid roots.

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

Taxa assessment based in 16S rRNA. Cultivable bacteria constitute only a biased subset of the total endophyte biodiversity, yet from our 16S rRNA sequences alone we found 470 isolates grouped into 242 OTUs, distributed in 17 families and 11 bacterial orders, with 27 genera in total, representing most of the known bacterial groups (**Table 1**. See also Additional file 3: **Table S3**). As seen in our 16S rRNA phylogenetic reconstruction (**Fig. 2**), all of our sequences grouped within monophyletic clades, and most trees within each clade show that there are new species that remain to be described, in almost all of the genera found within the cycad coralloid root (see also Additional file 4: **Fig. S1**). An 87% of the taxa identified can be taxonomically classified as diazotrophic plant endophytes, validating our endophyte isolation procedures (see Materials & Methods). Indeed, most OTUs grouped within the genera *Streptomyces, Bacillus, Rhizobium, Stenotrophomonas, Pseudomonas, Mitsuaria, Achromobacter* and *Burkholderia*, which are known for their extraordinary taxonomic diversity, their ability to establish symbiont relationships across the tree of life, or are commonly found in the soil or the plant rhizosphere.

We confirmed previous reports of other bacteria associated to the cycad coralloid root, namely, *Bacillus*, which was previously reported as associated to the outside of the coralloid root; *Streptomyces*, previously isolated as an epiphyte [23], which grew on our selective media (ISP4); and *Pseudomonas* [19] growing indistinctly in our four non-selective media. As expected, we confirmed endophytes that belong to *Nostoc* [5], but also found *Tolypothrix*, a previously unreported genus of Nostocales living in the coralloid root. We isolated six strains belonging to this genus according to 16S rRNA characterization.

Our results also show that OTUs are shared among samples and species, with no specific distribution among the various isolation culture media (**Fig. 2**). There are environment-specific trends such as higher diversity in the dry environment. We observed a tendency in the 16S rRNA data showing that 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*). In terms of species diversity and abundance, the Shannon-Weaver and Simpson biodiversity indices based on genera abundance from 16S rRNA sequences have higher diversity in the dry environment than in the humid environment (Additional file 5: **Table S4**). We consider these results preliminary and limited by the use of cultivable approaches, but valid as they compare samples treated under the same conditions and thus informative to define further ecological studies.

Taxa assessment based in co-cultures metagenomics. We extracted and sequenced whole-community metagenomic DNA from *t1* and *t2* subcommunity co-cultures with the aim of enriching for specific interactions in response to growth conditions. We were able to sequence metagenomes from six different individuals grown on eutrophic conditions after 72 hours, whose DNAs were pooled as limited diversity was expected (JPPOOL and RFPOOL); from four different individuals after 30 days of culture in oligotrophic conditions, two from each of the two environments (JP2, JP6 and RF1, RF3); and after 365 days, same conditions, one from each environment (JP6 and RF3) (**Table 2**. see also Additional file 6: **Table S5**).

In terms of taxonomic diversity, each OTU-assignment strategy recovered different taxa and in different proportion (Table 2). Notably, despite visual confirmation of the occurrence of heterocyst-forming cyanobacteria in green cultures (Additional file 7: **Fig. S2**), mOTUS revealed only a minor proportion of cyanobacteria, only 6%. In contrast, MG-RAST likely overestimated diversity at 39%. Kraken provided and intermediate result with 12%. Kraken is also a sequence classification technique that can exclude sequence contaminants from the draft assembly, allowing us to generate a symbiotic cyanobacteria marker database as reference for future classification. Thus, Kraken-identified OTUs were used for all subsequent analyses.

In Kraken-based OTUs, specifically associated to one of the metagenomes (JP), we also found *Calothrix*, previously reported in *Encephalartos* [16, 17] and in *Cycas revoluta* [18]; and *Caulobacter*, which can be found associated to cyanobacteria [19]. Of the Nostocales we were unable to recover *Tolypothrix* in the metagenomes. Notably, taxa identified in the four metagenomes mostly overlap (**Fig. 3**. See also Additional file 8: **Figure S3**). The few exceptions that were unique to a sample include species such as *Shewanella* specific to JP2 from the dry environment, and *Cronobacter* specific to RF3 in the humid environment. Likewise, the original taxonomic diversity from the environmental isolates (*t0*), as revealed by their 16S rRNAs sequences, and that found in the co-culture sub-communities (*t1*), measured as OTUs by Kraken, overlap only partially. Specifically, we recovered 12 OTUs with 16S rRNAs that were not recovered with Kraken, and 79 OTUs discovered only with Kraken, showing the complementarity of our approaches.

Biodiversity indices showed the same tendency as in the 16S rRNA results, in which the dry environment is more diverse than the humid (Additional file 5: **Table S4**). In all cases results from BG11 co-cultures show higher diversity than those obtained from the *Caulobacter* + mannitol medium. Similar to the process of eutrophication in biofilms, in which nutrient availability affects biofilm diversity and composition [53], rapid growers and presumably primary producers colonized and took over in the eutrophic medium, resulting in overall low diversity. In contrast, the results of the oligotrophic conditions suggest a cyanobacteria-centric community enables diversity. Indeed, rarefaction curves based on Kraken estimates suggest we captured 40-60% of the microbial community in the BG11 media (15 genera in JP6), with the least being the results obtained from the co-cultures grown on the *Caulobacter* + mannitol medium (Additional file 9: **Figure S4**).

Differences in genera identified with 16S rRNA and metagenomes could be explained because our metagenomes may not be 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. It is likely that all three factors influenced our results. Despite these issues and differences in the media, we confirmed the occurrence of many of the bacterial endophyte taxonomic groups in the metagenomes, which were previously isolated and characterized with 16S rRNA. In sum, it is clear from these results that we have captured a significant fraction of the taxonomic diversity of the endophytes in the cycad coralloid root, and that the combination of isolation and shotgun metagenomics results in a realistic representation of the cycad coralloid bacterial community.

Dioon cyanobacteria belong to the family *Nostocaceae* and are a monophyletic group

In order to explore the specificity of our cyanobacterial isolates, we reconstructed a phylogeny from five markers (**Fig. 4a**. See also Additional file 10: **Figure S5**). Although cyanobacteria phylogenetic history is likely reticulated [54], our tree is congruent with previous phylogenies that grouped cyanobacteria into mostly monophyletic clades, and we recover and support various known taxa relationships. For instance, we support the lack of monophyly of *Chlorogloeopsis* and *Fischerella* with *Chlorogloeopsis* strains grouped with the nostocalean *Scytonema* [55]. We also support the monophyly of heterocyst and akinete-bearing cyanobacteria of the sections IV and V [56, 57]. A deeper discussion of the phylogeny is out of the scope of this article, but it will serve as additional evidence in the complex relationships of the cyanobacteria. Hereafter we focus on the Nostocaceae as they are the closest to our samples, and species from the IV and V group are able to establish various types of symbiotic associations [58].

Previous molecular studies and our own data show that choice of genome-wide markers, and the type of OTU assignment methods, significantly affect the ability to recover Nostocaceae phylogenetic history. Our results were contingent on using 198 genome-wide orthologs from a broad and curated database (Additional file 1: **Table S1**; Additional file 2: **Table S2**), combined with Kraken to assign OTUs, which was best at detecting cyanobacteria. Overall, our phylogeny (**Fig. 4b**) shows that *Calothrix* PCC 7507 fails to group within the *Rivulariaceae* and is instead nested within the Nostocaceae. We confirmed the presence of *Anabaena* (metagenomes) first mentioned as algae in the cycad literature [13]; and of *Nostoc* (isolates) [18], and show that they each separate clearly in our phylogeny. Also, *Nostoc* is sister to *Anabaena, Aphanizomenon* and *Trichormus* [59, and references therein]. A previously recognized clade using 16S rRNA, constituted by *Anabaena* species associated to *Aphanizomenon* species, with *A. cylindrica* as sister to the rest [60], is also distinct in our phylogeny (Clade I). This group includes the fern endophyte *Nostoc azollae* 0708, supporting original descriptions of *Anabaena* fern symbionts [61] and similar findings with 16S rRNA [59]. The *Nostoc* free-living PCC 7120 grouped distantly to strains of symbiotic origin.

Importantly, our *Dioon* isolates from T09, 106C and RF31YmG form a monophyletic clade. This contradicts previous studies in which different species of cycads host multiple cyanobacteria and do not form cycad or host-specific clades [6, 62, 63]. The isolate T09 was obtained from coralloid roots of *Dioon caputoi,* collected previously by our group in dry shrubland from the Tehuacan Valley in Puebla, and added as a control. This result suggests specificity of Nostocaeae symbionts within *Dioon* species. It also shows diverging evolutionary trajectories of *Nostoc* species associated with cycads, from those of the free-living Nostocaceae (**Fig. 4b**). Congruent with these findings, a 16S rRNA phylogeny of Nostocacean cyanobacteria shows that hormogonia-producing species symbiotic to *Gunnera* ferns, *Anthoceros*, and cycads, tend to cluster together [59].

The name of the new *Dioon* cyanobacteria symbionts remains to be determined. *Tolypothrix* sp PCC 7601 is sister taxon to our *Dioon* isolates, and they are sister to two other plant symbionts: *Nostoc* sp KVJ20 (PRJNA310825), which lives in special cavities located on the ventral surface of the gametophyte of the Norway liverwort *Blasia pusilla* [64]; and *Nostoc punctiforme* PCC73102 (ATCC 29133), associated with the Australian cycad *Macrozamia* [65]. *Calothrix* sp. PCC 7507 and *Fortiea contorta* PCC7126 are sister taxa to our isolates clade (Clade II). Thus, it is concluded that *Dioon* cyanobacteria endophytes belong to the family Nostocaceae, and that they show a monophyletic origin. This suggests that our isolates may be specialized bacteria, with unique metabolic and other phenotypic features that warrant further characterization and polyphasic taxonomic determination.

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

Mapping the size of each bacterial genome onto the phylogeny showed that our *Dioon* coralloid endophytes have larger genomes sizes than all other close relatives, while maintaining their (G+C)-content (**Fig. 4b**). Large genomes correlate with the ability of bacteria to produce specialized metabolites. Thus, we aimed at exploring the coralloid root microbiome functions in detail by identifying examples of BGCs putatively directing the synthesis of specialized metabolites (**Fig. 5**). Genome mining of isolate 106C revealed 18 BGCs (Additional file 11: **Table S6**). The analysis of the distribution of these BGCs among the selected Nostocaceae genomes (Additional file 12: **Table S7**) revealed that the heterocyst glycolipid (BGC 16), the only BGC with a defined product [66], and BGC 2, a terpene of unknown structure, were present in all analyzed genomes.

Mining of other known molecules associated with cycad cyanobionts, such as nodularin [67], or other known BGCs found in members of the genus *Nostoc*, yielded negative results.

In contrast, half of the BGCs were uniquely found within *Dioon* symbionts including isolate 106C. Remarkably, these nine BGCs are absent in the well-annotated genome of *Nostoc punctiforme* PCC73102, a strain isolated from an Australian *Zamia*. These observations support the metabolic specialization of *Dioon* cyanobionts. Among the *Dioon*-specific cyanobacterial BGCs we found four coding for lantipeptides, namely, BGC 1, 9, 10, 17 (**Fig. 5**, see also Additional file 13: **Text S1**). BGC 20 includes genes coding for one adenylation domain, one thiolation domain and one thioesterase domain, which may be involved in the synthesis of modified amino acids, or in the formation of a yet-to-be discovered metabolite. The remaining four BGCs code for NRPSs, including one NRPS-PKS hybrid, BGC 21, which codes for a PKS-NRPS hybrid system potentially directing the synthesis of a hybrid peptide with three residues (Phe-Thr-Phe) and a hydroxyl-iso-butyrate group as the C-terminal substituent.

BGC12, which caught our attention, codes for an assembly line predicted to direct the synthesis of 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. The N and C terminal modifications on this peptide are typical of small peptide aldehyde protease inhibitors, which have been previously reported on cyanobacteria [68]. Alternatively, the product of this biosynthetic system may be a siderophore, as iron-related genes were found next to the NRPS coding-genes and previous reports have shown that reductase domain-containing NRPS systems such as in myxochelin [69], are linked to iron chelators. The BGC 22 encodes a small NRPS system for a dipeptide (Gly-Val), which in 106C and RF3Mg seems to be associated to genes coding for chemotaxis proteins, also present in the corresponding region in T09.

BGC 23, the most interesting of all, codes for a NRPS system putatively directing the synthesis of a tripeptide consisting of leucine, valine and tyrosine residues, as well as an N-terminal acylation, an N-methylation at an amide bond of the isoleucine residue, plus a domain of unknown function likely modifying the tyrosine residue. Remarkably, the order of the domains in the BGC suggests lack of co-linearity, which may imply domain skipping or recycling. A search for peptides containing such modifications, performed with the server PRISM that includes a feature for de-replication of known chemical structures [52], directed our attention to nostoginins, a specialized metabolite whose biosynthetic pathway remains unknown. Nostoginin A is an acylated tripeptide (Leucine-Valine-Tyrosine) with N-acylations at the isoleucine and tyrosine residues, originally isolated from a member of the genus *Nostoc* [70], and shown to be a protease inhibitor with specificity towards aminopeptidases. Similar bioactivity has been found for its congeners nostiginin B, microginins FR1 and SD755, and oscillaginins A and B [71]. Interestingly, a nostoginin congener (Nostoginin B), which includes an extra tyrosine group at the C- terminal end, was also isolated from the same *Nostoc* strain as nostoginin A. The amino acid specificity of BGC 23 adenylation domains, the location of the modification on the leucine and tyrosine residues, the lack of collinearity, the presence of N-terminal acylation domains, the occurence of peptidase coding genes in the BGC, and the taxonomic origin of nostoginins, strongly suggest that BGC 23 is linked to these metabolites (**Fig. 5**).

In addition to our genome-driven analysis, we also assembled, annotated and mined, *de novo*, the metagenomes of *t1* and *t2* oligotrophic 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. This approach only revealed in all metagenomes together of *t1* five short signal sequences (less than 3.5 Kbp) that are suggestive of enzyme genes that could be part of BGCs. It seems that although representative of the rich biological diversity of the root, the lower coverage of these metagenomes hampered our ability to obtain loci long enough to allow proper annotation of presumed BGCs. In contrast, for *t2*, where bacterial diversity has been enriched we found two complete BGCs in the RF3 sub-community metagenome, both clearly coming from cyanobacteria, the most abundant taxa in the co-culture (**Table 2**). Indeed, these BGCs coincided with those found in the RF31YmG genome extracted from RF3 metagenome, showing that a computational pangenomic analysis of metagenomes is a promising approach to capture the biosynthetic potential of co-cultures.

Discussion

Our combined strategy of co-cultures at different timescales and genomic and metagenomic sequencing analyzed with a phylogenomic framework enabled us to study bacterial endosymbionts that coexist in the same cycad host, and identify the BGCs associated to their coralloid root-specific niche. We focus our discussion on the taxa found in the bacterial isolates, and OTUs present in the metagenomes, and we refer to species and OTUs interchangeably.

The microbiome of the cycad coralloid root reveals a biodiverse community, with monophyletic grouping of cyanobacteria

Our evidence undoubtedly shows that within the cycad coralloid root there is a highly diverse bacterial community within the cycad coralloid root of at least 27 genera identified with 16S rRNA of which 12 were not recovered with Kraken, and 79 additional genera identified in the metagenomes, which includes all of the previously reported Nostocales and newly reported genera. We validated previous reports of taxa for which their endophytic origin and presence was unclear or doubtful. Cyanobacteria are present, but also many other taxa that interact in a community.

We also support previous morphological observations that showed that an individual cycad plant could harbor diverse communities that differ in their taxonomic composition and life-strategy [23], from soil dwellers to well-known plant symbionts. Morphological studies observing mucilaginous material inside the coralloid root [14, 20] are also congruent with the microbiome consortium we describe. However, most of the abundant genera were shared among samples, which suggests weak taxonomic specificity in different environments. Similarly, the majority of the taxa identified in the phylogeny can be taxonomically classified as diazotrophic plant endophytes, which points toward functional congruence associated with nitrogen fixation, rather than phylogenetic filtering, and suggests a taxonomic and functional core.

Although many other groups are worth exploring, we focused on cyanobacteria as the main group of interest given previous records of this group in cycads, their ability to establish symbiosis with most lineages of eukaryotes in many different types of tissues, and in plants with known co-evolutionary histories [72]. This bacterial group is also renowned for its potential to synthesize specialized metabolites of applied and evolutionary interest.

Among our most interesting findings is the monophyletic placement of our cyanobacterial samples, which confirm a single morphological observation of possible specificity among cyanobacteria coralloid root endophytes (then termed phycobionts), and their hosts, including *Dioon* [5], and contrasts with several previous notions regarding relationships between *Nostocaceae* and their hosts. Cyanobionts in other systems, such as cyanobacteria from a single lichen species, are often more closely related to free-living microorganisms, strains belonging to other species, or to plant symbionts, than to each other. Likewise, other studies of symbiotically competent *Nostoc* isolates suggest that they are not specialized and strains isolated from one plant species are capable of infecting phylogenetically distant hosts [59, 73, 74]. These contrasting previous observations could be biased by partial taxon identification in what we know now is a diverse cycad coralloid root microbiome, including several different cyanobacteria genera. Additionally, those phylogenies were based on samples collected growing outside of their place of the cycad's native distribution [75]. As data is gathered from more genomes of bacterial cycad symbionts, it will be possible to test for other co-evolutionary relationships, including horizontal gene transfer between bacteria and the eukaryote host, and other patterns that suggest close evolutionary histories.

Cultivated bacterial sub-communities are useful to assess functional interactions of the root microbiome

We found congruent results in diversity patterns among 16S rRNA and metagenomes, yet there are clear limitations of 16S rRNA and even genome-wide markers to carry out in-depth microbiome analyses, depending on how OTUs are assigned. There are even more limitations to understanding their functional interactions. We increased our ability to identify a diverse array of organisms using cultivated bacterial sub-communities (*t1, t2*) and exploring their metagenomes with phylogenomic tools. Most of the genera with only a few species were recovered in *t1*, and genera with many species were recovered in both *t0* and *t1*. The differences in composition with genera identified without enrichment (*t0*) was expected, because environmental sampling and enriched inoculant complement each other, and aim to recover distinct aspects of the microbiome's composition [27]. 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 to pre-adapted cyanobacteria-associated groups.

The long-term one-year co-culture (*t2*) allowed us to explore at least some of the aforementioned possibilities. Although dynamic, the initial amount of inorganic nitrogen available in these co-cultures became a limiting factor over time. Hence, the establishment of stable communities after a year with emerging and surviving taxa suggests that Nitrogen fixation is at least one of the main driving forces in the assembly of the coralloid root community. Plant-associated and slow-growing actinobacterial taxa, renowned for being prolific producers of specialized metabolites, are abundant in these communities. Further exploration of the metabolic-driven hypotheses emerging from these observations in different conditions, with an emphasis on Nitrogen fixation and physiological studies of the community, is required to understand the complexity of such community. For now, we can conclude that co-cultures are a strategy that allows assessing deeper sub-community functional interactions within the microbiome of a specialized organ, as it is the cycad coralloid root.

Large genome size as a signature of facultative lifestyles in cycad cyanobacteria symbionts

Most bacterial endosymbionts of plants or animals show a reduction in genome size compared to free-living relatives [76], yet our endosymbiont samples have larger genome sizes than all other closely related taxa in their phylogeny. Large genome sizes in endosymbionts are usually attributed to a facultative relationship that requires retaining free-living stages. For instance, rhizobial nitrogen-fixing bacteria in root-nodules of legumes that exhibit multiple lineages with genome expansions compared to closely related taxa ([77] and references therein), are also more similar in genome content and size to other plant symbionts than to closely related species [78]. Other facultative symbionts which form Nitrogen-fixing root nodules in angiosperms have large genome sizes adapted to shifting from the soil to the plant environment [79], while others such as *Brucella, Wolbachia* or *Agrobacterium* have favored expansions of genome size to cope with complex and varying life-styles [80]. Thus, a feasible hypothesis is that the *Nostocaceae* taxa we found associated to the cycad coralloid root, have experienced a large genome expansion driven by selection to initially survive the structural, ecological and biological complexity of the soil from which they are recruited.

Additionally, a large repertoire of genes would be required to maintain the developmental phenotypic plasticity of the cyanobiont cells to adapt to the inside of the cycad host. Extremely plastic symbionts, such as *Nostoc* species, have notorious complex life cycles that require cell differentiation of the organism to be able to enter the host plant and disperse [81]. The only other cyanobacteria cycad symbiont sequenced, *Nostoc punctiforme* from an African cycad *Macrozamia* [65], is phenotypically plastic and ranges from photoautotrophic to diazotrophic, to facultatively heterotrophic. Its vegetative cells can develop into nitrogen-fixing heterocysts and have transient differentiation into hormogonia. Its genome shows 29% unique protein-encoding sequences of known function, with roles in its cell differentiation and symbiotic interaction properties [65]. It also has numerous insertion sequences and multilocus repeats, as well as genes encoding transposases and DNA modification enzymes, which would be congruent with genomic plasticity required to sense and respond to the environment outside and inside the plant [65].

In sum, taxonomic diversity of the coralloid root, combined with monophyly of the large *Nostocaceae* genomes found in the cycad coralloid root, could be a result of imposed constrains of the facultative symbiotic lifestyle, and the broad symbiotic competence with the plant host. The facultative nature of cyanobionts of *Dioon* would suggest they are secondary endophytes acquired from environmental sampling with host-specificity to *Dioon*.

It remains to be examined how the genomes of our *Dioon* cyanobionts expanded. Upcoming work on the comparative genomics of the cycad coralloid root microbiome should test for trends in genome size, AT content, changes in the content and distribution of repeats and mobile elements, distribution of accumulated mutations and type of genes gained or lost and pseudogenization. All these factors could inform the nature of the cycad-bacterial interactions in ecological and evolutionary time. Of particular interest to us, is how metabolic functions are retained or acquired in relation to loci present within the root microbiome. We begin exploring this by identifying and analyzing the distribution of BGCs in our bacterial genomes, which we discuss in the final section below.

BGCs are conserved and unique to the cycad cyanobionts

The bacterial repertoire of specialized metabolites can correlate to environmental selective pressures [82] and result in conserved metabolic and genetic repertoires among species facing similar challenges, including those from plant symbiotic relationships. In Nostocales, although free-living strains are often competent and will form symbiotic interactions under laboratory conditions with many hosts [83], most recruited cyanobacteria are capable of producing specific compounds to survive within the plant. A remarkable example of a specialized metabolite involved in symbiosis is nosperin, a polyketide produced by a lichen-associated *Nostoc* cyanobacteria [84]. This molecule belongs to the pederin family, which includes molecules produced by non-photosynthetic bacterial symbionts from beetles and sponges [84], suggesting a role on eukaryote-prokaryote interaction. Nosperin has also been found in the liverwort *Blasia*-

associated and in free-living *Nostoc* cyanobacteria [64] suggesting that in cycads, nosperin producers are selected for symbiosis, although production is not necessarily induced while inside the coralloid roots.

None of the BGCs for specialized metabolites previously reported for *Nostoc* cyanobionts of lichens, bryophytes or other cycads, namely, nosperin, mycocystin or nodularin, could be found in the *Dioon* cyanobionts. Our unique biosynthetic repertoire of several BGCs provides an example of metabolic specialization that correlates more with the plant host biology than with the environmental conditions or geography.

A chemical insight derived from our genome mining efforts, which may have a strong bearing on the evolution and biology of the *Dioon*-bacteria symbiosis, relates to the potential of *Dioon* cyanobionts to produce at least two small peptide protease inhibitors: the nostoginin-like peptides predicted to be produced by BGC 23; and the acylated penta-peptide aldehyde predicted to be produced by BGC 12. The specific presence of these metabolites in the cyanobionts may imply that proteolysis is involved in the cyanobacteria-cycad interaction. Protease activity in the coralloid roots may be linked to the reconfiguration of the root architecture or the filtering of the microbiome. This is an interesting possibility as the involvement of proteases in root nodule symbiosis has been observed previously between arbuscular mycorrhiza and legumes [85]. Within this context, our sub-community metagenomics approach provided a platform for BGC discovery that can be applied to other microbial-host interactions. Also, the BGC patterns found in the coralloid root add to the growing notion that symbiotic relations occur under heavy influence of chemical interactions, providing a rich source of novelty for drug discovery [84].

Conclusions

Our work shows that the coralloid root microbiome is a highly diverse community, with most genera shared within *Dioon* species regardless of their original environment or plant host. Our methods of enriched sub-community metagenomics and phylogenomics were able to recover a good portion of the taxonomic and phylogenetic diversity and reveal genes underlying the production of previously unreported specialized metabolites that result from bacterial functional interactions. We also provide emerging evidence of co-evolution between cyanobacteria and their plant hosts, suggested by monophyly of the samples and the presence of unique BGCs to their clade.

The coralloid root microbiome is likely established by dual forces of host-driven selection and environmental recruitment of cyanobacteria and possibly other taxa that are capable of transitioning from free-living to endosymbiotic lifestyles, and the functional capacities of the bacterial consortium itself. Future phylogenomic work on the cycad coralloid root microbiome via an integrated analysis of genome organization and expression of specialized metabolite production, as well as of their relationship to the fitness of the host, will further facilitate our understanding of the evolutionary history of the cycad microbiome.

References

1. Norstog KJ and Nicholls TJ, The Biology of the Cycads. Cornell University Press: New York. 1997. p. 504

2. Bergensen F, Lindblad P, and Rai A. Nitrogen fixation in coralloid roots of *Macrozamia communis*. L. Johnson. Aus J Bio Sc. 1986.18:1135-42.

3. Halliday J and Pate J. Symbiotic nitrogen fixation by blue algae in the cycad *Marozamia riedlei*: Physiological characteristics and ecological significance. Aus J Plant Phys. 1976.3:349-58.

4. Grove T, O'connell A, and Malajczuk N. Effects of fire on the growth, nutrient content and rate of nitrogen fixation of the cycad *Macrozamia riedlei*. Australian Journal of Botany. 1980.28:271-81.

5. Caiola M. On the phycobionts of the cycad coralloid roots. New Phytologist 1980.85:537-44

6. Zimmerman WJ and Rosen BH. Cyanobiont diversity within and among cycads of one field site. Canadian J Microbiol 1992.38:1324-8.

7. Costa JL and P L, Cyanobacteria in Symbiosis with Cycads, in Cyanobacteria in Symbiosis. Kluwer Academic Publishers: Dordrecht. 2002. p. 195–205.

8. Costa J, Romero E, and Lindblad P. Sequence based data supports a single *Nostoc* strain in individual coralloid roots of cycads. FEMS Microbiol Ecol. 2004.49:481-7.

9. Costa J, Paulsrud P, and Lindblad P. Cyanobiont diversity within coralloid roots of selected cycad species. FEMS Microbiol Ecol 1999.28:85-91.

10. Thajuddin N, Muralitharan G, Sundaramoorthy M, Ramamoorthy R, Ramachandran S, et al. Morphological and genetic diversity of symbiotic cyanobacteria from cycads. J Basic Microbiol. 2010.50:254-65.

11. Gehringer M, Pengelly J, Cuddy W, Fieker C, Forster P, et al. Host selection of symbiotic cyanobacteria in 31 species of the Australian cycad genus: *Macrozamia* (Zamiaceae). Molecular Plant-Microbe Interactions 2010.23:811-22.

12. Cuddy W, Neilan B, and Gehringer M. Comparative analysis of cyanobacteria in the rhizosphere and as endosymbionts of cycads in drought-affected soils. FEMS Microbiol Ecol. 2012.80:204-15.

13. Chaudhuri HaA, A.R. The coral-like roots of *Cycas revoluta, Cycas circinalis* and *Zamia floridana* and the alga inhabiting them. J Indian Bot Soc. 1931.10:43-59.

14. Baulina O and Lobakova E. Atypical cell forms overproducing extracellular substances in populations of cycad cyanobionts. Microbiology. 2003.72:701-12.

15. Zvyagintsev D, Zenova G, Lobakova E, and Savelyev I. Morphological and physiological modifications of cyanobacteria in experimental cyanobacterium-actinomycete associations. Microbiology. 2010.79:314-20.

16. Grobbelaar N, Scott WE, Hattingh W, and Marshall J. The identification of the coralloid root endophytes of the southern African cycads and the ability of the isolates to fix dinitrogen. South African J Bot. 1987.53:111-8.

17. Huan T and Grobbelaar N. Isolation and characterization of endosymbiotic *Calothrix* (Cyanophyceae) in *Encephalartos hildenbrandii* (Cycadales). Phyocologia. 1989 28:464-8.

18. Thajuddin N, Muralitharan G, Sundaramoorthy M, Ramamoorthy R, Ramachandran S, et al. Morphological and genetic diversity of symbiotic cyanobacteria from cycads. J Basic Microbiol. 2010.50:254-65.

19. Bershova O, Kopteva Z, and Tantsyyurenko E, The interrelations between the blue-green algae -the causative agents of the water 'bloom' - and bacteria. , in 'Tsvetenie' Vody, A. Topanchevsky, Editor. Naukova Dumka: Kiev, USSR.1968. p. 159-71.

20. Ow M, Gantar M, and Elhai J. Reconstitution of a cycad-cyanobacterial association. Symbiosis. 1999.27:125-34.

21. Yamada S, Ohkubo S, Miyashita H, and Setoguchi H. Genetic diversity of symbiotic cyanobacteria in *Cycas revoluta* (Cycadaceae). FEMS Microbiol Ecol 2012.81:696-706.

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

23. Lobakova ES, Orazova, MK and Dobrovol'skaya, TG. Microbial complexes occurring on the apogeotropic roots and in the rhizosphere of cycad plants. Microbiology. 2003.72:628.

24. De Luca P, Sabato S, and Vazquez-Torres M. *Dioon meroale* (Zamiaceae), a new species from Mexico. Brittonia. 1981.33:179-85.

25. Lázaro-Zermeño JM, González-Espinosa M, Mendoza A, and Martínez-Ramos M. Historia natural de *Dioon merolae* (Zamiaceae) en Chiapas, México. Botanical Sciences. 2012.90:73-87.

26. Traxler M and Kolter R. Natural products in soil microbe interactions and evolution. Nat Prod Rep. 2015.32:956-70.

27. Cibrián-Jaramillo A and Barona-Gómez F. Increasing metagenomic resolution of microbiome interactions through functional phylogenomics and bacterial sub-communities. Frontiers in Genetics. 2016.7:4.

28. Atlas, RM, Handbook of Microbiological Media, CRC press: Florida. 2004. ISBN 9781439804087.

29. Collection ATC, ATCC Catalogue of Bacteria and Bacteriophages. 1992: Rockville, MD.

30. Subba-Rao N, Soil Microorganisms and Plant Growth: Science Publishers, Inc. 1995 p. 350. ISBN 1886106185.

31. Shirling E and Gottlieb D. Methods for characterization of *Streptomyces* species. Int J Syst Evol Microbiol. 1966.16:313-40.

32. Rippka R, Stanier R, Deruelles J, Herdman M, and Waterbury J. Generic assignments, strain histories and properties of pure cultures of Cyanobacteria. Microbiology. 1979.111:1-61.

33. Bolger AM, Lohse M, and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014.30:2114-20.

34. Zerbino DR and Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Research. 2008.18:821-9.

35. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics. 2008.9:75.

36. Li H and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. 2009.25:1754-60.

37. Lane, D. J. 16S/23S rRNA sequencing. In Stackebrandt, E and Goodfellow, M, editors. Nucleic acid techniques in bacterial systematics. John Wiley & Sons, New York. 1991. p. 115-175

38. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucl Acids Res. 2013.41:D590--D6.

39. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Meth. 2013.10:1196-9.

40. Oksanen J. BFG, Kindt R., Legendre P., Minchin P. R., O'Hara R. B., et al. . Vegan: community ecology package. R Packag. version 2. 2015.

41. Hurlbert SH. The nonconcept of species diversity: a critique and alternative parameters. Ecology. 1971.52:577-86.

42. Wood D and Salzberg S. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology. 2014.15:R46.

43. Wilke A, Bischof J, Gerlach W, Glass E, Harrison T, et al. The MG-RAST metagenomics database and portal in 2015. Nucl Acids Res. 2016.44:D590-D4.

44. Shannon P, Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Admin, N., Schwikowski, B., Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research. 2003.13:2498–504.

45. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004.32:1792-7.

46. Ronquist F, Teslenko M, van der Mark P, Ayres D, Darling A, et al. MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology. 2012.61:539–42.

47. Posada D and KA C. ModelTest: testing the model of DNA substitution. Bioinformatics 1998. 1998.14:817-8.

48. Huerta-Cepas J, Dopazo J, and Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010.11:24.

49. Capella-Gutierrez S, Kauff F, and Gabaldón T. A phylogenomics approach for selecting robust sets of phylogenetic markers. Nucl Acids Res. 2014.42:e54-e.

50. Weber T, Blin K, Duddela S, Krug D, Kim H, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucl Acids Res 2015.43:W237-W43

51. Bachmann B and Ravel J. *In silico* prediction of microbial secondary metabolic pathways from DNA sequence data. Methods in Enzymology. 2009.458:181-217.

52. Skinnider M, Dejong C, Rees P, Johnston C, Li H, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res 2015.43:9645–62.

53. Costerton J, Lewandowski Z, Caldwell D, Korber D, and Lappin-Scott H. Microbial biofilms. Annu Rev Microbiol 1995.49:711–45.

54. Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, & Papke RT. Phylogenetic analyses of cyanobacterial genomes: Quantification of horizontal gene transfer events. Genome Research. 2006.16:1099–108.

55. Tomitani A, Knoll AH, Cavanaugh CM, and Ohno T. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. PNAS. 2006.103:5442-7.

56. Tomitani A, Knoll AH, Cavanaugh CM and Ohno T. The evolutionary diversification of cyanobacteria: Molecular–phylogenetic and paleontological perspectives. PNAS. 2006.103:5442-7.

57. Turner S, Pryer K, Miao V, and Palmer J. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. J Eukaryot Microbiol. 1999.46:327-38.

58. Rai AN, Bergman, B., Rasmussen, Ulla, editors. Cyanobacteria in Symbiosis. Springer: Netherlands. 2002 p. 355. ISBN 9780306-48005-8.

59. Papaefthimiou D, Van Hove C, Lejeune A, Rasmussen U, Wilmotte A. Diversity and host specificity of *Azolla* cyanobionts. J Phycol. 2008.44:60-70.

60. Lyra C, Suomalainen S, Gugger M, Vezie C, Sundman P, Paulin L and Sivonen K. Molecular characterization of planktic cyanobacteria of *Anabaena*, *Aphanizomenon, Microcystis* and *Planktothrix* genera. Int J Syst Evol Microbiol 2001.51:513-26.

61. Strasburger E. Die Controversen der indirecten Keimtheilung. Arch Mikrob Anat 1884.23:301.

62. Lindblad P, Haselkorn R, Bergman B, Nierzwicki-Bauer SA, and Rica C. Microbiology. Symbiosis. 1989:20- 4.

63. Zheng W ST, Bao X, Bergman B, Rasmussen U. High cyanobacterial diversity in coralloid roots of cycads revealed by PCR fingerprinting. FEMS Microbiol Ecol. 2002.40:215-22.

64. Liaimer A, Jensen JB and Dittmann E. A genetic and ghemical perspective on symbiotic recruitment of Cyanobacteria of the genus *Nostoc* into the Host Plant *Blasia pusilla* L. Frontiers in Microbiology. 2016.7.

65. Meeks JC, Elhai J, Thiel T, et al. An overview of the genome of *Nostoc punctiforme,* a multicellular, symbiotic cyanobacterium. Photosynthesis Research. 2001.70:85-106.

66. Soriente A, Sodano G, Cambacorta A, and Trincone A. Structure of the “heterocyst glycolipids” of the marine cyanobacterium *Nodularia harveyana*. Tetrahedron. 1992.48:5375–84.

67. Gehringer M, Adler L, Roberts A, et al. Nodularin, a cyanobacterial toxin, is synthesized in planta by symbiotic *Nostoc* sp. The ISME Journal. 2012.6:1834–47.

68. Fewer DP, Jokela J, Paukku E, et al. New Structural variants of aeruginosin produced by the toxic bloom forming cyanobacterium *Nodularia spumigena*. PLoS ONE. 2013.8:e73618.

69. Li Y, Weissman K, and Müller R. Myxochelin biosynthesis: direct evidence for two- and four-electron reduction of a carrier protein-bound thioester. J Am Chem Soc. 2008.130:7554–5.

70. Ploutno A and Carmeli S. Modified peptides from a water bloom of the cyanobacterium *Nostoc* sp. Tetrahedron. 2002.58:9949-57.

71. Sano T and Kaya K. A 3-amino-10-chloro-2-hydroxydecanoic acid-containing tetrapeptide from *Oscillatoria agardhii*. Phytochemistry. 1998.44:1503-5.

72. Rai AN, Soderback E, Bergman B. Cyanobacterium–plant symbioses. New Phytologist. 2000.147:449-81.

73. Johansson C and Birgitta B. Reconstitution of the symbiosis of *Gunnera manicata* Linden: cyanobacterial specificity. New Phytologist. 1994.126: 643-652.

74. Whitton BA, editor. Ecology of Cyanobacteria II: Their Diversity in Space and Time: Springer Science & Business Media: Netherlands. 2012. p.760. ISBN 97894007-3855-3.

75. Papaefthimiou D, Mugnai HPM, Lukesova A, et al. Differential patterns of evolution and distribution of the symbiotic behaviour in nostocacean cyanobacteria. Int J Syst Evol Microbiol. 2008.58 553–64.

76. McCutcheon J. The bacterial essence of tiny symbiont genomes. Curr Opin Microbiol. 2010.13:73-8.

77. MacLean A, Finan T, and Sadowsky M. Genomes of the Symbiotic Nitrogen-Fixing Bacteria of Legumes. Plant Physiol. 2007.144:615-22.

78. Bentley S and Parkhill J. Comparative genomic structure of prokaryotes. Annu Rev Genet 2004.38:771-92.

79. Normand P, Lapierre P, Tisa L, Gogarten J, Alloisio N, et al. Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. Genome Res. 2007.17:7-15.

80. Tsolis R. Comparative genome analysis of the alpha-proteobacteria: relationships between plant and animal pathogens and host specificity. PNAS. 2002.99:12503–5.

81. Meeks J, Campbell E, Summers M, and Wong F. Cellular differentiation in the cyanobacterium *Nostoc punctiforme*. Arch Microbiol 2002.178:395–403.

82. Ziemert N, Alanjaryab M, and Weber T. The evolution of genome mining in microbes – a review. Nat Prod Rep. 2016.33:988.

83. West NJ and Adams DG. Phenotypic and genotypic comparison of symbiotic and free-living cyanobacteria from a single field site. Appl Environ Microbiol 1997.63:4479-84.

84. Kampa A, Gagunashvili AN, Gulder TAM, et al. Metagenomic natural product discovery in lichen provides evidence for a family of biosynthetic pathways in diverse symbioses. PNAS. 2013.110:E3129-E37.

85. Takeda N, Kistner C, Kosuta S, et al. Proteases in plant root symbiosis. Phytochemistry. 2007.68:111-21.

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

Phylum	Class	Order	Family	Genus	OTUs [a]
Bacteroidetes	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae	*Mucilaginibacter*	3
				Sphingobium	1
				Sphingomonas	2
				Variovorax	**1**
	Cytophagales	Cytophagales	Cytophagaceae	*Dyadobacter*	1
Cyanobacteria	Cyanobacteria	Nostocales	Microchaetaceae	*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
			Brucellaceae	*Ochrobactrum*	1
	Betaproteobacteria	Burkholderiales	Alcaligenaceae	*Achromobacter*	***33***
			Burkholderiaceae	*Burkholderia*	***39***
				Ralstonia	***2***
				Mitsuaria	**8**
			Comamonadaceae	*Variovorax*	1
	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	*Enterobacter*	***3***
				Luteibacter	1
				Pantoea	***1***
		Pseudomonadales	Pseudomonadaceae	*Pseudomonas*	***21***
		Xanthomonadales	Xanthomonadaceae	*Luteibacter*	**2**
				Stenotrophomonas	***35***
				Xanthomonas	**2**
Actinobacteria	Actinobacteria	Micrococcales	Microbacteriaceae	*Microbacterium*	5
		Streptomycetales	Streptomycetaceae	*Streptomyces*	***19***

[a] Taxa identified in the literature as endophytes (italics) and/or diazotroph (bold) are shown.

Table 2. Taxonomic composition of sub-communities isolated from *Dioon* coralloid roots

Sample [a]	Growth conditions	Genera identified with different methods: total number (bold), most abundant (%)		
		mOTUs	**Kraken**	**MG-RAST**
JPPOOL	Eutrophic, 72 hours, *Caulobacter* medium + mannitol	**6**, *Bacillus* (87%)	**22**, *Bacillus* (84%)	**512**, *Bacillus* (86%)
RFPOOL		**8**, *Bacillus* (99%)	**25**, *Bacillus* (65%)	**524**, *Bacillus* (80%)
JP2	Oligotrophic, 30 days, BG-11	**42**, *Agrobacterium* (45%)	**57**, *Rhizobium* (7%)	**1273**, *Nostoc* (21%)
JP6		**38**, *Pseudoxanthomonas* (22%)	**69**, *Xanthomonas* (2%)	**1253**, *Xanthomonas* (8%)
RF1		**33**, *Stenotrophomonas* (83%)	**63**, *Nostoc* (3%)	**1157**, *Stenotrophomonas* (20%)
RF3		**25**, *Stenotrophomonas* (42%)	**61**, *Xanthomonas* (7%)	**1065**, *Xanthomonas* (22%)
JP6	Oligotrophic, 1 year, BG-11	**70**, *Deinococcus* (25%)	**69**, *Deinococcus* (4%)	**1957**, *Deinococcus* (26%)
RF3		**67**, *Stenotrophomonas* (33%)	**63**, *Nostoc* (3%)	**1592**, *Nostoc* (13%)

[a]JPPOOL= JP1, JP2, JP6; RFPOOL = RF1, RF3, RF9.

Legends to Main Figures

Figure 1. Pipeline to capture and characterize bacterial microbiome diversity. Coralloid roots from cycads growing naturally in dry and humid deciduous tropical forests were sampled (photo of coralloid root of approx. 9cm in length shown, not to scale). Endophytes from the macerated root were isolated, following two strategies: directly from the sample (*t0*) and after enrichment using co-cultures of sub-communities, and sampled after 30 days (*t1*), although sampling can be done anytime (*t1…tn*). Cultivable bacteria were obtained using an array of six different media. Co-cultures were characterized using shotgun metagenomics, and the resulting data was used to select representative genomes from the endophyte culture collection that we mined for functional information using a phylogenomic and comparative genomic approaches.

Figure 2. 16S rRNA Bayesian phylogeny of endophytes from coralloid roots of *Dioon merolae*. The external ring refers to the two environments sampled: dry or JP (D - orange) and humid of RF (H - blue) deciduous tropical forests. The inner ring refers isolation strategy: directly from the sample (*t0* - white) or after enrichment using co-cultures of sub-communities (*t1* - gray). Major bacterial groups are highlighted in different colors across the tree.

Figure 3. Network of taxa co-occurrence from different coralloid root samples. The lines connecting the circles represent shared taxa identified with Kraken from the metagenomes. Orange lines correspond to samples from the dry (JP) forest and blue to

samples from the humid (RF) forest. The most abundant genera in the four metagenomes are represented by circles. Circle diameters are scaled in accordance with the number of reads associated to each genus.

Figure 4. Phylogeny of *Cyanobacteria*. A. Multilocus phylogeny. The tree was constructed with five molecular markers and genomes obtained from GenBank, plus our genomes from T09, 106C and Rf31Y. Branches names have been colored according to the genera originally assigned in GenBank (a larger version of the tree is available as additional file 10: Figure S5); **B. Genome-wide phylogeny of the family *Nostocaceae*.** The tree was constructed with 45 conserved proteins, and includes *Dioon* cyanobionts 106C, T09 and Rf31Ymg. The habitat type of each taxa is indicated with colored bullets. The bars show a relatively homogeneous (G+C)-content among *Nostocaceae* cyanobacteria, and a trend for larger genomes in *Dioon*-associated cyanobacteria.

Figure 5. *Dioon*-specific cyanobiont biosynthetic gene clusters for specialized metabolites predicted from their genomes. Genes are shown as colored boxes, the tips of the boxes indicate the direction of their translation. Annotation color key is provided. Domain organization, biosynthetic logic and products are indicated below each BGC, except for lantipeptide encoded by BGCs 1, 9, 10 and 17, whose predicted products are shown as additional file 13: Text S1.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The 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. RF31Y

Metagenomes are available at sequence read archive (ID number pending), and directly from the corresponding author. Other data generated or analyzed during this study are included in this published article and its supplementary information or additional files, as enlisted:

Additional file 1: Table S1.docx/ Proteins in the cyanobacterial core genome. Annotated proteins used to reconstruct the cyanobacteria phylogenetic tree of 198 conserved proteins which represent the core of a set of 77 cyanobacterial genomes. We provide the name of the protein and the aminoacid sequence.

Additional file 2: Table S2.docs/Genomes used to obtain the core proteome. List of species and their larger classification used to obtain the core genome.

Additional file 3: Table S3.xlsx/List of 470 isolated bacteria with their 16S rRNA. We enlist all of the identified taxa isolated from the *t0* samples and identified with 16S rRNA Sanger-sequencing.

Additional file 4: Figure S1.pdf/Graphic representation of each group identified with 16S rRNA from isolates. A) We generated individual phylogenies for each of the genera in our main tree and represented them graphically as shown here. **B)** We also show individual trees with support values. A full resolution of both figures as individual files is available at: https://www.dropbox.com/sh/ss5mmwujnynyc7m/AABqABxc5wS_wjd8NzkarHTca?dl=0.

Additional file 5: Table S4.docx/ Biodiversity indices of 16S rRNA and OTUs. Diversity indices estimated for samples from 16S rRNA data, and from the four metagenomes (MET) we sequenced. We calculated Shannon-Weaver *H'* (1962) and Simpson *L* (1964).

Additional file 6: Table S5.docx/Statistics of metagenomes sequenced. We provide detail on the sequencing depth, contigs, quality of contigs and other basic statistics on sequenced metagenomes.

Additional file 7: Figure S2.jpg/ Pictures of cyanobacteria-centric co-cultures. Co-cultures in 1L flasks. In the insets is a close up of the culture, where a mucilaginous biofilm mass can be observed, presumably polysaccharides generated by the cyanobacteria.

Additional file 8: Figure S3. Kraken-based taxonomic diversity of metagenomes. Taxa abundance from the metagenome mOTUs defined as the percentage of the genera present in each sample. Jiquipilas (JP) is the dry environment, while Raymundo Flores (RF) individuals are found in the humid environment. JP or RFPOOL refers the samples sequenced in pools from media No. 6.

Additional file 9: Figure S4.pdf/ Rarefaction analysis of 16S rRNA and OTUs data. Shown is the proportion of OTUs represented by sample, by type of culture and by environment for each of the metagenomes sequenced, and a total of possible samples (All samples) according to a rarefaction estimate.

Additional file 10: Figure S5.pdf/Concatenated species-tree of cyanobacteria. Complete phylogeny of the Nostocales using five molecular markers, RPOB, GyrB, CPS, PheT and Tig. See text for technical details.

Additional file 11: Table S6.docx/Prediction of BGCs on the genome of isolate 106C. Biosynthetic Gene Clusters predicted by antiSMASH on the genome of isolate 106C are enlisted, with their corresponding length in Kp.

Additional file 12: Table S7.docx/106C-specific BGCs throughout Nostocales. We show the presence or absence of the 18 BGCs found throughout the Nostocales, to emphasize their presence of only some of them in our samples.

Additional file 13: Text S1.docx/ Predicted lantipeptide from *Dioon* cyanobionts. We show the sequence corresponding to the lantipeptides from the unique BGCs, whose prediction could not be fully shown in the main figures.

Any additional datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

Funding from this work is from CONACyT #169701 to ACJ, CONACyT #179290 and #177568 to FBG.

Authors' contributions

PC-M executed laboratory work, analyzed and interpreted data, and was a major contributor in writing the manuscript. AC-M executed laboratory work and analyzed data. NSM analyzed data. MAP-F identified and collected the plants. AC-J and FB-G equally co-designed and executed the study. AC-J was the main contributor in writing the manuscript. FB-G revised the manuscript critically for intellectual content. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge Rafael Rincón for his help with preparation of supplementary material. We also thank Juan Palacios, Pablo Suarez-Moo and Antonio Hernández for help during field collections, as well as Hilda E. Ramos-Abiotes and Flor Zamudio for technical support.

Authors' information

PC-M is a biochemist with a PhD in plant biotechnology. He is focused on the evolutionary mechanisms behind the chemical diversity of bacterial metabolism and the effect of natural selection upon chemical structures and biosynthetic pathways of NPs. He strongly believes that integrative biology approaches will have a direct impact on the discovery of novel molecules.

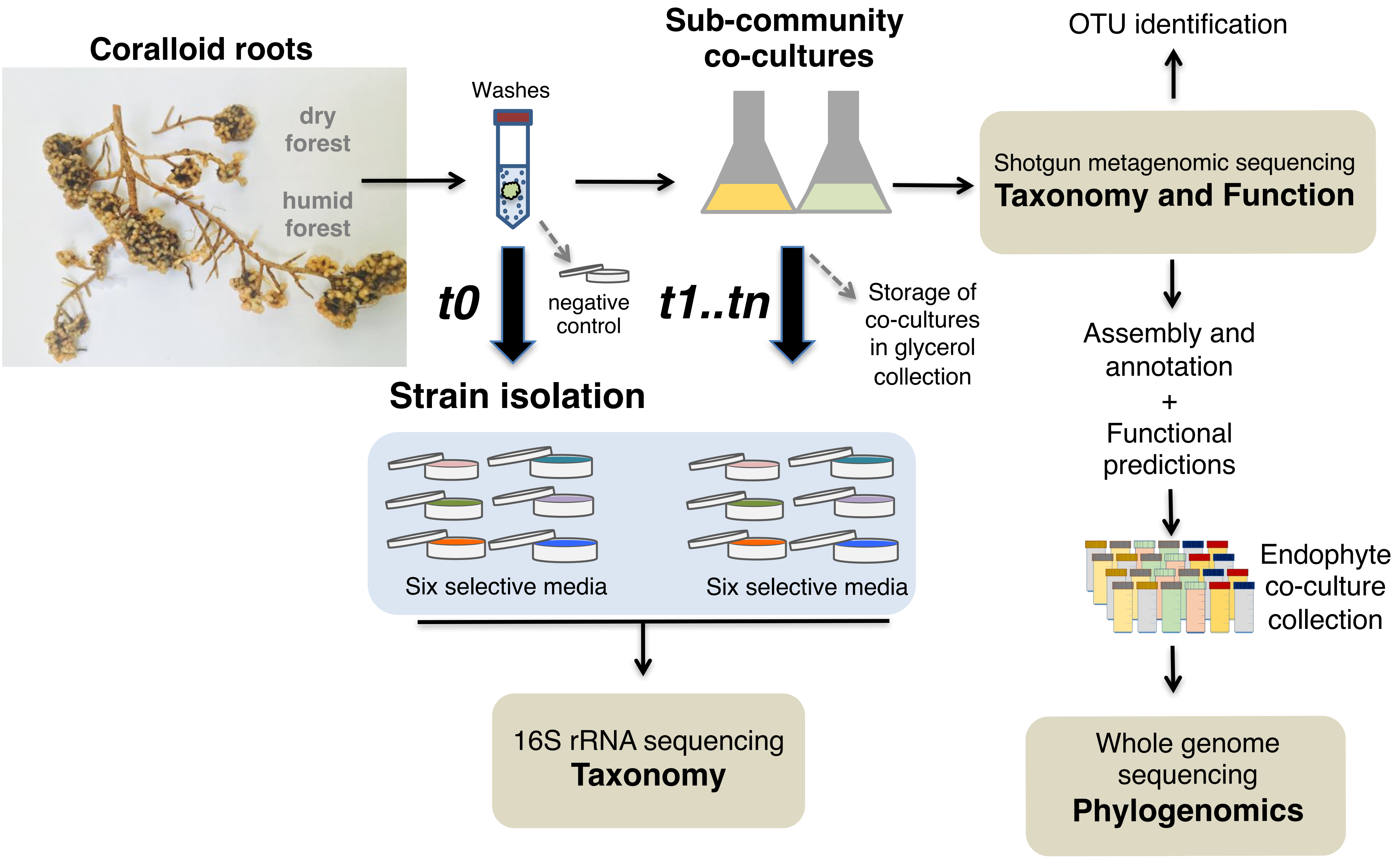
AC-G is a biologist specializing in the field of bioinformatics, with a master's degree in biotechnology in plants dedicated to the study of microbiota of plants.

NSM is a mathematician by training. She is currently in the last year of her PhD in Integrative Biology, studying the relationship between genome dynamics and enzyme promiscuity. Her aim is to develop new approaches with predictive power for functional annotation of enzymes and metabolic pathways.

MAPF is a biology, researcher and professor of Herbarium Eizi Matuda and Evolutionary ecology laboratory in the Universidad de Ciencias y Artes de Chiapas. He is studying the biology, systematics and ecology of Mexican cycads and palms, and the analysis of communities of plants in the tropical region of Mexico.

FB-G is a chemist with interest in the evolutionary and mechanistic aspects that allowed for the appearance of bacterial metabolism from a phylogenomics perspective. He runs a concept-driven multi- and inter-disciplinary research program that integrates different scales and types of data. http://www.langebio.cinvestav.mx/?pag=120

AC-J is an evolutionary biologist with specialization in plant population genetics and phylogenomics. She is interested in integrating multiple disciplines to understand the adaptive value of microbiomes in plant ecological and evolutionary history, in cycads in particular. http://www.langebio.cinvestav.mx/?pag=426.



Coralloid roots

dry forest

humid forest

Washes

negative control

t0

Strain isolation

Six selective media

Six selective media

16S rRNA sequencing
Taxonomy

Sub-community co-cultures

t1..tn

Storage of co-cultures in glycerol collection

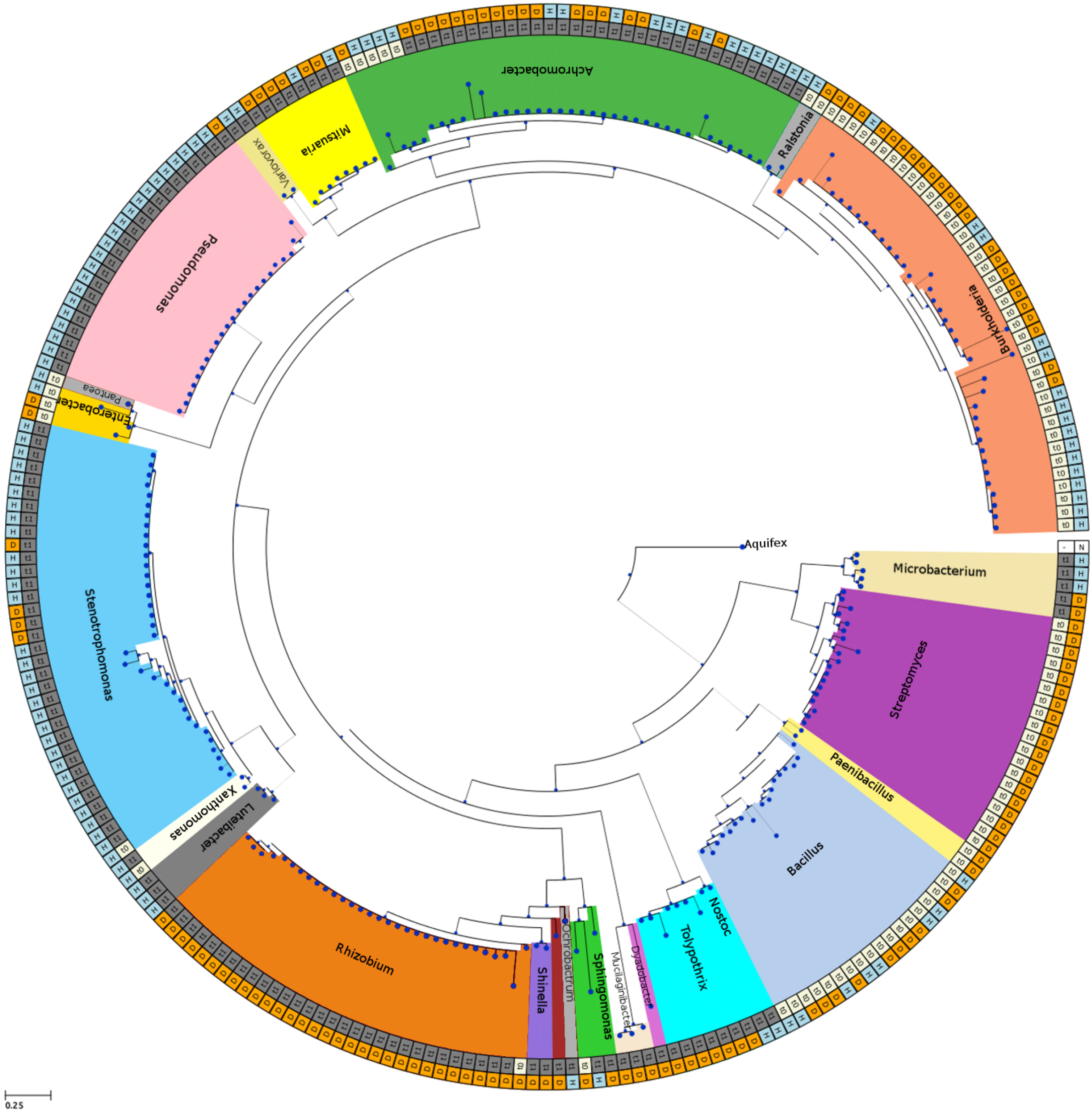
Shotgun metagenomic sequencing
Taxonomy and Function

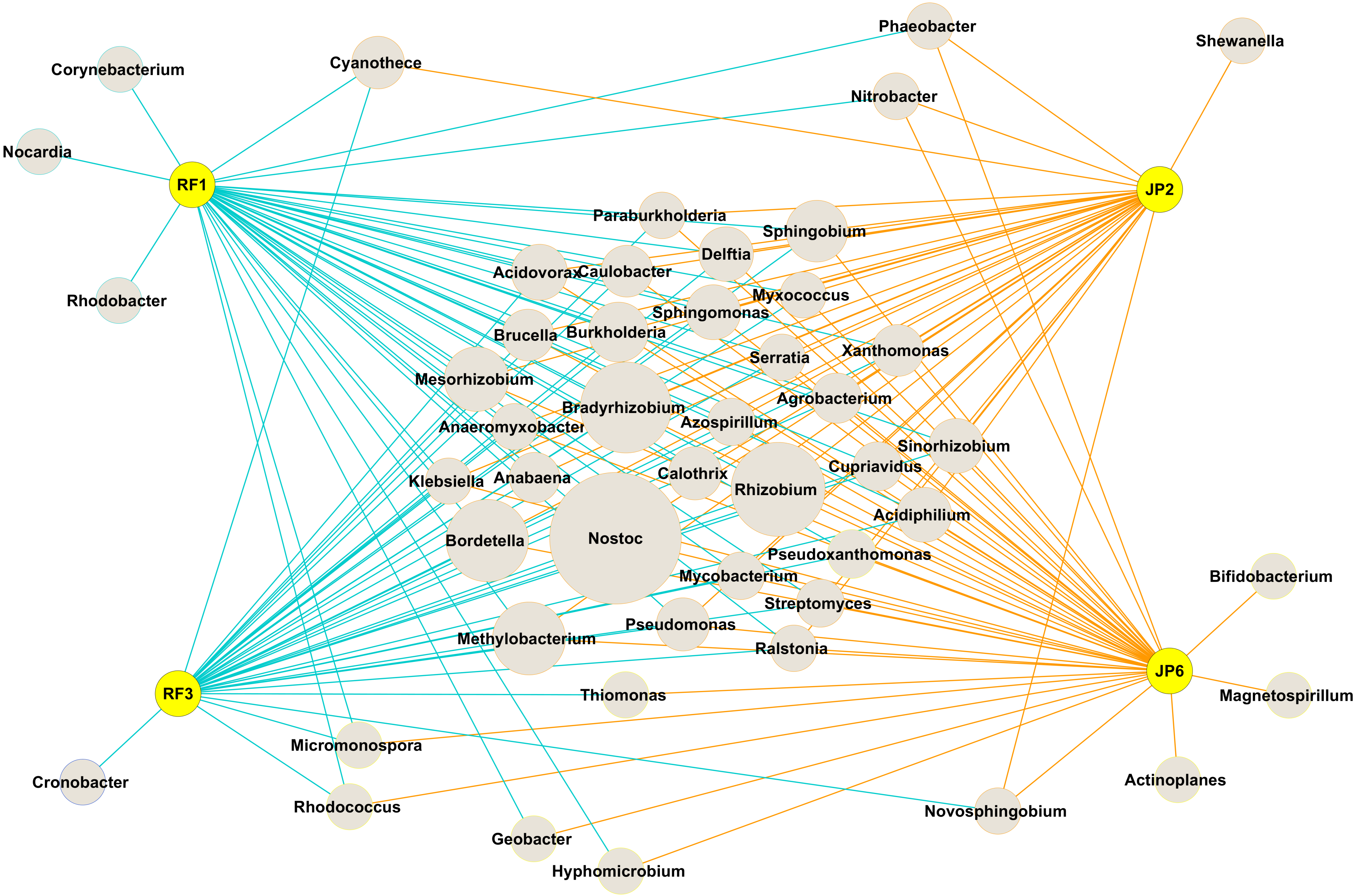
OTU identification

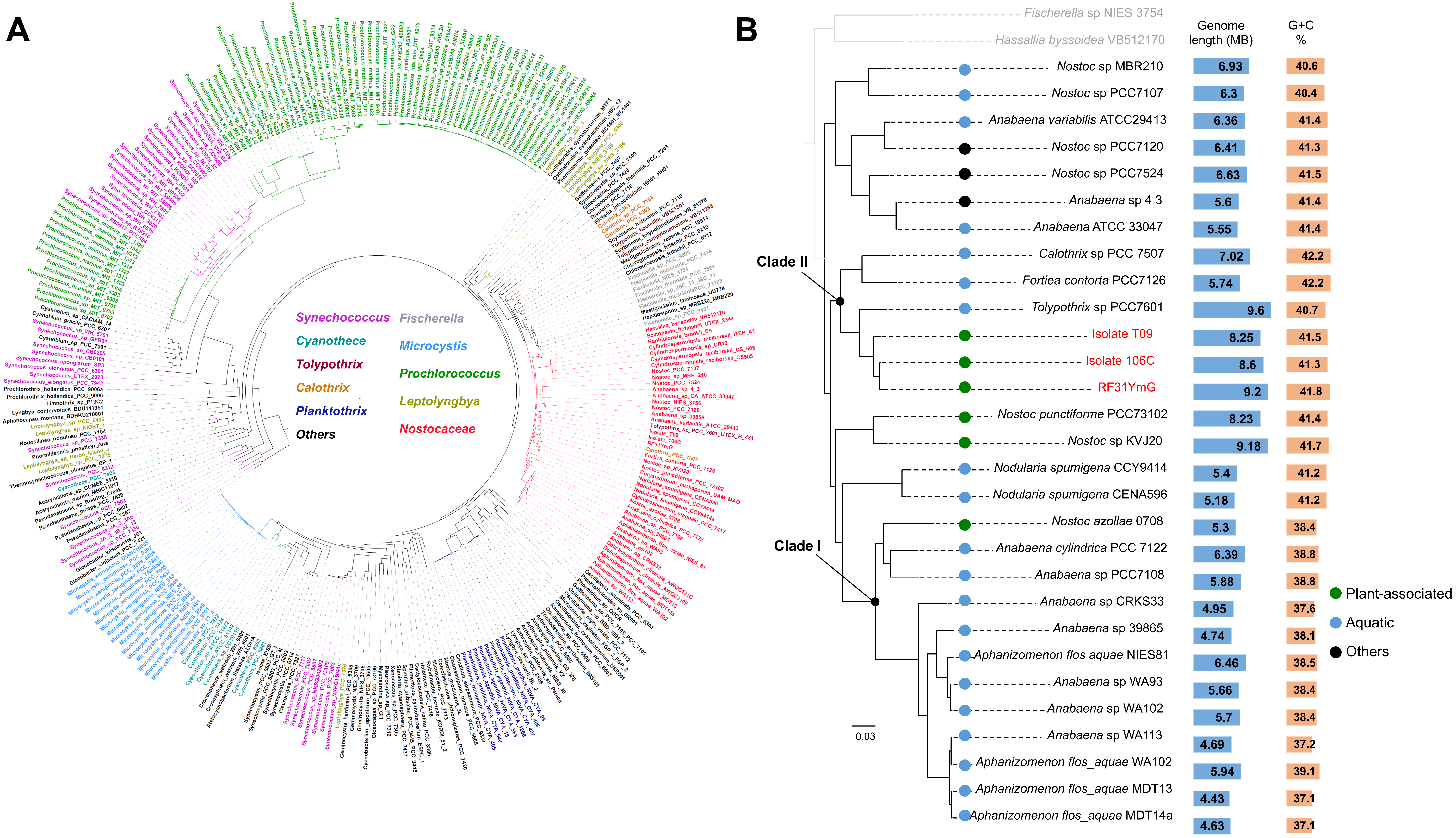
Assembly and annotation
+
Functional predictions

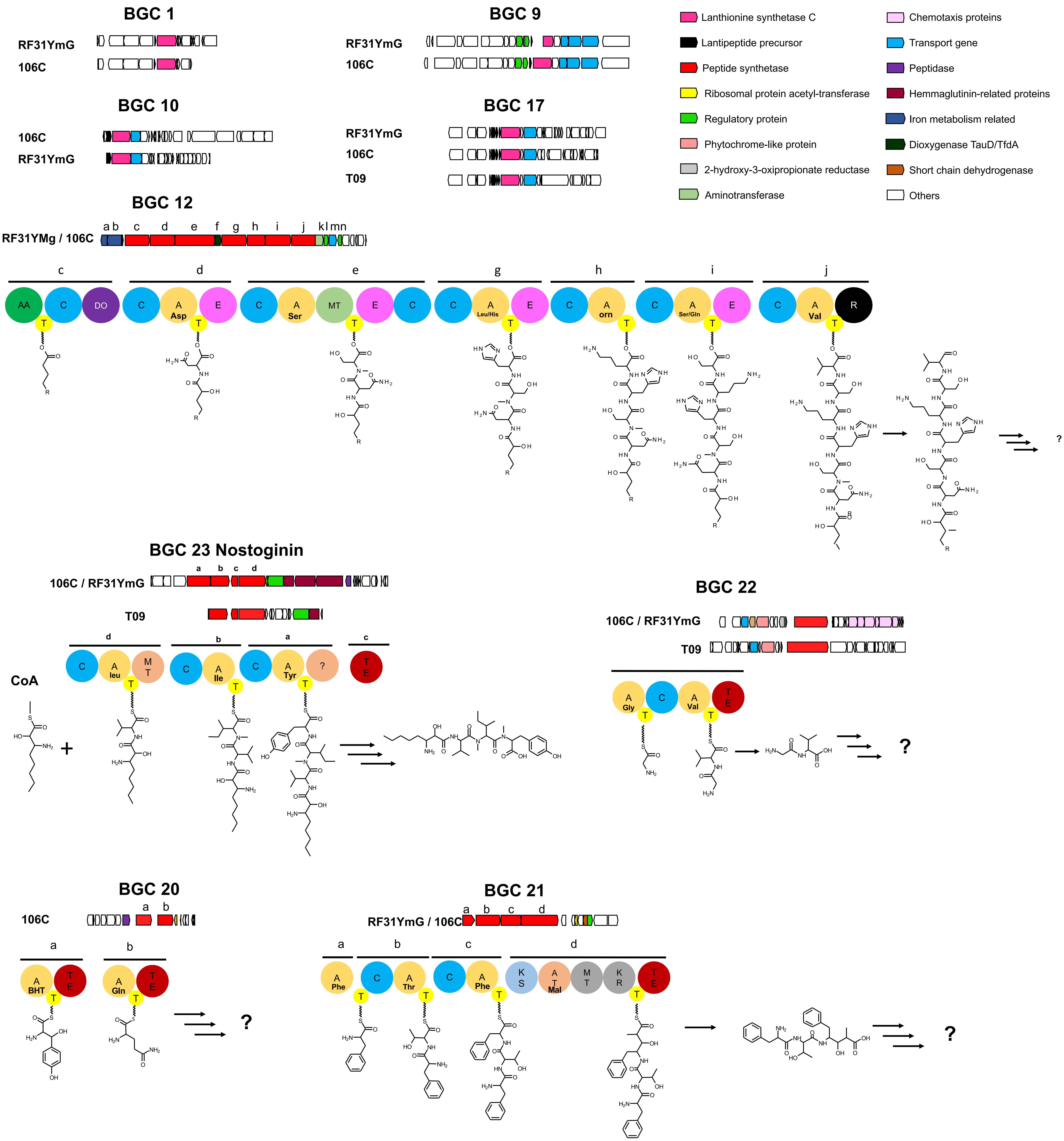
Endophyte co-culture collection

Whole genome sequencing
Phylogenomics


Achromobacter
Ralstonia
Burkholderia
Mitsuaria
Variovorax
Pseudomonas
Pantoea
Enterobacter
Stenotrophomonas
Xanthomonas
Luteibacter
Rhizobium
Shinella
Ochrobactrum
Sphingomonas
Mucilaginibacter
Dyadobacter
Tolypothrix
Nostoc
Bacillus
Paenibacillus
Streptomyces
Microbacterium
Aquifex
t0
t1
H
D
-
N
0.25


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


A
Synechococcus
Cyanothece
Tolypothrix
Calothrix
Planktothrix
Others
Fischerella
Microcystis
Prochlorococcus
Leptolyngbya
Nostocaceae
B
Fischerella sp NIES 3754
Hassallia byssoidea VB512170
Genome length (MB)
G+C %
Nostoc sp MBR210 6.93 40.6
Nostoc sp PCC7107 6.3 40.4
Anabaena variabilis ATCC29413 6.36 41.4
Nostoc sp PCC7120 6.41 41.3
Nostoc sp PCC7524 6.63 41.5
Anabaena sp 4 3 5.6 41.4
Anabaena ATCC 33047 5.55 41.4
Calothrix sp PCC 7507 7.02 42.2
Fortiea contorta PCC7126 5.74 42.2
Tolypothrix sp PCC7601 9.6 40.7
Isolate T09 8.25 41.5
Isolate 106C 8.6 41.3
RF31YmG 9.2 41.8
Nostoc punctiforme PCC73102 8.23 41.4
Nostoc sp KVJ20 9.18 41.7
Nodularia spumigena CCY9414 5.4 41.2
Nodularia spumigena CENA596 5.18 41.2
Nostoc azollae 0708 5.3 38.4
Anabaena cylindrica PCC 7122 6.39 38.8
Anabaena sp PCC7108 5.88 38.8
Anabaena sp CRKS33 4.95 37.6
Anabaena sp 39865 4.74 38.1
Aphanizomenon flos aquae NIES81 6.46 38.5
Anabaena sp WA93 5.66 38.4
Anabaena sp WA102 5.7 38.4
Anabaena sp WA113 4.69 37.2
Aphanizomenon flos_aquae WA102 5.94 39.1
Aphanizomenon flos_aquae MDT13 4.43 37.1
Aphanizomenon flos_aquae MDT14a 4.63 37.1
Clade II
Clade I
0.03
● Plant-associated
● Aquatic
● Others


BGC 1
RF31YmG
106C
BGC 9
RF31YmG
106C
BGC 10
106C
RF31YmG
BGC 17
RF31YmG
106C
T09
Lanthionine synthetase C
Lantipeptide precursor
Peptide synthetase
Ribosomal protein acetyl-transferase
Regulatory protein
Phytochrome-like protein
2-hydroxy-3-oxipropionate reductase
Aminotransferase
Chemotaxis proteins
Transport gene
Peptidase
Hemmaglutinin-related proteins
Iron metabolism related
Dioxygenase TauD/TfdA
Short chain dehydrogenase
Others
BGC 12
RF31YMg / 106C
a b c d e f g h i j k l m n
c
AA C DO T
d
C A Asp E T
e
C A Ser MT E C T
g
C A Leu/His E T
h
C A orn T
i
C A Ser/Gln E T
j
C A Val R T
?
BGC 23 Nostoginin
106C / RF31YmG
a b c d
T09
d
C A leu M T T
b
C A Ile T
a
C A Tyr ? T
c
T E
CoA
+
BGC 22
106C / RF31YmG
T09
A Gly C A Val T E T T
?
BGC 20
106C
a b
a
A BHT T E T
b
A Gln T E T
?
BGC 21
RF31YmG / 106C
a b c d
a
A Phe T
b
C A Thr T
c
C A Phe T
d
K S A T Mal M T K R T E T
?