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], Pablo Cruz-Morales[1,2], Karina G. Gutiérrez-García[1], 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 Lab, 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 bacterial microbiome, 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 the lens of phylogenomics and detailed metabolic annotation. Most of the characterized bacteria were identified as diazotroph plant endophytes belonging to 18 different bacterial families. The genomes of three Cyanobacteria strains were obtained, and after whole-genome inferences were found to form a monophyletic group, suggesting a level of specialization characteristic of co-evolved symbiotic relationships. In agreement with their large size (8.3-9.2 Mbp), the 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, and we hope to contribute 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 clustered 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, and thus with a unique evolutionary history. In natural habitats coralloid roots appear in the most vulnerable early life stages (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 (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. 2010b, 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

69 and transmission of cycad bacterial endophytes (Cuddy et al. 2012). Interestingly, anatomical
70 studies of these Cyanobacteria have shown the presence of mucilaginous or protein-rich material
71 that hosts associated bacteria (Chaudhuri et al. 1931, Caiola 1980, Baulina et al. 2003), whose
72 taxonomic associations have only been suggested on the basis of low resolution data (Bershova
73 et al. 1968, Grobbelaar et al. 1987, Huan et al. 1989 , Thajuddin et al. 2010a, Zvyagintsev et al.
74 2010). Thus, despite progress, there is little investigation of the bacterial endophyte biodiversity
75 other than Cyanobacteria, and almost none from natural habitats.

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

92 Our goal in this study is to investigate the microbiome of the coralloid roots of *Dioon*
93 *merolae* (De Luca et al. 1981), which is a long-lived, entomophilous, dioecious, and arborescent
94 cycad native to Mexico (Lázaro-Zermeño et al. 2012). We collected coralloid root samples from
95 wild populations in two different habitats from their natural range, currently distributed in
96 moderate population sizes of a few hundreds of individuals throughout Chiapas and Oaxaca in
97 the South of Mexico (Lázaro-Zermeño et al. 2012). To overcome technical difficulties in
98 characterizing the breadth of microbial diversity in these environmental samples, we used an
99 enrichment co-culture strategy of sub-communities obtained from the original sample. This
100 approach, termed EcoMining, uses complementing microbiological, (meta)genomic sequencing
101 to characterize the coralloid microbiome's taxonomic and functional diversity. Our study
102 provides evidence of the existence of a diverse endophytic bacterial community inside the
103 coralloid root beyond Cyanobacteria, and suggests functional specificity, i.e. unique BGCs,
104 throughout the evolutionary history of the cycad-cyanobacteria coralloid root system.
105

106 ## Methods

107

108 ***Overall strategy.*** We used a combined sub-community co-culture, metagenomics and
109 phylogenomics strategy to detect and measure taxonomic diversity and phylogenetic
110 relationships in the endophytes of the cycad coralloid root, as previously described (Cibrián-
111 Jaramillo et al. 2016) (fig. 1). In this approach, we grew and isolated bacteria from
112 environmental samples using a diverse set of semi-selective media that aim to capture as much of
113 the cultured bacterial diversity (*t0*). Simultaneously, we enriched the same samples for specific
114 sub-communities in co-cultures grown under conditions for cyanobacteria using BG11 media. In

115 addition to this autotrophic bacterial group, this second approach aims to capture other bacterial
116 groups that interact with and depend on cyanobacteria, present in the original sample at low titers
117 (rare taxa). We allowed the sub-community co-culture to grow over time and sampled it after one
118 month (*t1*) and at the end of a year (*t2*). This involved isolation of bacteria grown in selective
119 media over time (*t0* and *t1*) and sequencing sub-communities in co-cultures using shotgun
120 metagenomics (*t1* and *t2*). The phylogenetic relationships of the isolated bacteria were
121 reconstructed using amplified 16S rRNA sequences and OTUs obtained from metagenomes. To
122 further assess the natural history of the cyanobacteria-cycad interaction, two cyanobacterial
123 genomes and one genome reconstructed from a metagenomic sub-community, were obtained and
124 analyzed. Specific methods, materials and analytical tools are described in the following sub-
125 sections.
126
127 ***Field collections***. We sampled coralloid roots from two wild cycad populations previously
128 reported (Lázaro-Zermeño et al. 2012). In March of 2014 we sampled from two sites in
129 deciduous tropical forests, at Jiquipilas, Oaxaca Mexico (JP or dry) at 560m above mean sea
130 level, with an average annual precipitation of 320 mm and average annual temperature of 18 °C;
131 and Raymundo Flores, Oaxaca, Mexico (RF or humid) at 900m above mean sea level, with 2500
132 mm and 25°C annual average precipitation and temperature, respectively (exact coordinates upon
133 request to preserve the integrity of these populations). In some cycad plants, coralloid roots were
134 easily visible aboveground, while in others it was necessary to dig to about 30 cm around the
135 main root until coralloid roots were found. In a population of approximately 40 individuals, we
136 generally found 10-12 plants with coralloid roots, in almost exclusively juvenile plants. A total
137 of 10 coralloid apogeotropic roots were cut from 10 plants, cleaned with sterile distilled water to

138 remove soil excess, placed in 15 ml sterile Falcon tubes (Beckton Dickinson), and transported
139 immediately to the laboratory at room temperature.
140
141 ***Coralloid root processing.*** We focused our effort on two samples of two individuals with the
142 largest coralloid roots, in each of the two sites, Jiquipilas (JP or dry) and Raymundo Flores (RF
143 or humid) for a total of four coralloid root samples (JP2 and JP6 plus RF1 and RF3,
144 respectively), and stored the remaining samples at -80 °C for subsequent studies. We treated the
145 coralloid root in a laminar flow hood (Nuaire Model Nu-126-400) with a series of washes to
146 remove exogenous bacteria from the rhizosphere or epiphytic contamination. Each root was
147 introduced in 50 ml sterile Falcon tubes containing 10 ml of each of the following solutions, and
148 gently stirred for: three minutes in hydrogen peroxide (H_2O_2), seven minutes in 70% ethanol, 30
149 seconds in sterile dd-MilliQ water, four minutes in 6% sodium hypochlorite (NaClO), and three
150 one-minute washes in sterile dd-MilliQ water. After this procedure, we plated out water from the
151 last wash in Petri dishes containing the six media described below. Lack of growth in the last
152 wash was considered a negative control. Only samples complying with this criterion were used
153 for endophyte isolation. We undertook two approaches to bacterial isolation (fig. 1): sampling
154 without enrichment directly from field samples (*t0*), and sampling from the enriched co-cultures
155 (*t1*), as described in the following sections.
156
157 ***Bacterial growth and isolation in selective media.*** To isolate bacteria from field samples before
158 (*t0*) and after (*t1, t2*) enrichment, macerated roots or co-culture broth were used as inoculant,
159 respectively. Loss of some bacterial groups that were present in the sample collected from the
160 environment (*t0*) is expected. However, after enrichment (*t1*) we recover bacteria expected to be

161 initially present in low abundances and required time to grow, and that did so as a response to the
162 community nutritional interactions (e.g. amino acids derived from the process of fixing nitrogen)
163 (Cibrián-Jaramillo et al. 2016). Roots were macerated in 10 ml of sterile water using a pestle and
164 mortar until plant material was completely disintegrated. We used 100 μl from the root macerate
165 to directly isolate bacteria in Petri dishes containing a total of six different media, chosen to
166 selectively (oligotrophic, four media) or non-selectively (eutrophic, two media) recover bacterial
167 diversity as much as possible. The four selective media used were chosen to target bacterial
168 groups that are known to be either plant endophytes or rhizosphere bacteria, and included: 1)
169 *Caulobacter* medium (glucose: 1 g/L; peptone: 1g/L; yeast extract: 1.5 g/L; trace metals solution:
170 1 mL/L; and 10 g/L of agar for solid medium); 2) *Rhizobium* medium (mannitol: 10 g/L;
171 dipotassium phosphate: 0.5 g/L; magnesium sulfate: 0.2 g/L; yeast extract: 1 g/L; sodium
172 chloride: 0.1 g/L; final pH 6.8; and 20 g/L for solid medium; 3) ISP4, for isolation of
173 actinomycetes (starch: 10.0 g/L; dipotassium phosphate: 1 g/L; magnesium sulfate: 1 g/L;
174 sodium chloride: 1 g/L; ammonium sulfate: 2 g/L; calcium carbonate: 2 g/L; ferrous sulfate: 1
175 mg/L; magnesium chloride: 1 mg/L; zinc sulfate: 1 mg/L; final pH 7.0; and 20 g/L for solid
176 media); 4) BG-11, a Cyanobacteria medium (sodium nitrate: 1.5 g/L; dipotassium phosphate:
177 0.04 g/L; magnesium sulfate: 0.075 g/L; calcium chloride: 0.036 g/L; citric acid: 0.006 g/L;
178 ferric ammonium citrate: 0.006 g/L; EDTA (disodium salt): 0.001 g/L; sodium carbonate: 0.02
179 g/L; final pH 7.1 and agar solid media 10.0 gr/L. The non-selective, rich media, included: 5)
180 Nutrient Broth (BD Bioxon, Mexico); and 6) As in *Caulobacter* medium, but supplemented with
181 mannitol (*Caulobacter* + mannitol medium): 1g/L, with aim of providing a carbon source closer
182 to that hypothetically encountered inside the cycad root.

183 ***Bacterial growth and cultivation in sub-communities.*** 100 μl of the macerated roots from
184 samples JP2, JP6 and RF1 and RF3, which passed the negative growth control after the final
185 washing step, were used to inoculate 100 ml of media in 250 ml flasks. The remaining macerated
186 roots not used for fresh cultures were kept as frozen stocks for future studies (-80 °C in 10%
187 glycerol), although community viability after freezing was found to diminish over time. We used
188 selective oligotrophic medium, i.e. BG11 (medium No. 4) for these experiments. This medium
189 lacks a carbon source but contains a limited amount of inorganic nitrogen. BG11 cyanobacteria-
190 centric co-cultures were grown for up to one year with constant stirring, with cycles of 16/8
191 hours of light/darkness per day. Sampling of the oligotrophic sub-communities was done after
192 one month (*t1*) and one year (*t2*), and treated independently. Bacterial isolates were only
193 obtained from *t1* whereas shotgun metagenomics were obtained for both time points.
194
195 ***Genomics and shotgun metagenomics.*** The genomic sequence of the cyanobacterial isolate
196 106C was obtained from sample JP6. To test for host differences, we also obtained a genome
197 from coralloid roots of *Dioon caputoi* from an unrelated environment (Xeric shrubland,
198 Tehuacan valley Puebla). Both organisms were grown on BG11 plates before DNA extraction,
199 which was done using a CTAB-phenol chloroform standard protocol. Likewise, metagenomic
200 DNA of all enriched sub-community co-cultures, coming from RF and JP samples, were
201 extracted using the same protocol, with the excpetion that biomass was obtained from liquid
202 cultures by high-speed centrifugation during 15 minutes. The RF31Ymg genome was obtained
203 from one of this metagenomic DNA samples, i.e. RF3-1yr or *t2*, after bioinformatic treatment as
204 described next. DNA samples were processed with truseq nano kit Q28 and sequenced at
205 Langebio, Cinvestav (Irapuato, Mexico) using the MiSeq Illumina platform in the 2X250 paired

206 end reads format (T09) and the NextSeq mid output 2X150 paired end read format (106C y RF3-
207 1yr or *t2*). The reads for each library were filtered with fastQ and trimmed using Trimommatic
208 version 0.32 (Bolger et al. 2014). Assembly of 106C and T09 was done *de novo*, whereas
209 RF31YmG genome was obtained from metagenomic reads of the sub-community co-culture
210 RF3-*t2* of *D. merolae*, using as reference the newly obtained sequence of 106C. The
211 metagenomic reads were filtered by mapping them against the assembly of *Nostoc* sp. 106C with
212 BWA (Li et al. 2009). In all cases the resulting reads were assembled with Velvet (Zerbino et al.
213 2008) using different *k*-mers: the assemblies with the largest length and the smaller number of
214 contigs were selected and annotated using RAST (Aziz et al. 2008).
215

216 ***Taxonomic diversity***. We first estimated taxonomic diversity using the 16S rRNA gene as a
217 marker, obtained after PCR amplification, for our entire bacterial endophyte collection (*t0* and
218 *t1*). PCR products of 1.4 Kbp in length, obtained using the F27 and R1492 primers (Lane 1991),
219 were obtained and sequenced using the Sanger method (ABI 3730xl). The taxonomic
220 identification was made using Blastn with an initial cut-off e-value of 1e-5 against the SILVA
221 database (Quast et al. 2013) and above 98% identity levels. We used the phylogenetic position of
222 the top 10 hits from each search without duplicated matches, to determine both taxonomic
223 diversity and phylogenetic relationships. We only report genus-level taxonomy as species
224 determination can be biased by reference 16S rRNA databases in bacteria.
225 We used Kraken (http://ccb.jhu.edu/software/kraken/) to assign taxonomic labels to
226 metagenomic DNA sequences based on exact alignment of *k*-mers (Wood et al. 2014). Kraken
227 was used to exclude sequence contaminants from the draft assembly, allowing us to generate a
228 symbiotic cyanobacteria marker database as reference for future classification. We made a

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

235 We estimated the efficiency of our sequencing effort with respect to the total possible
236 taxa per metagenome using a previously reported rarefaction method (Hurlbert 1971). To do this
237 we calculated the proportional number of sequences for each metagenome, in which the richness
238 of mOTUs is sub-sampled randomly from the total of the community. The Shannon-Weaver *H'*
239 and Simpson *L* indices for OTUs from different environments were calculated. In both cases we
240 used th Vegan v2.3-5 package in R (Oksanen J 2015). To visualize shared taxa among
241 metagenomes, and their abundance, we used Cytoscape v3.4.0 (Shannon 2003), where each node
242 and its size represent the abundance of an OTU, and lines represent shared taxa between
243 metagenomes. The network was made by an interaction matrix, where each of the OTUs that had
244 more than 14 readings assigned directly by Kraken, was linked to its cognate metagenome.
245 Identified nodes were manually ordered to prevent visual overlap.
246

247 ***Reconstruction of phylogenomic relationships.*** We aligned annotated 16S rRNA sequences
248 trimmed to 1.1 Kbp, using MUSCLE v3.8.31 with default parameters (Edgar 2004). This matrix
249 was used for phylogenetic reconstruction using MrBayes v3.2 (Ronquist et al. 2012) with a
250 gamma distribution type range with 1,000,000 generations. ModelTest (Posada et al. 1998)
251 showed that Kimura 2 parameters was the best substitution model. To explore major clades in

252 more detail, we estimated individual phylogenies for each of the genera in our main tree and
253 represented them graphically. To do this we first recovered a tree by generating a consensus
254 sequence from all genera within each clade, then a Bayesian phylogeny with a gamma
255 distribution and a million generations (additional generations did not change our results) was
256 reconstructed using MrBayes v3.2 for each individual genus dataset. The resulting trees were
257 edited and sorted with Environment for Tree Exploration Toolkit v3.0.0b35 (Huerta-Cepas et al.
258 2010) in Python v2.7.6.

259 To construct a complete phylogeny of Cyanobacteria strains using their whole genomes,
260 we adopted a phylogenomics approach that we termed OrthoCores based on our in-house scripts
261 (https://github.com/nselem/orthocore/wiki, dockerhub: nselem/orthocores). Our method is based
262 on the premise that single gene or even housekeeping multilocus phylogenies are often not
263 enough to resolve the evolutionary history of certain taxonomic groups, and thus total evidence
264 tree evidence approaches must be used (Narechania et al. 2012). We first obtain a set of
265 orthologous groups that are shared among all high-quality genomes, or the 'core genome',
266 excluding paralogs and signatures of expansions (i.e. not all genes from a gene family are part of
267 the Best Bidirectional Hits, BBH set). We define orthologs from the core genome as hits that
268 satisfy that all gene members are BHHs all vs all (fig. s1). To place our samples within the
269 Cyanobacteria phylogeny we used OrthoCores to generate a tree using representative genera
270 from different families sampled from all Cyanobacteria groups. We selected closed and well
271 annotated genomes, although for Nostocales we used all the publicly available genomes. We
272 obtained representatives from Chroococcidiopsidales, Oscillatoriales, Synechococcales,
273 Chroococcales, Pleurocapsales, Gloeobacterales and Nostocales publically available (July 2016),
274 and included the three genome sequences obtained in this study in the phylogeny.

275 The OrthoCores cut-off e-value was 1e-6, although most genes had an 1e-50 score given
276 our restrictive all vs all criteria. Set of sequences were concatenated and aligned using MUSCLE
277 v3.8.31 with default parameters (Edgar 2004), and trimmed by Gblocks (Talavera et al. 2007)
278 with five positions as minimum block length, 10 as a maximum number of contiguous non-
279 conserved positions and only considering positions with a gap in less than 50% of the sequences
280 in the final alignment. The final matrix included 76 taxa and 198 core protein sequences with
281 45475 amino acids that represent a set of 76 cyanobacterial genomes including our two isolates
282 (T09 and 106C) (table s1), plus the RF31YmG metagenomic assembly, and *Gloeobacter*
283 *kilaueensis* JS1 and *Gloeobacter violaceus* PCC 7421 as outgroups. This matrix was used to
284 reconstruct a tree with MrBayes, using a mixed substitution model based on posterior
285 probabilities (aamodel[Wag]1.000) for proteins for a 1 million generations. The tree was
286 visualized with jsPhyloSVG (Smits Sa et al. 2010), and once we identified our samples as
287 Nostocales, we highlighted key traits of these organisms, such as genome size, habitat, and GC
288 content of them and their closest taxa. Metadata were added using PhyD3 (Kreft L et al. 2017).
289 Finally, we examined the level of colinearity among the genomes closest to our samples, using
290 R2cat with default paramenters (Husemann et al. 2010).
291
292 ***Genome mining for BGCs.*** To identify BGCs potentially directing the synthesis of specialized
293 metabolites among selected Cyanobacteria, we annotated the genome of the isolate 106C with
294 antiSMASH (Weber et al. 2015). The predicted BGCs in this strain were used as a reference for
295 subsequent searches in all of the selected genomes, which allowed us to identified both
296 conserved and unique BGCs in each genome. Each BGC from antiSMASH was searched against
297 our Cyanobacteria genome database as on NCBI (July 2016) (table s2). We considered that a

298 BGC hit was found whenever a selected enzyme, plus another enzyme for the BGC, were found
299 in a vicinity of 10 genes on a cyanobacterial genome, within a blast e-value cut-off of .001, using
300 in-house scripts. Prediction of the chemical structures of the putative specialized metabolites
301 associated with these BGCs was done after domain identification and specificity prediction,
302 mainly of adenylation and acyl transfer domains, with the non-ribosomal peptide synthetase
303 (NRPS) and polyketide synthases (PKS) NRPS-PKS server (Bachmann et al. 2009), PRISM
304 (Skinnider et al. 2015) and antiSMASH. In addition to analyzing single genomes including
305 RF31YmG, we also assembled, annotated and mined *de novo*, the complete metagenomes of *t1*
306 and *t2* sub-community co-cultures in an iterative fashion. First, by identifying sequence
307 signatures of biosynthetic enzymes using antiSMASH, and second, by extending the contigs with
308 hits by iterative mapping and assembly. We only recovered five short signal sequences (less than
309 3.5 Kbp) for all metagenomes obtained at *t1*, that are suggestive of enzyme genes that could be
310 part of BGCs. Thus, these datasets were not used as part of our genome mining efforts, and only
311 whole genome sequences were mined for BGCs.
312

313 **Results**

314

315 Our experimental strategy to characterize the taxonomic and functional diversity of the coralloid
316 root endophytic microbiome led to hundreds of bacterial isolates obtained directly from the
317 original sample (*t0)*; and from enriched sub-communities in oligotrophic (BG11) medium (*t1*)
318 aimed at promoting interactions between members of the coralloid root community. Individual
319 markers and genomic sequences obtained from these isolates captured the taxonomic diversity of
320 endophytes living in the coralloid root, including bacteria initially present in low titers (rare) in

321 the original sample (*t2*). In the following sections we describe the results obtained from this
322 effort, including the overall bacterial taxonomic diversity, and the phylogenomic relationships of
323 the cyanobacteria present in the *Dioon* coralloid roots; as well as insights of metabolic
324 specialization of the cyanobacteria-cycad interaction, after comprehensive genome mining of
325 natural products BGCs.
326
327 ***Dioon* coralloid roots show ample endophyte diversity of taxa beyond and within**
328 **cyanobacteria**
329
330 *Taxa assessment using 16S rRNA*. Cultivable bacteria constitute only a biased subset of the total
331 endophyte biodiversity, yet from our 16S rRNA sequences alone we found 470 isolates grouped
332 into 242 OTUs, distributed in 18 families and 11 bacterial orders, with 27 genera in total,
333 representing most of the known bacterial groups (table 1; see also table s3). As seen in our 16S
334 rRNA phylogenetic reconstruction (fig. 2), all of our sequences grouped within monophyletic
335 clades, and most trees within each clade show that there are novel isolates that remain to be
336 described, in almost all of the genera found within the cycad coralloid root (see also fig. s2 for a
337 detailed phylogeny). An 87% of the taxa identified can be taxonomically classified as
338 diazotrophic plant endophytes, validating our endophyte isolation procedures and supporting the
339 notion of N_2 fixation. Most OTUs grouped within the genera *Streptomyces, Bacillus, Rhizobium,*
340 *Stenotrophomonas, Pseudomonas, Mitsuaria, Achromobacter* and *Burkholderia*.
341 Our results also clarified earlier reports of bacteria associated to the cycad coralloid root,
342 including *Bacillus*, which was reported as associated to the outside of the coralloid root;
343 *Streptomyces*, isolated as an epiphyte (Lobakova et al. 2003), which grew on one of our selective

344 media (ISP4); and *Pseudomonas* (Bershova et al. 1968), growing indistinctly in our four semi
345 selective media. We identified six isolates as *Tolypothrix* spp, a previously unreported genus of
346 Nostocales living in the coralloid root, and isolated two strains identified as *Nostoc* spp, which
347 confirmed earlier reports of this genus present in the coralloid root (Caiola 1980). We sequenced
348 the genome of one of these isolates, strain 106C, as further discussed in following sections.
349 Presumably, if the coralloid root is adapted to specific environments, then species
350 composition, diversity and abundance should differ among collection sites (dry vs humid), and
351 should also be independent of the media they were cultivated in. Our phylogenetic results show
352 that OTUs are mostly shared with no specific distribution among the various isolation culture
353 media (fig. 2). Some genera occur only in dry (JP; e.g. *Rhizobium*), or only in humid (RF; e.g.
354 *Xanthomonas*) forest environments, with a few genera occurring in both (e.g. *Burkholderia*).
355 Diversity and abundance measured as the Shannon-Weaver and Simpson biodiversity indices
356 were H'=1.50, L=3.8 in the dry environment and H'=1.57, L= 3.76 in the humid environment
357 (table s4) are similar. Thus it seems that there are some trends in different environments, but we
358 consider these results preliminary and limited by the use of cultivable approaches.
359
360 *Taxa assessment based in sub-community co-cultures characterized with metagenomics*. We
361 extracted and sequenced whole-community metagenomic DNA from *t1* and *t2* subcommunities
362 with the aim of enriching for specific interactions in response to oligotrophic (BG11) growth
363 conditions. Basic indicators of metagenome quality are shown in (table s5), and rarefaction
364 curves suggest we captured 40-60% of the microbial community growing in the BG11 media
365 (fig. s3). Metagenomic sequencing of endophytic sub-communities was successful for the
366 following individuals, grown under the same conditions: two from each of the two environments

367 (JP2, JP6 and RF1, RF3) after 30 days (*t1*); and two after 365 days (*t2*) one from each
368 environment (JP6 and RF3).
369 With respect to identity and abundance of taxa, in the 30 days sample (*t1*), we found the
370 following (total taxa/genus/most abundant in percentage): JP2, 57/*Rhizobium* (7%); JP6,
371 69/*Xanthomonas* (2%); RF1, 63/*Nostoc* (3%); and RF3 61/*Xanthomonas* (7%). As expected, we
372 observed different bacterial proportions for the 1-year samples (*t2*): JP6, 69/*Deinococcus* (4%),
373 RF3, 63/*Nostoc* (3%), and in both *t1* and *t2* samples *Nostoc* is present. When the overall taxa is
374 evaluated, the most abundant genera from the four metagenomes from humid and dry
375 environments mostly overlap (fig. 3, table s6). There are some exceptions that were unique to an
376 environment, which include genera such as *Shewanella*, only found in JP2 from the dry
377 environment, and *Cronobacter*, only found in RF3 in the humid environment. These results
378 support the trends found with 16S rRNA, in which the abundant taxa are shared within the
379 analyzed coralloid root samples, with exceptions of some genera.
380 Analyses using Kraken revealed an overall proportion of 12% Cyanobacteria, yet we
381 failed to detect *Tolypothrix* in any of the metagenomes, identified with 16S rRNA in the previous
382 section. We did find *Anabaena* (Chaudhuri et al. 1931), whose presence in cycads was
383 previously questioned; *Nostoc* (Thajuddin et al. 2010a), the mainly reported genus in cycads
384 before this study; and *Calothrix*, previously reported in *Encephalartos* (Grobbelaar et al. 1987,
385 Huan et al. 1989) and in *Cycas revoluta* (Thajuddin et al. 2010a). Notably, we identified
386 *Caulobacter*, which to our knowledge is the first report of this genus as a cycad endophyte. The
387 importance of this observation relies on the fact that *Caulobacter* can be found associated to
388 unicellular green algae (Kim Bh et al. 2014) and has been found to enhance growth of *Anabaena*
389 cyanobacteria (Berg et al. 2008).

390 The taxonomic diversity from the cultivated environmental isolates (*t0*) using 16S rRNA,
391 and the diversity found in the sub-community co-cultures (*t1*) measured by Kraken, overlap only
392 in 20.8% of the genera. Differences in genera, such as the absence of *Tolypothrix* from the
393 metagenomes, could be explained because our sequencing efforts were not deep enough to
394 recover cyanobacteria-associated OTUs; because taxa presence may fluctuate in the cultures;
395 and/or because Cycanobacteria sequences are too divergent to be captured or there are errors of
396 annotation in the 16S rRNA databases. It is likely that all these factors influenced our results.
397 Despite these issues, we confirmed the occurrence of several of the bacterial endophyte
398 taxonomic groups in the metagenomes, which we also isolated and characterized with 16S
399 rRNA. Thus, it can be concluded that the taxonomic diversity of the endophytes present in the
400 cycad coralloid root captured by our combined approach, is a valid representation of the natural
401 circumstances.
402
403 ***Dioon* cyanobacteria are a monophyletic group with large genomes within the Nostocales**
404
405 Given that Cyanobacteria phylogenetic history is likely reticulated (Zhaxybayeva et al. 2006) we
406 adopted a whole-genome inference approach to explore the phylogenetic position of our
407 cyanobacterial isolates. Using OrthoCores we reconstructed a phylogeny from 198 proteins
408 identified as the core genome in 76 selected genomes that provide a broad taxonomic coverage
409 and represent good quality sequences. We then used this phylogeny to map general genome
410 features.
411 We found lack of monophyly of several Cyanobacteria groups in our whole-genome
412 phylogeny, with the exception of the Nostocales that were monophyletic (fig. 4a). Our isolates

413 T09 from *D. caputoi*, and 106C / RF31YmG from *D. merolae,* form a monophyletic clade, and
414 group with *Tolypothrix* sp PCC 7601. The *Dioon*-specific clade is sister to *Calothrix* sp. PCC
415 7507 and *Fortiea contorta* PCC7126. When examining the synteny of our isolates with their
416 closest relatives (fig. s4), we observe that they are more similar to each other, and to *Tolypothrix*
417 sp PCC 7601, and have less genome conservation when compared to *Calothrix* sp. PCC 7507,
418 *Fortiea contorta* PCC7126, and surprisingly, to *Nostoc punctiforme* from the cycad *Macrozamia*.
419 Thus, although the genus and species of our isolates remains to be determined, we can conclude
420 that our cyanobacteria endophytes belong to the Nostocales, have genome conservation, are
421 closely related despite different host origin, and appear to be specific to *Dioon*.
422 These results contrast with previous studies in which different species of cycads host
423 multiple cyanobacteria and do not form cycad or host-specific clades (Lindblad et al. 1989,
424 Zimmerman et al. 1992, Zheng W et al. 2002), but are congruent with a 16S rRNA phylogeny of
425 Nostocaceae cyanobacteria which shows that hormogonia-producing species symbiotic to
426 *Gunnera* ferns, *Anthoceros*, and cycads, tend to cluster together (Papaefthimiou et al. 2008).
427 Our phylogeny also shows diverging evolutionary trajectories of *Nostoc* species
428 associated with cycads, from those of the free-living Nostocaceae and other *Nostoc* associated
429 with plants (fig. 4b). The *Nostoc* free-living PCC 7120 grouped distantly to strains of symbiotic
430 or plant-associated origin, while the *Nostoc azollae* and other plant-associated *Nostoc* are closely
431 related to some *Anabaena* and *Aphanizomenon,* congruently with previous reports
432 (Papaefthimiou et al. 2008).
433 When we mapped genome size and G+C content traits onto the phylogeny, we found that
434 although there are no differences in G+C content, our coralloid endosymbiont cyanobacteria and
435 related taxa have larger genomes compared to other clades (fig. 4b). The exception to this is

436 *Hassallia byssoidea,* a soil-dwelling bacteria that has a genome of similar size to our cycad
437 symbionts, but it is phylogenetically distant. Notably, the other taxa with large genomes in
438 Nostocales include their sister taxa *Tolypothrix* sp PCC 7601, and two other plant symbionts:
439 *Nostoc* sp KVJ20 which lives in special cavities located on the ventral surface of the
440 gametophyte of the Norway liverwort *Blasia pusilla* (Liaimer 2016); and *Nostoc punctiforme*
441 PCC73102, associated with the Australian cycad *Macrozamia* (Meeks et al. 2001). Interestingly,
442 the closest genome at the synteny level outside of our own samples is KVJ20, suggesting
443 convergence between the divergent plant niches hosting these bacteria.
444

445 **Identification of BGCs in sub-community metagenomes suggests metabolic specialization of**
446 ***Dioon* cyanobacteria.**
447

448 Given that large genomes correlate with the ability of bacteria to produce specialized metabolites
449 (Doroghazi et al. 2014), we aimed at exploring the functional specialization of the coralloid root
450 microbiome by identifying examples of BGCs putatively directing the synthesis of specialized
451 metabolites. Genome mining of isolate 106C revealed 18 BGCs, which are overall well
452 conserved in the related RF31YmG genomic sequence (table s7). Interestingly, none of the
453 known molecules previously associated with cycad cyanobionts, such as nodularin (Gehringer et
454 al. 2012), were identified. Moreover, the analysis of the distribution of these 18 BGCs
455 throughout the selected genomes (table s8) revealed that only the well-characterized heterocyst
456 glycolipid (Soriente et al. 1992), termed here BGC 16; and BGC 2, putatively directing the
457 synthesis of a terpene of unknown structure, are conserved in the analyzed genomes. In contrast,
458 we did find that up to half of the identified BGCs were uniquely present within *Dioon* symbionts.

459 Among these BGCs we could annotate four coding for lantipeptides, namely, BGC 1, 9, 10, 17;
460 and five coding for NRPS-containing systems, of which some may contain additional PKSs,
461 namely, BGC 12, 20, 21, 22, 23 (fig. 5, text s1). It should be noted that all 9 BGCs are absent
462 from *Nostoc punctiforme* PCC73102, a strain isolated from an Australian *Macrozamia* that has
463 been recently re-sequenced (Moraes et al. 2017).

464 BGC 12 may provide an example of a specialized metabolite in the coralloid root with a
465 functional role. Our sequence-based predictions suggest that this BGC may code for an assembly
466 line capable of synthesizing an N-terminal acylated hexapeptide with several modifications, such
467 as the epimerization of four of its residues, the N-acylation of its second amidic bond, and the
468 reduction of its C-terminal end to yield an aldehyde group. The N and C terminal modifications
469 on this peptide would resemble small peptide aldehyde protease inhibitors, which have been
470 previously reported on cyanobacteria (Fewer et al. 2013). Alternatively, the product of this
471 biosynthetic system may be a siderophore, as iron-related genes were found next to the NRPS
472 coding-genes, which include a reductase domain. The latter observation is significant as previous
473 reports implicate reductase-containing NRPS systems in the synthesis of iron chelators, e.g.
474 myxochelin biosynthesis (Li et al. 2008).

475 BGC 23, the most interesting of all our predictions, seems to code for a NRPS system
476 putatively directing the synthesis of a tripeptide consisting of leucine, valine and tyrosine
477 residues. This prediction also includes informative modifications, such as N-terminal acylation,
478 N-methylation at an amide bond of the isoleucine residue, and modification of the tyrosine
479 residue by a domain of unknown function. A search for peptides containing such modifications,
480 performed with the server PRISM with the feature for de-replication of known chemical
481 structures (Skinnider et al. 2015), directed our attention to nostoginins, specialized metabolites

482 originally isolated from *Nostoc* (Ploutno et al. 2002), whose biosynthetic pathway remains to be
483 identified. Indeed, nostoginins A and B, which are N-acylated Leucine-Valine-Tyrosine peptides
484 have been shown to have protease inhibitory activity, as well as their congeners microginins FR1
485 / SD755 and oscillaginins A / B (Sano et al. 1998). BGC 23 includes a peptidase-coding gene
486 (fig. 5), which warrants further biochemical and functional characterization of 106C with regards
487 to this prediction. Thus far, it is clear that the *Dioon*-specific cyanobionts have distinctive
488 metabolic capabilities that may result from the evolution of these organisms within a diverse
489 microbiome present in the coralloid root.
490
491 **Discussion**
492
493 Our evidence undoubtedly shows that within the cycad coralloid root there is a diverse bacterial
494 community that includes all of the previously reported Nostocales, other taxa for which their
495 endophytic origin and presence was unclear, and newly reported genera. We support previous
496 morphological observations that showed that an individual cycad plant could hold diverse
497 communities that differ in their taxonomic composition and life strategies (Lobakova et al.
498 2003). Morphological studies observing mucilaginous material inside the coralloid root (Ow et
499 al. 1999, Baulina et al. 2003) are also congruent with the microbiome consortium we describe.
500 Overall, many shared genera among samples obtained from plants growing in different
501 environmental conditions, including mostly diazotrophic plant endophytes, point toward
502 functional congruence in a taxonomically diverse coralloid microbiome.
503 In this study we increased our ability to identify a diverse array of organisms using co-
504 cultured bacterial sub-communities (*t1, t2*). The differences in genera identified following this

505 approach in comparison with direct isolation (*t0*) was expected, because environmental sampling
506 and enriched inoculant complement each other, and aim to recover distinct aspects of the
507 microbiome's composition (Cibrián-Jaramillo et al. 2016). These patterns can also be explained
508 by various scenarios: i) rare groups present in low abundance can only be recovered in sub-
509 community co-cultures on which they increase in biomass; ii) some organisms are fast growers
510 irrespective of media, and will dominate in OTUs, simply by chance, iii) some groups are more
511 media-specific; and/or iv) groups in BG11 (*t1*) are recovered as a result of functional interactions
512 in cyanobacteria-associated groups.

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

523 Notably, we recover a monophyletic placement of our cyanobacterial samples, which
524 confirms a previous single morphological observation of possible specificity among
525 cyanobacteria coralloid root endophytes (previously termed phycobionts), and their hosts,
526 including *Dioon* (Caiola 1980). This notion contrasts with other reports that assume no
527 specificity based on the observation that different strains isolated from one plant species are

528 capable of infecting phylogenetically distant hosts (Johansson et al. 1994, Papaefthimiou et al.
529 2008, Whitton 2012). Partial taxon identification and lack of samples from their native habitats
530 could confound the identification of co-evolutionary signatures between plant host and the
531 bacterial microbiome in those previous studies.

532 Interestingly, our endosymbiont cyanobacterial isolates have larger genome sizes than all
533 other closely related taxa in their phylogeny. This observation contrasts with most bacterial
534 endosymbionts in eukaryotes that show a reduction in genome size compared to free-living
535 relatives (Mccutcheon 2010). Large genome sizes in endosymbionts are usually attributed to a
536 facultative relationship that requires retaining free-living stages (Maclean et al. 2007). In
537 angiosperms facultative symbionts that form nitrogen-fixing root nodules have large genome
538 sizes adapted to shifting from the soil to the plant environment (Normand et al. 2007).
539 Congruently, the only other cyanobacteria cycad symbiont whose genome has been sequenced,
540 *Nostoc punctiforme* from an African cycad *Macrozamia*, has variation within its large genome
541 required to sense and respond to the environment outside and inside the plant (Meeks et al.
542 2001). Thus, a feasible yet-to-be-tested hypothesis is that the Nostocaceae taxa associated to the
543 cycad coralloid root have experienced genome expansion driven by selection to initially survive
544 the complexity of the environment from which they are recruited, but retain the complex life
545 cycles involving cell differentiation needed to establish an advantageous relationship with the
546 cycad plant (Meeks et al. 2002), and with other members of the microbiome inside the coralloid
547 root.

548 The genome synteny observed within the cycad isolates, together with the nine BGCs
549 unique to the *Dioon*-specific cyanobionts reported here, is in agreement with the hypothesis of a
550 specialized community driven by metabolic adaptation. Of the nine unique BGCs found in our

551 isolates, BGC 12 and BGC 23 putatively coding for two small peptide protease inhibitors, or a
552 siderophore, could have a strong bearing on the evolution and biology of the *Dioon*-bacteria
553 symbiosis. Peptide production may imply that proteolysis is involved in the cyanobacteria-cycad
554 interaction, linked to the reconfiguration of the root architecture or the filtering of the
555 microbiome, as it has been observed in arbuscular mycorrhiza and legumes (Takeda et al. 2007).
556 An obvious starting point to explore this hypothesis would be to confirm the predicted gene-
557 metabolite relationship proposed here, by sequencing the genomes of nostoginin-producing
558 organisms (Ploutno et al. 2002). If a siderophore, BGC12 could be mediating bacterial
559 community interactions (Traxler et al. 2015) and/or interactions with the host, adding to the
560 growing notion that symbiotic relations occur under heavy influence of metabolic exchange.
561 In sum, our study demonstrates the broad taxonomic diversity of the coralloid root,
562 which combined with monophyletic Nostocales large genomes with at least nine unique BGCs,
563 could be a sign of the constraints imposed by the transition from a facultative symbiotic lifestyle
564 into a specialized organ. As data is gathered from more genomes of bacterial cycad symbionts
565 isolated from different micro-niches of the rhizosphere-coralloid root ecological system, it will
566 be possible to test for co-evolutionary relationships, including horizontal gene transfer between
567 bacteria and the eukaryote host, and other patterns indicative of close evolutionary histories.
568 These patterns will inform the evolutionary and ecological nature of cycad-bacterial interactions,
569 which in turn could direct detailed functional characterization of these plant specialized bacterial
570 communities.

571 **Table 1.** Taxonomic composition of endophytes cultivated 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***

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

573

574

575 **Legends to Figures**

576

577 **Figure 1. Diagram of experimental strategy to capture bacterial diversity.** Coralloid roots
578 from cycads growing naturally in deciduous tropical forests were sampled. Endophytes from the
579 macerated root were isolated, following two strategies: directly from the sample (*t0*) and after
580 enrichment of sub-communities co-cultures in BG11 (*t1*). In both cases, cultivable bacteria were
581 obtained using an array of six different media. Co-cultures were characterized using shotgun
582 metagenomics, and the resulting data was used to select representative genomes from the
583 endophyte culture collection that we mined for functional and evolutionary information.

584

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

590

591 **Figure 3. Network of taxa co-occurrence from different coralloid root samples.** The most
592 abundant genera present in the four metagenomes representative of dry (JP) and humid (RF) are
593 represented by circles. Circle diameters are scaled in accordance with the number of reads
594 associated to each genus. The lines connecting the circles represent shared taxa, orange lines
595 correspond to dry environments and blue to wet environments.

596

Figure 4. A. Phylogenomic tree of the Phylum Cyanobacteria. Tree reconstructed with 198 conserved proteins to place 106C, T09 and RF31Ymg. Branches are colored according to the genera originally assigned in GenBank. **B. Zoom into the Nostocales**. Habitats for each species are indicated with colored bullets. Bars show a relatively homogeneous (G+C)-content among cyanobacteria, and a trend for larger genomes in plant symbionts including genomes obtained in this study.

Figure 5. *Dioon*-specific BGCs predicted from cyanobacterial genomes. 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, except for lantipeptide BGCs 1, 9, 10 and 17, whose products are shown as supplementary text S1.

Data deposition

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. RF31YmG

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

620 Supplementary files:

621 Cibrian_SupplementaryTables.xls, divided in sheets

622 Table S1. Organisms and proteins in the cyanobacterial core genome

623 Table S2. Genomes used to obtain the core proteome

624 Table S3. List of 470 isolated bacteria with 16S rRNA

625 Table S4. Biodiversity indices of 16S rRNA and OTUs

626 Table S5. Statistics and details of metagenomes sequenced

627 Table S6. Distribution and abundance of OTUs from the metagenomes

628 Table S7. Prediction of BGCs from the 106C isolate

629 Table S8. 106C-specific BGCs throughout Nostocales

630

631 Cibrian_SupplementaryText.txt

632 Text S1. Predicted lantipeptide from *Dioon* cyanobionts

633

634 Cibrian_SupplementaryFigures.pdf

635 Figure S1. OrthoCores diagram

636 Figure S2. Rarefaction of metagenomic OTUs defined by Kraken

637 Figure S3. Detailed 16S rRNA phylogenetic tree (two panels that can be zoomed in)

638 Figure S4. Syntenic comparisons among selected genomes

639

640 **Acknowledgements**

641 This work was supported by CONACyT #169701 and FON.INST./265/2016 (SA-FC2015-2-

642 901) to ACJ. CONACyT #179290 and #177568 to FBG. We acknowledge Rafael Rincón for his

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

646

647 **References**

648 Adnani, N, et al. (2017). Symbiosis-inspired approaches to antibiotic discovery. *Nat Prod Rep.*
649 34: 784-814.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

679 De Luca, P, et al. (1981). *Dioon meroale* (Zamiaceae), a new species from Mexico *Brittonia* 33:
680 179-185.

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

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

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

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

689 Gehringer, M, et al. (2010). Host selection of symbiotic cyanobacteria in 31 species of the
690 Australian cycad genus: *Macrozamia* (Zamiaceae). *Molecular Plant-Microbe Interactions* 23:
691 811-822.

692 Grobbelaar, N, et al. (1987). The identification of the coralloid root endophytes of the southern
693 African cycads and the ability of the isolates to fix dinitrogen. *South African Journal of Botany*
694 53: 111-118.

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

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

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

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

703 Huerta-Cepas, J, et al. (2010). ETE: a python Environment for Tree Exploration. *BMC*
704 *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, J. B., & 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.

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

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

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

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

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

748 Oksanen J, BF, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. (2015). Vegan: community
749 ecology package. R Packag. version *2*: 264.

750 Ow, M, et al. (1999). Reconstitution of a cycad-cyanobacterial association. *Symbiosis* 27: 125-
751 134.

752 Papaefthimiou, D, et al. (2008). Diversity and host specificity of *Azolla* cyanobionts. *J Phycol*
753 44: 60-70.

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

756 Posada, D, et al. (1998). ModelTest: testing the model of DNA substitution. *Bioinformatics.*
757 *1998* 14: 817-818.

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

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

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

764 Shannon, P, Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Admin, N.,
765 Schwikowski, B., Ideker, T (2003). Cytoscape: a software environment for integrated models of
766 biomolecular interaction networks. *Genome Research* 13: 2498–2504.

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

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

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

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

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

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

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

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

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

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

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

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

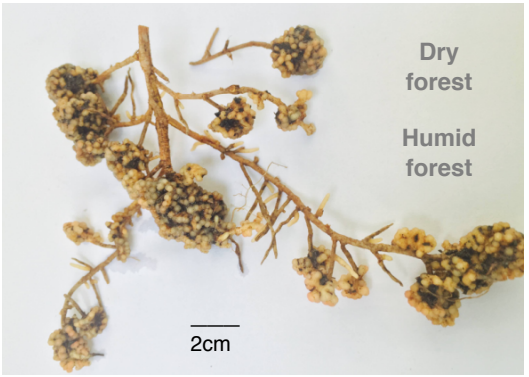
791 Zerbino, DR, et al. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn
792 graphs. *Genome Research* 18: 821-829.

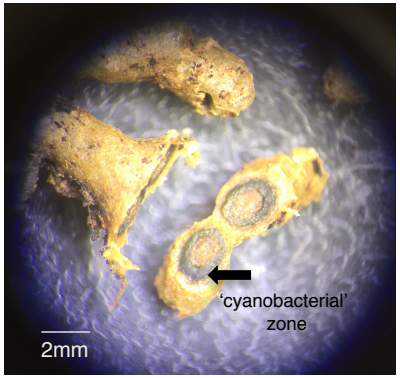
793 Zhaxybayeva, O, et al. (2006). Phylogenetic analyses of cyanobacterial genomes: Quantification
794 of horizontal gene transfer events. *Genome Research* 16: 1099–1108.

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

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

799 Zvyagintsev, D, et al. (2010). Morphological and physiological modifications of cyanobacteria in
800 experimental cyanobacterium-actinomycete associations. *Microbiology* 79: 314-320.
801



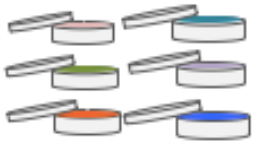


16S rRNA Sanger

Amplicon sequencing

t0

Strain isolation



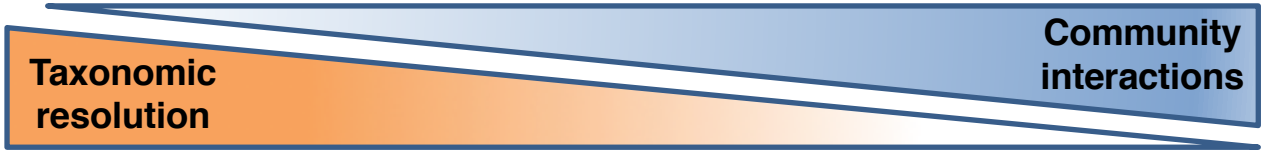
Six selective media

16S rRNA Sanger
Whole genome sequencing

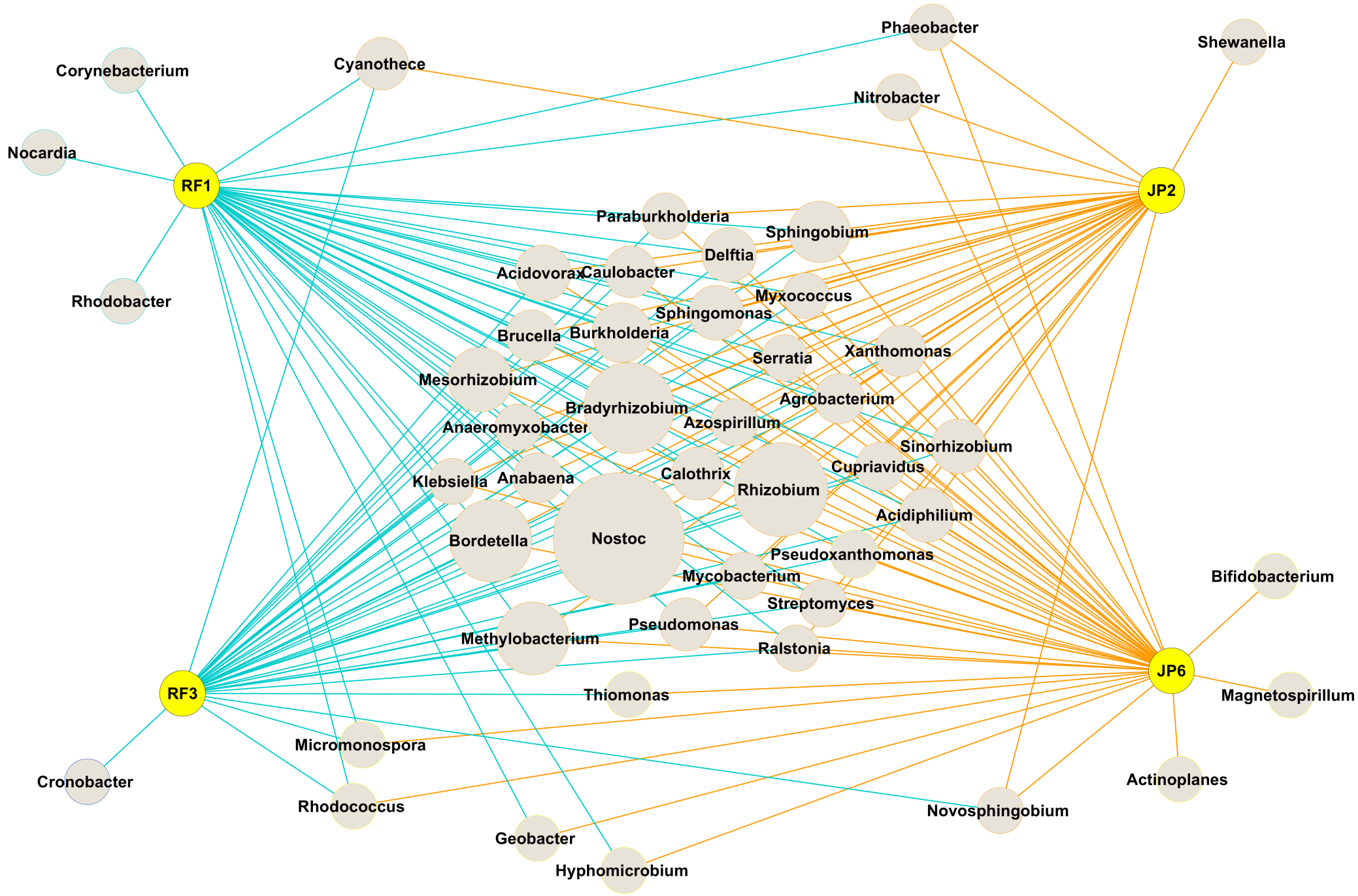
Sub-community co-cultures

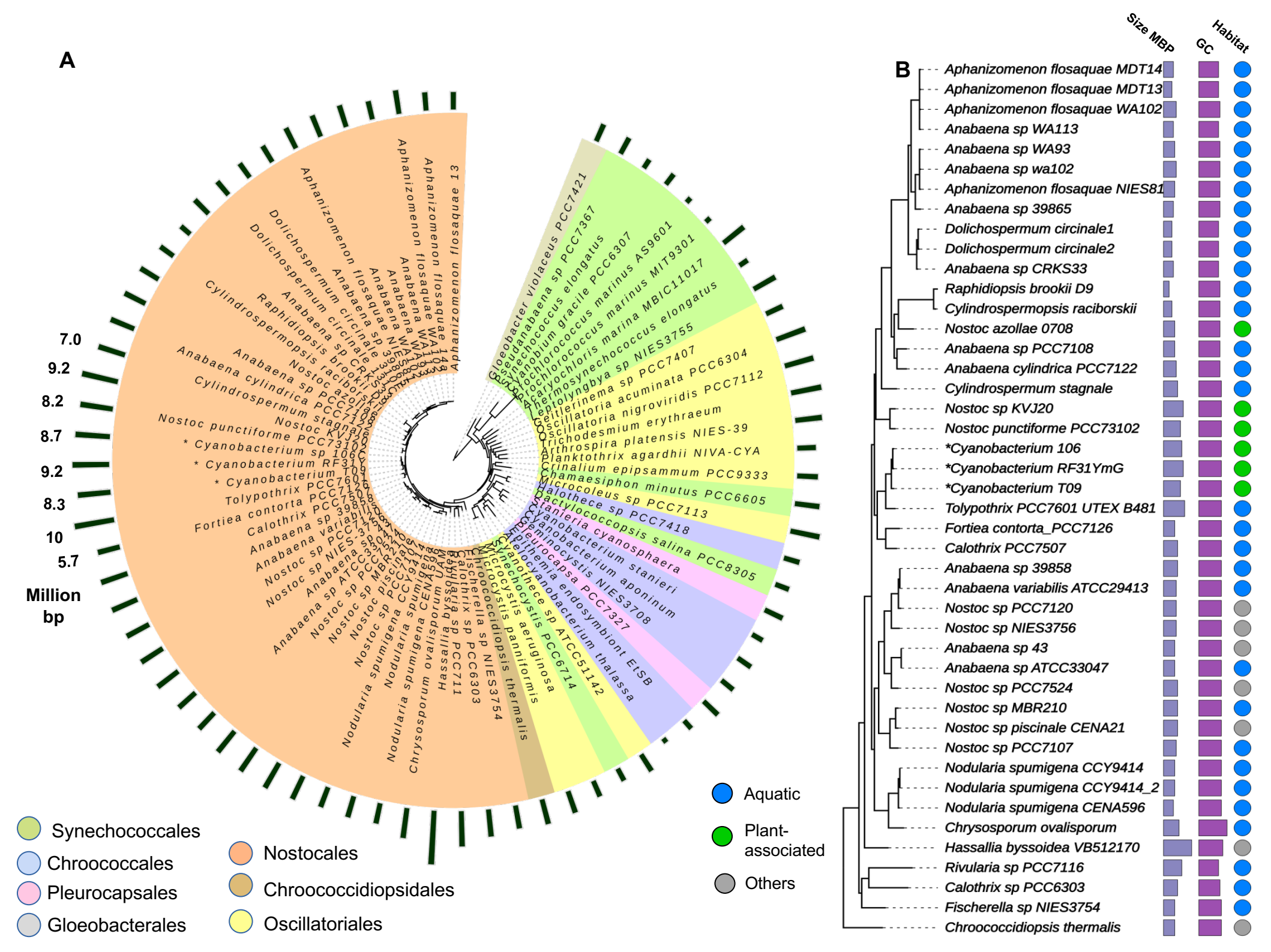
t1, t2..tn

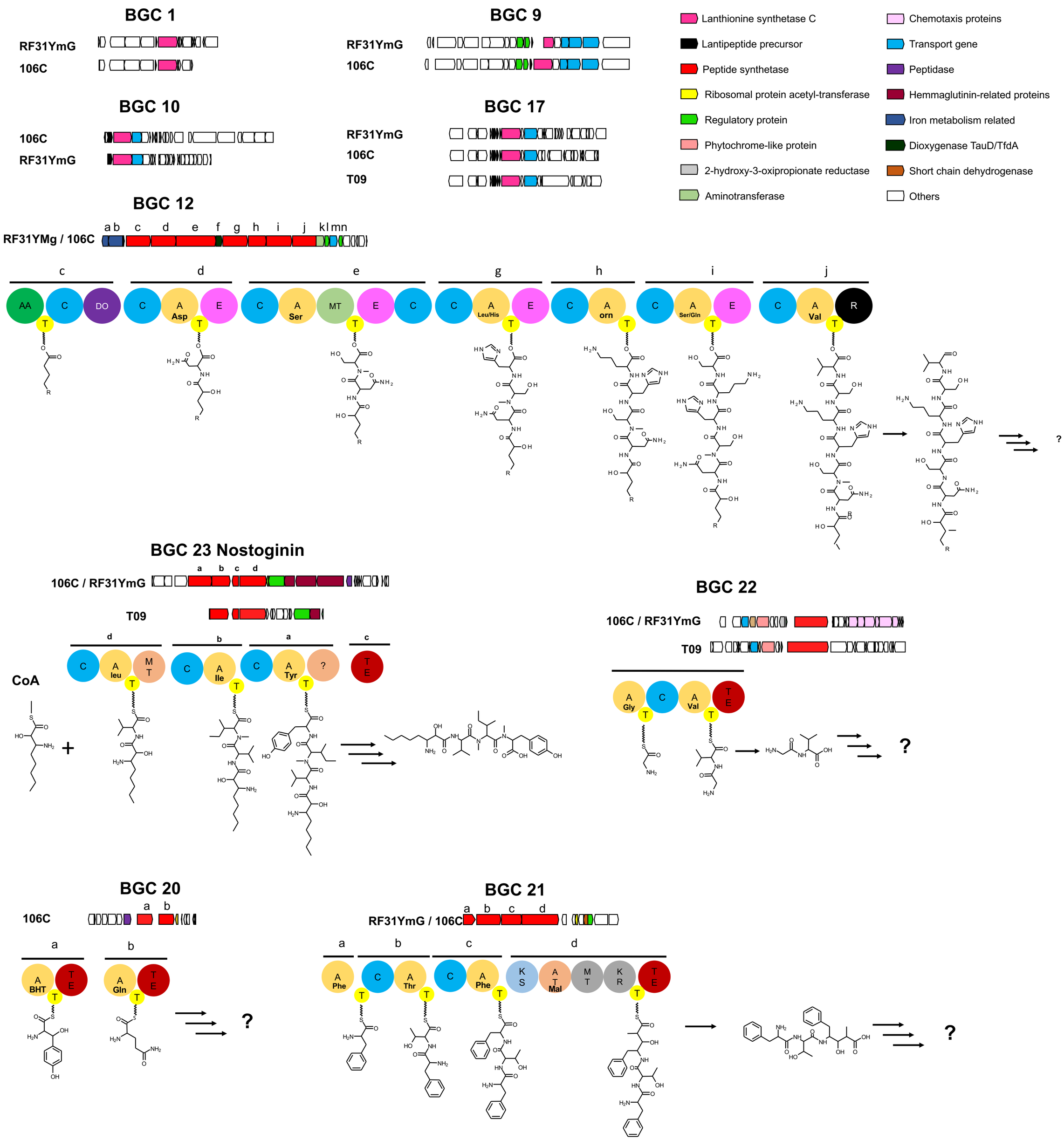
Shotgun metagenomic sequencing



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


A
B
Size MBP
GC
Habitat
7.0
9.2
8.2
8.7
9.2
8.3
10
5.7
Million bp
Aphanizomenon flosaquae 13
Aphanizomenon flosaquae 14a
Aphanizomenon flosaquae WA102
Anabaena WA113
Anabaena WA93
Anabaena 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
Synechococcales
Chroococcales
Pleurocapsales
Gloeobacterales
Nostocales
Chroococcidiopsidales
Oscillatoriales
Aquatic
Plant-associated
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
d
e
g
h
i
j
AA
C
DO
A
Asp
E
A
Ser
MT
A
Leu/His
A
orn
A
Ser/Gln
A
Val
R
T
?
BGC 23 Nostoginin
106C / RF31YmG
a b c d
T09
d
b
a
c
CoA
C
A
leu
M
T
C
A
Ile
C
A
Tyr
?
T
E
BGC 22
106C / RF31YmG
T09
A
Gly
C
A
Val
T
E
?
BGC 20
106C
a b
A
BHT
T
E
A
Gln
T
E
?
BGC 21
RF31YmG / 106C
a b c d
A
Phe
C
A
Thr
C
A
Phe
K
S
A
T
Mal
M
T
K
R
T
E
?