

1 **Title: Diversity and specificity of molecular functions in cyanobacterial**
2 **symbionts**

3 **Authors:** Ellen S. Cameron^{1,2}, Santiago Sanchez¹, Nick Goldman¹, Mark L. Blaxter²,
4 Robert D. Finn¹

5 ¹ European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-
6 EBI), Hinxton, Cambridge CB10 1SD, United Kingdom

7 ² Tree of Life, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton,
8 Cambridge, CB10 1SA, United Kingdom

9

10	Abstract	3
11	1. Introduction	4
12	2. Methods	7
13	<i>2.1 Cyanobacterial genomes, habitat annotation & quality control</i>	7
14	<i>2.2 Phylogenetic tree reconstruction</i>	7
15	<i>2.3 Genome annotation and KEGG completeness estimation</i>	8
16	<i>2.4 Biosynthetic gene cluster prediction and classification</i>	8
17	3. Results	10
18	<i>3.1 Enrichment of Molecular Functions and Biosynthetic Gene Clusters in Host-Associated Cyanobacterial Symbionts</i>	10
20	Figure 1: Phylogeny and distribution of host-associated lifestyles in the phylum, Cyanobacteria.	11
22	Figure 2: Host-associated enrichment of KEGG pathways and secondary metabolite production potential.	16
24	<i>3.2 Host-Associated Lifestyle Appears Non-Specific with Multiple Origins in the Nostocaceae</i>	17
26	Figure 3: Distribution of host-types in the order Nostocales and the origin of host associations in Nostocaceae	20
28	<i>3.3 Host-specific molecular specialization in Nostocaceae symbionts</i>	21
29	Figure 4: Distribution of significant KEGG functions and groups of biosynthetic gene clusters impacted by isolation source in Nostocaceae genera which include host-associated cyanobacterial symbionts.	23
32	4. Discussion	24
33	5. References	29
34	Competing Interests	36
35	Data Availability	36
36	Acknowledgements	36
37		

38 **Abstract**

39 Cyanobacteria are globally occurring photosynthetic bacteria notable for their
40 contribution to primary production and their production of toxins which have
41 detrimental impacts on ecosystems. Beyond this, cyanobacteria can form mutualistic
42 symbiotic relationships with a diverse set of eukaryotes, ranging from land plants to
43 fungi. Nevertheless, not all cyanobacteria are found in symbiotic associations
44 suggesting symbiotic cyanobacteria have evolved specializations that facilitate host-
45 interactions. Photosynthetic capabilities, nitrogen fixation, and the production of
46 complex biochemicals are key functions provided by host-associated cyanobacterial
47 symbionts. To explore if additional specializations are associated with such lifestyles
48 in cyanobacteria, we have conducted comparative phylogenomics of molecular
49 functions and of biosynthetic gene clusters (BGCs) in 977 cyanobacterial genomes.
50 Cyanobacteria with host-associated and symbiotic lifestyles were concentrated in the
51 family Nostocaceae, where eight monophyletic clades correspond to specific host
52 taxa. In agreement with previous studies, symbionts are likely to provide fixed nitrogen
53 to their eukaryotic partners. Additionally, our analyses identified chitin metabolising
54 pathways in cyanobacteria associated with specific host groups, while obligate
55 symbionts had fewer BGCs. The conservation of molecular functions and BGCs
56 between closely related symbiotic and free-living cyanobacteria suggests that there is
57 the potential for additional cyanobacteria to form symbiotic relationships than is
58 currently known.

59 **Keywords:** mutualism, facultative symbioses, host-association, biosynthetic gene
60 clusters, phylogenomics

61 1. Introduction

62 Cyanobacteria, a group of photosynthetic bacteria, have a long evolutionary
63 history with fossil evidence dating back up to 1.9 billion years ago¹. These organisms
64 are found globally in diverse habitats, from aquatic to terrestrial landscapes and from
65 polar to tropical climates¹⁻³. Cyanobacteria have evolved adaptations for survival
66 under numerous types of stresses including desiccation, extreme temperatures,
67 salinity, UV radiation and pathogenic infections⁴. Additionally, they can threaten
68 ecosystems and human health through their production of potent toxins when
69 blooms contaminate water sources⁵. As such, much of cyanobacterial research has
70 focused on public health impacts of freshwater and marine strains⁶. However, the
71 impact of cyanobacteria on ecosystem health extends beyond degrading water
72 quality, as many of the members in this taxonomic group have been found to be
73 critical partners in mutualistic symbiotic associations with a diverse range of
74 eukaryotic hosts.

75 The keystone example of cyanobacterial symbioses is that of the endosymbiotic
76 event that occurred some 2.1 billion years ago and led to the development of
77 chloroplasts and photosynthetic eukaryotes⁷. Beyond the endosymbiont origin of
78 chloroplasts, cyanobacteria are also found in symbiotic associations with diverse
79 hosts such as protists, metazoans, fungi, macroalgae and land plants in both
80 terrestrial and aquatic environments⁷⁻⁹. These symbionts provide hosts with
81 beneficial services including photosynthetic products¹⁰ and fixed nitrogen^{11,12}. The
82 mode of host association is also variable, including epiphytic growth (e.g. on
83 feathermoss), intra-organismal location in specialised symbiotic structures and
84 intracellular incorporation^{9,11}. These associations can be ancient with examples of
85 cyanobacterial symbionts found in a fossilised lichen from 400 million years ago⁷.
86 Such long associations raise the potential for coevolution between the eukaryotic
87 host and specialised cyanobacterial partners⁹, selecting for symbiotic
88 competence^{13,14}.

89 Host-symbiont interactions require pathways for communicating and detecting
90 signals¹⁴ which may involve secondary metabolites. Secondary metabolites,
91 compounds that are not essential for primary growth and reproduction¹⁵, are
92 produced by co-localized genes organized as biosynthetic gene clusters (BGC)¹⁶.

93 These compounds are often specialized for species interaction and survival in
94 stressful environments, and can include bioactive compounds with antibacterial,
95 antifungal and cytotoxic properties^{4,15,16}. Secondary metabolites have previously
96 been shown to impact symbiotic associations such as diatoms producing compounds
97 to promote growth and attachment of beneficial bacteria¹⁷, or coral microbiomes
98 producing a high diversity of antimicrobial products¹⁸. However, secondary
99 metabolites are often produced for a specific physiological or ecological reasons and
100 are often taxon specific¹⁹, with this specificity potentially being a mechanism for
101 symbiont communication to their potential host²⁰.

102 Even amongst microbes known for their production of diverse secondary
103 metabolites, cyanobacteria alone are known to produce over 1,100 unique
104 secondary metabolites and their genomes frequently contain a high number of
105 BGCs^{21,22}. The majority of cyanobacterial genomes contain polyketide synthase and
106 nonribosomal peptide synthetase pathways that account for up to 5% of their total
107 genome sizes²³. The compounds that cyanobacteria produce span diverse roles
108 ranging from UV protection (mycosporines and scytonemin) to grazing deterrents
109 and nutrient scavenging⁶ which may provide additional competitive advantages to
110 hosts²⁴. The compounds may also mediate inter-partner communication in
111 symbioses. For example, the production of nostopeptolide in the cyanobacterium
112 genus, *Nostoc*, is associated with repression of formation of infectious differentiated
113 cells and is down-regulated in the presence of plant hosts^{25,26}. While genome mining
114 approaches have identified many cyanobacterial biosynthetic gene clusters of
115 unknown function^{27,28}, the potential for symbiosis-specific secondary metabolites and
116 their distribution among lineages of cyanobacteria has not been fully explored.

117 Cross-talk between cyanobacteria and their host-species has been previously
118 reported, ranging from the upregulation of transcription of ammonium and nitrate
119 transporters²⁹ to influencing cell differentiation in the life cycle of *Nostoc*¹¹. However,
120 varying reports of host-specificity and phylogenetic clades of symbiont
121 cyanobacteria^{9,11,30–32} requires a phylum-wide study to explore the origins of host-
122 association in this ancient lineage. Uniquely, *Nostoc* has shown broad symbiotic
123 competence with different eukaryotic hosts, yet there still remains questions on
124 molecular drivers of these associations due to the potential of non-host specific
125 responses as isolates from cycads have previously been shown to also enter into

126 symbiotic associations with mosses, fungi and angiosperms (*Gunnera*)¹³. Previous
127 research has identified niche-specific BGCs that have been connected to individual
128 host-specific associations in cyanobacteria³³ suggesting the presence of specialized
129 secondary metabolites associated with cyanobacterial symbionts. However, a large-
130 scale analysis of all available cyanobacterial genomes within the context of symbiotic
131 associations has not yet been conducted. In this work, we utilize comparative
132 phylogenomic approaches to identify trends in distribution of (i) molecular functions
133 and (ii) biosynthetic gene clusters which may mediate host-symbiont interactions in
134 this phylum.

135 **2. Methods**

136 **2.1 Cyanobacterial genomes, habitat annotation & quality control**

137 Assembled genome sequence data for 1078 species belonging to the phylum
138 Cyanobacteria were downloaded from NCBI RefSeq in January 2023 (Table S1). An
139 additional 27 metagenomic assembled genomes (MAGs) taxonomically classified as
140 cyanobacteria from lichen sources³⁴ were included to provide additional examples of
141 host-associated symbionts for a total of 1105 cyanobacterial genomes.

142 Wherever possible the sampled cyanobacteria were assigned to their source
143 habitat(s) based on available sample metadata, associated publication(s) or
144 metadata describing the original isolation reported by culture collections. These
145 habitat assignments include aquatic (e.g., freshwater, marine and man-made aquatic
146 sources) and terrestrial (e.g. soils), as well as host-associated environments. Host
147 associations include vascular and non-vascular plants, protists, fungi, macroalgae,
148 and marine mammals (epidermal mats). Individual host species were grouped into
149 broad taxonomic categories including bryophytes, cycads, fruit trees, diatom
150 endosymbionts, and lichens. Water fern (*Azolla*) cyanobacterial symbionts were
151 placed in their own category. These habitat annotations were also used for grouping
152 the cyanobacteria into two broader lifestyle classifications: free-living and host-
153 associated. Cyanobacterial genomes of which no specific source habitat could be
154 discovered were excluded, leaving 1026 cyanobacterial genomes for comparative
155 analyses.

156 Quality control filtering was performed using CheckM³⁵ (Version 1.1.3) and 977
157 high-quality (>90% complete; <5% contamination; Table S2) cyanobacterial
158 genomes were retained for phylogenetic tree reconstruction and downstream
159 analysis. Representatives of Melainabacteria (n=37), a basal non-photosynthetic
160 lineage of cyanobacteria, were included as an outgroup.

161 **2.2 Phylogenetic tree reconstruction**

162 Taxonomic classification of genomes and generation of marker gene alignments
163 was conducted using GTDBtk³⁶(v. 2.3.0; Table S3). Phylogenetic trees were
164 constructed for the final high-quality set of cyanobacterial genomes using IQ-
165 TREE³⁷(v. 2.2.0). The analysis used the LG+R10 model as identified in the IQ tree

166 model finder based on the Bayesian Information Criterion (BIC). A family-level
167 phylogenetic tree for the family Nostocaceae (n = 300), rooted with representatives
168 of the order Elainellales, was constructed using IQ-TREE and the LG+F+R7 model
169 determined by BIC. Phylogenetic trees were visualised using iTOL³⁸(v.5). For phylum
170 and family level trees, non-parametric bootstraps (n = 1000) were conducted with IQ-
171 TREE to assess the robustness of phylogenetic inferences.

172 **2.3 Genome annotation and KEGG completeness estimation**

173 Cyanobacterial genomes were annotated with Prokka³⁹(v.1.14.6) and the
174 resulting gene predictions were functionally annotated with KofamScan(v.1.3.0) to
175 derive Kyoto Encyclopaedia of Genes and Genomes (KEGG) module annotations.
176 KofamScan predictions were used with KEGG-Decoder⁴⁰(v. 1.3) to generate a table
177 representing molecular function completeness across samples (Table S4). KEGG
178 functions were classified as being present using two thresholds, either >98%
179 complete for a more stringent analysis of distribution and complete function, or >50%
180 complete for lower stringency examination for the potential presence of molecular
181 functions, herein referred to as indicative functions (Figure S1). Presence/absence
182 matrices generated for KEGG functions were used in a phylogenetic logistic
183 regression⁴¹ to identify enrichment of molecular functions based on lifestyle
184 classification at the phylum level (Table S5, S6) and enrichment of molecular
185 functions in individual isolation sources in the family Nostocaceae (Table S7, S8).
186 Phylogenetic logistic regressions were conducted using the *phyloglm* function in the
187 R package *phylolm*⁴², using the penalised likelihood with Firth's correction and 100
188 bootstraps. Responses of lifestyle classification and isolation sources were defined
189 as significant if the p-value was less than 0.05.

190 **2.4 Biosynthetic gene cluster prediction and classification**

191 BGCs were predicted on cyanobacterial genomes using SanntiS⁴³ (v. 0.9.1) due
192 to high performance on both isolate genomes and MAGs, thus providing consistent
193 annotations across all genome types. The predictions were subsequently filtered to
194 remove those occurring at the edges of contigs and those which were less than 3000
195 bp in length, reflective of the minimum length of BGCs observed in the MIBiG
196 database¹⁶. BGCs were initially classified by SanntiS into standard classes such as
197 ribosomally synthesised and post-translationally modified peptides (RiPPs),

198 terpenes, nonribosomal peptides, polyketides, alkaloids, saccharides, and hybrid
199 classes which represent BGCs that cover multiple biochemical classes (Table S9).
200 To detect enrichment of total and specific BGC classes in host-associated
201 symbionts, phylogenetic linear regression was conducted at the phylum level (Table
202 S10) and in the Family Nostocaceae (Table S11). This was performed with the
203 *phylolm* function using 100 bootstraps and a lambda model for covariance.

204 To expand upon the basic BGC classifications provided by SanntiS and identify
205 diversity in potential products, predicted BGCs in cyanobacteria were clustered with
206 a large, reference set of biosynthetic gene clusters (the ‘reference BGC collection
207 termed RefBGC hereafter). RefBGC includes BGC predictions from running SanntiS
208 on MGnify⁴⁴ and RefSeq genomes⁴⁵, as well as the BGCs found in MiBIG¹⁶, and
209 subsequently refined to only include complete predictions. This clustreing enabled
210 the assignment of BGCs to more specific groups based on functional domain
211 composition, utilizing the Louvain community detection method⁴⁷ and the Sørensen-
212 Dice similarity coefficient⁴⁸. To refine the SanntiS BGC classification assigned to
213 each group antiSmash⁴⁶(v.7.0.0) predictions were also generated for RefSeq and
214 used to provide more specific natural product annotations, thereby combining the
215 breadth offered by SanntiS and the accurate BGC product assignments provided by
216 antiSMASH. Groups of BGCs containing antiSmash predictions were retained as the
217 final set of BGCs (Table S12). The habitat source of each BGC group was use in
218 phylogenetic logistic regression to identify enrichment of specialized biosynthetic
219 gene clusters in cyanobacteria with different lifestyles (Table S13, Table S14). This
220 was performed with the *phyloglm* function maximizing the penalized likelihood with
221 Firth’s correction across 100 bootstraps. Groups found to be significantly enriched at
222 the phylum level were used to assess phylogenetic signal in the family Nostocaceae
223 using the D-statistic⁴⁹ with the *phylo.d* function in the R package *caper*⁵⁰(v.1.0.2) of
224 lifestyle classification and isolation sources were defined as significant if p-value was
225 less than 0.05.

226 3. Results

227 3.1 Enrichment of Molecular Functions and Biosynthetic Gene Clusters in 228 Host-Associated Cyanobacterial Symbionts

229 Using the taxonomic classifications based on GTDB the cyanobacterial
230 genomes were assigned to 18 taxonomic orders and 42 families, which were
231 monophyletic based on the GTDBtk phylogeny thus facilitating rigorous interpretation
232 of evolutionary relatedness of these organisms. Of these, Cyanobacteriales (n = 576)
233 and PCC-6307 (representative of *Cyanobium gracile*; n = 261) comprised over 85%
234 of available genome assemblies (Figure 1A). Habitat sources were highly skewed,
235 with aquatic environments (n = 753) representing >75% of environmental sources for
236 all genome assemblies. Notably, only 6% (n = 62) of assessed cyanobacterial
237 genomes were isolated from host-associated environments including non-vascular
238 and vascular plants, protists, macroalgae, metazoan epidermal mats and fungi.
239 Within this, Cyanobacteriales accounted for 93% [5.9% of host-associations in all
240 assessed cyanobacterial genomes; n = 58] of all host-associated cyanobacterial
241 symbiont genomes including representatives from all detected habitat source
242 classifications (Figure 1B). NCBI taxonomy was also considered, however due to
243 challenges with nested, non-monophyletic groupings based on current taxonomic
244 nomenclature, comparisons based on 'taxonomic identity' were not possible.
245 Nevertheless, similar trends were shown with NCBI taxonomy with a high proportion
246 of genomes arising from the orders Synechococcales (n = 428) and Nostocales (n =
247 300) comprising nearly 75% of available reference genome assemblies with host-
248 associations concentrated in the Nostocales (Figure S2).

249

A

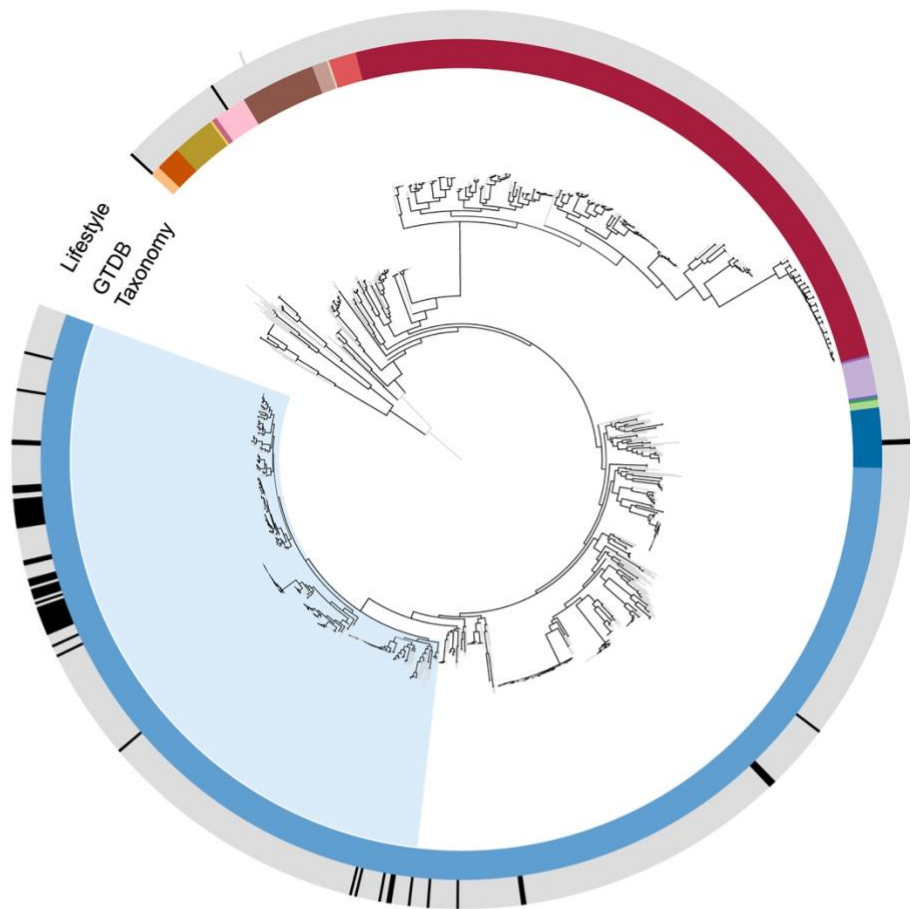
Tree scale: 1

GTDB Phylum Classification

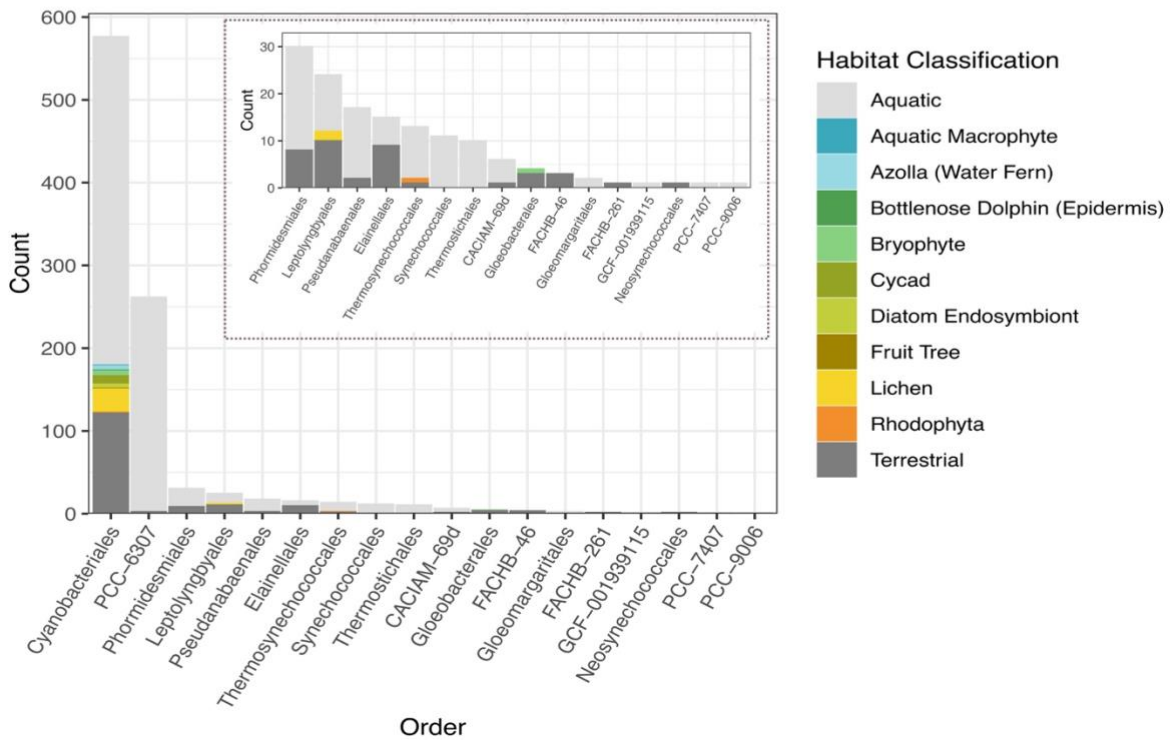
- Gloeobacterales
- Thermotrichales
- Pseudanabaenales
- FACHB-261
- Gloeomargaritales
- Thermosynechococcales
- Phormidesmiales
- CACIAM-69d
- PCC-9006
- Synechococcales
- PCC-6307
- PCC-7407
- Elainellales
- Neosynechococcales
- GCF-001939115
- FACHB-46
- Leptolyngbyales
- Cyanobacteriales

Lifestyle Classification

- Free-Living
- Host-Associated



B



250

251 *Figure 1: Phylogeny and distribution of host-associated lifestyles in the phylum,*
252 *Cyanobacteria.*

253 (A) Phylogeny generated using concatenated marker genes of genome sequences
254 of strains from phylum Cyanobacteria, rooted with representatives of the sister
255 group, Melainabacteria, with 1000 bootstraps. Branches with high bootstrap support
256 (>80%) are shown with black. The outer annotation track depicts the lifestyle
257 classification to highlight host-associated cyanobacterial symbionts. The inner
258 annotation track depicts the classified taxonomic order assigned by GTDB.
259 Nostocaceae, a family containing the majority of host associations, are shaded in
260 light blue. (B) Frequency counts distributed across taxonomic orders for habitat
261 classifications highlighting the different host sources including vascular plants (water
262 fern, cycad, fruit trees, aquatic macrophytes), non-vascular plants (bryophytes),
263 protists (e.g. diatoms), macroalgae (Rhodophyta), fungi, and epidermal mats of
264 aquatic mammals such as dolphins.

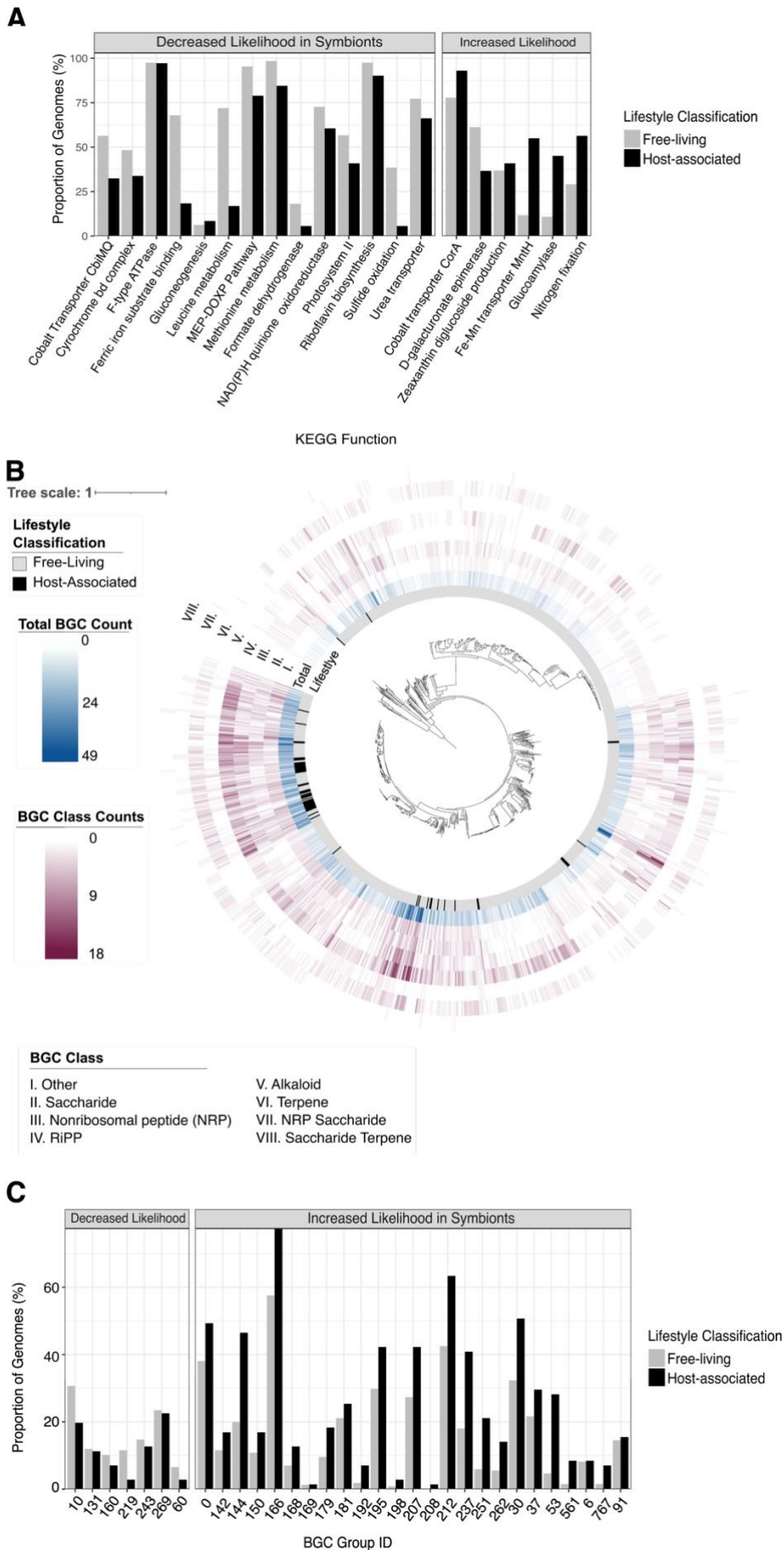
265

266 KEGG functional annotations were analysed to identify molecular functions
267 enriched in symbiont genomes by exploring the distribution of complete KEGG
268 functions. In total, 77 complete KEGG functions were variably present across the
269 phylum, of which 20 were significantly associated with lifestyle classification (Figure
270 2A; Figure S3). Host-associated lifestyles were found to have a significantly higher
271 level of occurrence of functions including those of glucogenesis ($p=0.042$; Est. 0.85),
272 Fe-Mn transporter ($p=5.53e-07$; Est. 1.45), glucoamylase ($p=4.86e-04$; Est. 1.31),
273 zeaxanthin diglucoside production ($p=6.36e-04$; Est. 1.01), cobalt-magnesium
274 transporters ($p=3.46e-02$; Est. 0.54) and nitrogen fixation ($p=4.77e-02$; Est. 0.59).
275 While statistically non-significant, chemotaxis ($p=0.056$; Est. 0.41) was also found to
276 have a higher likelihood of occurrence in host-associated cyanobacteria. Certain
277 complete functions were also found to have a significantly lower occurrence in host-
278 associated symbionts including photosystem II ($p=7.97e-14$; Est. -2.74), the MEP-
279 DOXP pathway ($p=1.87e-11$; Est. -2.37), methionine ($p=1.24e-09$; Est. 3.07) and
280 leucine metabolism ($p=2.36e-03$; Est. -0.83), F-type ATPase ($p=5.54e-06$; Est. -
281 2.03), NAD(P) H-quinone oxidoreductase ($p=1.25e-05$; Est. -1.18), riboflavin
282 biosynthesis ($p=3.25e-05$; Est. -1.59), sulfide oxidation ($p=3.58e-04$; Est. -1.17), urea
283 transporters ($p=1.68e-03$; Est. -0.81), cytochrome bd complex ($p=2.87e-03$; Est.
284 0.87), cobalt transport proteins (CbiMQ; $p=5.42e-03$; Est. 0.79), D-galacturonate
285 epimerase ($p=6.21e-03$; Est. 0.71), formate dehydrogenase ($p=5.36e-03$; Est. -1.11),
286 cytochrome bd complex ($p=2.87e-03$; Est. -0.8375) and iron transport system binding
287 proteins ($p=2.56e-03$; Est. -0.85). Assessment of indicative functions also revealed a
288 significantly higher likelihood of occurrences for Type I Secretion systems ($p=0.024$;
289 Est. 0.53) and SecSRP secretion pathways ($p=0.01$; Est. 0.82).

290 To further explore specializations associated with host-associated
291 relationships in cyanobacteria, 8,815 BGCs were identified across 98% ($n=961$) of all
292 cyanobacterial genome assemblies. In total, 21 classes of biosynthetic gene clusters
293 were identified including hybrid classes which span biochemical properties of
294 multiple classifications. Lifestyle classification was found to significantly associate
295 with the number of detected BGCs. Host-associated cyanobacteria were found to
296 have a significantly lower numbers of BGCs in total ($p = 1.02e-06$; Est. -4.13) (Figure
297 2B), and this trend was paralleled at the level of individual BGC class. Host-
298 associated cyanobacterial symbionts were found to have significantly lower count of

299 individual biosynthetic gene cluster classes including nonribosomal peptides (NRP; p
300 = 0.016; Est. -0.52), RiPPs (p = 1.44e-04; Est. -1.21), alkaloids (p=1.01e-03; Est. -
301 0.15), terpenes (p=2.84e-03; Est. -0.43), saccharides (p=8.7e-03; Est. -0.47),
302 saccharide terpenes (p=0.019; Est.-0.038), NRP saccharides (p =0.042; Est. -0.082),
303 and other (p = 0.0012.87e-05; Est. -0.75639), a class of BGC that does not fit into
304 properties of otherwise described secondary metabolites.

305 The 8815 biosynthetic gene clusters identified were classified into 124 unique
306 groups representative of BGCs, which are likely to produce similar secondary
307 metabolites based on similarity of the protein domain annotations. Although host-
308 associated symbionts were found to have a significantly lower count of BGCs and
309 classes of BGCs as a whole, individual BGC groups were found to be positively
310 associated with cyanobacterial symbionts. Overall, 61 groups were found to be
311 present in both free-living and host-associated cyanobacteria, 61 groups were found
312 only in free-living cyanobacteria, and only 2 groups were found exclusively in host-
313 associated symbionts corresponding to a terpene in a cycad symbiont and 'other'
314 classification in aquatic macrophytes symbionts. Of the 61 groups found in both free-
315 living and host-associated cyanobacteria, 25 were found to have a significantly
316 higher prevalence in host-associated cyanobacteria (Figure 2C; Figure S4), while
317 only 7 showed a significantly decreased prevalence in host-associated symbionts
318 (Table S10).



320 *Figure 2: Host-associated enrichment of KEGG pathways and secondary metabolite*
321 *production potential.*

322 (A) Proportion of genomes of each lifestyle type containing KEGG pathways shown
323 to be significantly impacted by life-style classification including key functions
324 corresponding to beneficial ecosystem services including nitrogen fixation. (B)
325 Distribution of counts of total detected biosynthetic gene clusters and classes of
326 biosynthetic gene clusters shown to be significantly impacted by lifestyle
327 classification across the phylum Cyanobacteria (C) Proportion of genomes for each
328 lifestyle type with unique groups of biosynthetic gene clusters that are significantly
329 impacted by lifestyle classification.

330

331 **s3.2 Host-Associated Lifestyle Appears Non-Specific with Multiple Origins in** 332 **the Nostocaceae**

333 Cyanobacteriales-classified cyanobacteria were recovered as a well-
334 supported monophyly (Figure 1A) and contained the majority of the symbionts
335 analysed. Within the Cyanobacteriales, the host-associated lifestyle was found to be
336 concentrated in the family Nostocaceae (Figure 3A). Phylogenetic reconstruction
337 based on marker genes from publicly available high-quality cyanobacterial genomes
338 belonging to the family Nostocaceae revealed a family-wide distribution of host-
339 associated growth forms (Figure 3B). Eight monophyletic clades corresponding to a
340 unique host category (Table S15) ranging in levels of host specificity. Denoted
341 clades I–VIII, they derive from: diatom endosymbionts; Peltigeraceae lichens
342 *Solorina crocea* and *Peltigera malacea*; the lichen *Peltigera membranacea*; *Azolla*
343 ferns; an unspecified lichen thallus cyanobiont culture ATCC 53789); the lichen
344 *Peltigera*; Peltigeraceae lichens *Collema furfuraceum*, *Leptogium*
345 *austroamericanum*, *Lobaria pulmonaria*, *Peltigera membranacea*, *Peltigera aphthosa*
346 and *Peltigera malacea*; and *Dioon* cycads, respectively.

347 Ten cyanobacterial genomes were sourced from cycad symbioses but only
348 three of these were found to form a monophyletic clade. *Aulosira*, previously
349 classified as *Nostoc*, comprised monophyletic clade VIII. These symbionts were all
350 from a *Dioon* host supporting previous reports of monophyletic origin of endophytic
351 cyanobacteria with this host species (Gutierrez-Garcia et al., 2019). Cyanobacteria
352 from other cycad hosts (*Cycas revoluta* (n = 3), *Macrozamia* (n = 1), *Zamia*
353 *pseudoparasitica* (n = 1), *Encphalartos horridus* (n = 1), and *Euterpe edulis* (n = 1))
354 were distributed across the phylogeny. The genomes sourced from *Cycas revoluta*
355 did not form a monophyletic clade and were distributed across the Nostocaceae tree.
356 The cyanobacterium from the Arecales palm, *Euterpe edulis*, was found in a clade
357 with the cyanobacterium from *Garcinia macrophylla*, a dicot (Malpighiales) fruit tree.

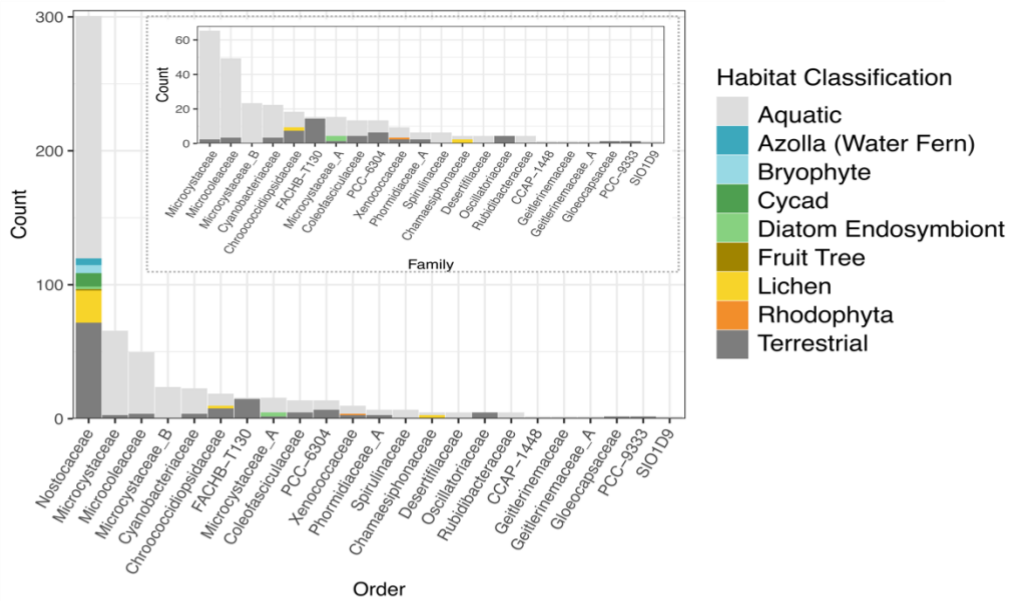
358 Clade IV contained 3 of the 5 analyzed *Azolla* cyanobionts. Notably, the
359 cyanobiont isolated from an epiphytic growth form on *Azolla* was not found with other
360 true *Azolla* cyanobionts.

361 Five of the monophyletic clades, denoted II, III, IV, V and VII, contained 66%
362 (n = 16) of the analysed lichen cyanobionts, and their hosts were all Peltigeraceae

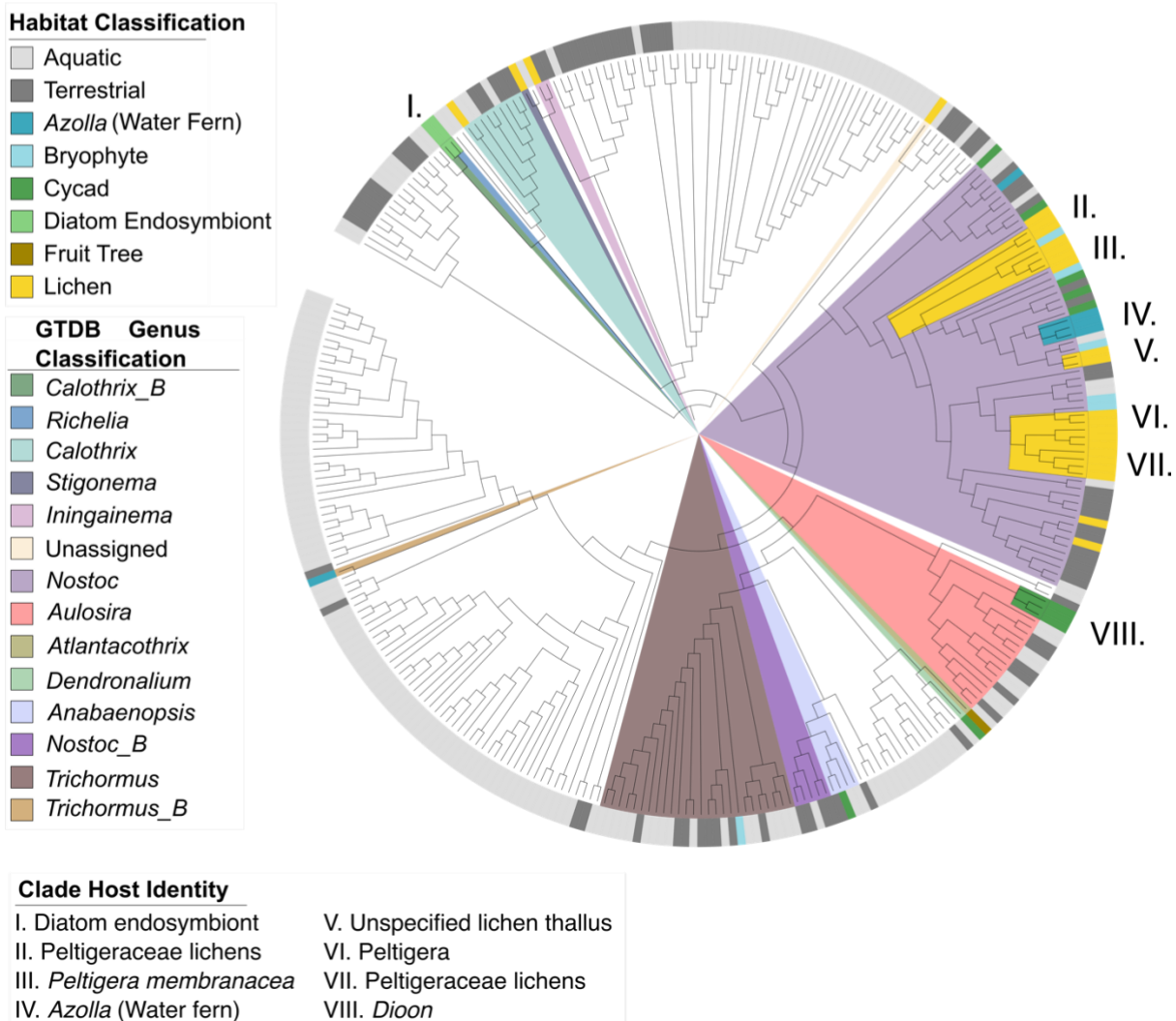
363 fungi. Lichen cyanobionts most distant to the main lichen clades arose from lichens
364 of different family lineages including *Coccocarpia palmicola* (Coccocarpiaceae) and
365 *Placynthium petersii* (Placynthiaceae) in more basal origins of the Nostocaceae.
366 While all lichens observed in this analysis were of the order Peltigerales, the
367 mycobiont from these lichens are in a different fungal family compared with those in
368 the other analysed cyanolichens (Peltigeraceae), suggesting the potential for
369 genomic diversity in cyanobionts depending on host identity.

370 Bryophyte cyanobionts did not form host-specific clades, but instead were often
371 found in clades containing lichen cyanobionts or terrestrial isolates. Bryophyte
372 cyanobionts were limited to three host species: *Blasia pusilla* (n=3), *Phaeoceros*
373 (n=1), and *Leiosporoceros dussi* (n=3). The multiple isolates from *Blasia pusilla* and
374 *Leiosporoceros dussi* were distributed across the tree but commonly observed in
375 clades with lichen cyanobionts.

A



B



377 *Figure 3: Distribution of host-types in the order Nostocales and the origin of host*
378 *associations in Nostocaceae*

379 (A) Frequency counts distributed across taxonomic families in the order Nostocales
380 which includes the majority of host-associated cyanobacterial symbiont genomes
381 spanning a high diversity of eukaryotic hosts in the family, Nostocaceae. Families
382 with low frequency counts are displayed as an inset panel. (B) Cladogram of
383 Nostocaceae generated from an alignment of marker genes rooted with the outgroup
384 of Elainellales (n = 15) to explore the origin of host-specific association. Genera with
385 host-associations are highlighted, as well as a non-host associated genus of *Nostoc*
386 (*Nostoc_B*). Colour block shading on branches represent eight monophyletic clades
387 containing symbionts arising from single host classifications.

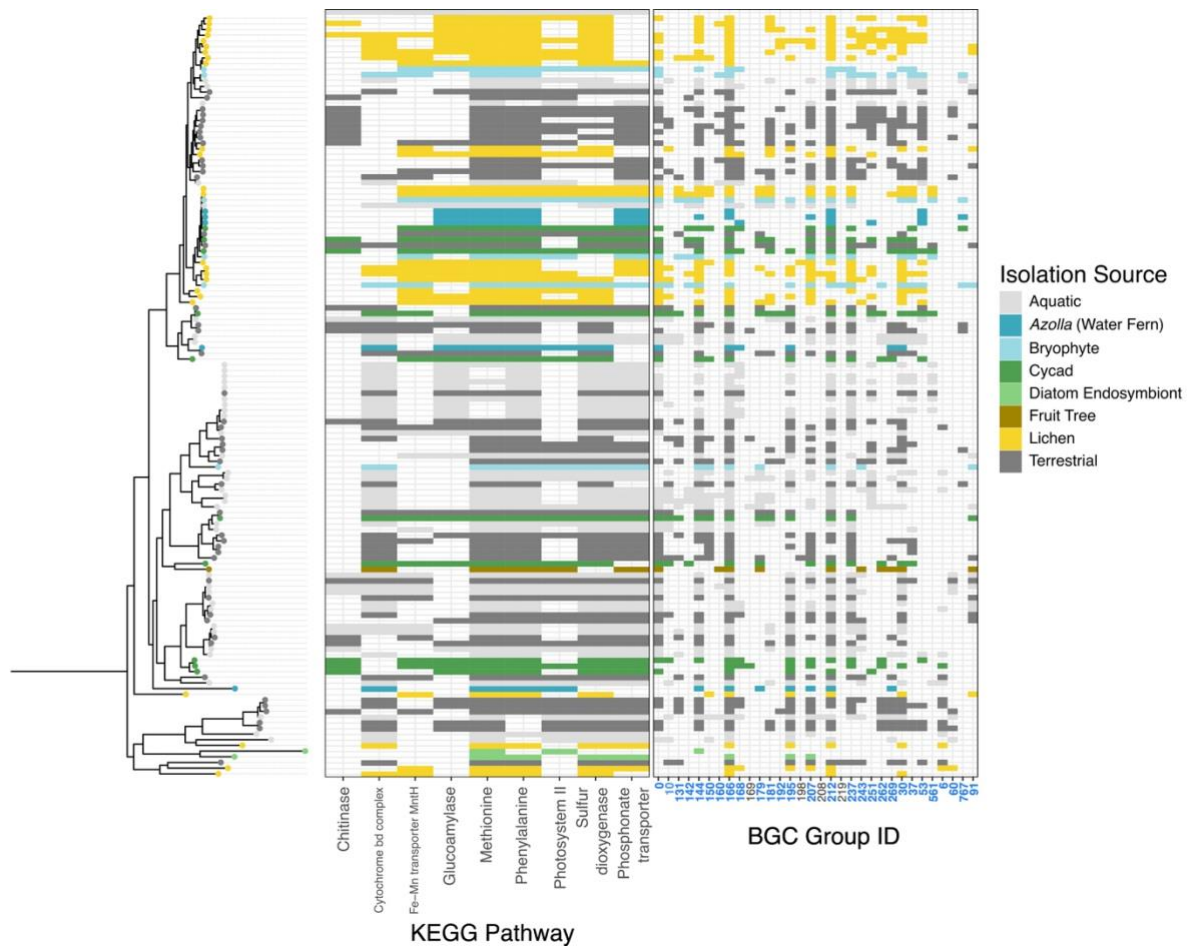
388 **3.3 Host-specific molecular specialization in Nostocaceae symbionts**

389 To identify host specialization of cyanobacterial symbionts in the family
390 Nostocaceae, the occurrence of KEGG functions across specific isolation sources
391 was assessed. A total of 69 complete KEGG functions were found across
392 Nostocaceae genomes. 5 of these were found in 99% (n=299) of Nostocaceae
393 genomes including functions of histidine, tyrosine and arginine metabolism,
394 nostoxanthin production and retinal biosynthesis. An additional 30 were found in
395 more than 90% of Nostocaceae genomes with functions including amino acid
396 metabolism, astaxanthin production, starch and glycogen degradation, riboflavin
397 biosynthesis and sulfolipid biosynthesis, and Type I secretion systems. Exploration
398 of indicative functions identified the ubiquitous distribution of many additional
399 functions present in 90% of Nostocaceae genomes, including nitrogen fixation, Sec-
400 SRP secretion pathways, chemotaxis, and cobalamin and thiamine biosynthesis.
401 Some of these ubiquitous functions had also been observed to be significantly
402 enriched in host-associated genomes at the phylum level. In addition to the
403 ubiquitous distribution of certain molecular functions, specific isolation sources were
404 also found to be associated with the prevalence of certain molecular functions
405 (Figure 4A; Table S7,S8; Figure S5). Sulfur dioxygenase (0.027; Est. -2.52) had a
406 significantly lower prevalence in symbionts isolated from the water fern, *Azolla*.
407 Lichen cyanobionts were shown to have functions that had either significantly
408 increased or decreased likelihood of occurrence. Phosphonate transporters
409 (p=0.012; Est. -1.23), methionine synthesis (p=0.026; Est. -1.82) and cytochrome bd
410 complex (p=0.037; Est=-1.03) were found to have a significantly lower prevalence in
411 lichen cyanobionts. Conversely, lichen cyanobionts had significantly higher likelihood
412 for Fe-Mn transporters (p=2.71e-03; Est. 1.48), glucoamylase (p=0.049; Est. 1.03)
413 and photosystem II (p=3.29e-03; Est. 1.61). Similarly, cycads symbionts were also
414 found to have significantly higher likelihood for complete pathways for glucoamylase
415 (p=2.48e-03; Est. 2.23) and chitinase (p=0.021; Est. 1.252), and nearly significant
416 increased likelihood for Fe-Mn transporters (p=0.057; Est. 1.26).

417 The distribution and occurrence of classes of BGCs in the family Nostocaceae
418 revealed trends correlated with host identity (Table S11; Figure S6). The
419 cyanobacterial symbionts of the water fern, *Azolla*, were found to consistently have a
420 significantly lower number of total BGCs (p=5.19e-05; Est. -12.32), nonribosomal

421 peptides ($p=2.54e-03$; Est. -2.26), nonribosomal peptide polyketides ($p=2.54e-03$;
422 Est. -1.53), RiPP ($p=8.45e-03$; -2.86), terpenes ($p=9.72e-03$; Est. -1.25), 'other'
423 ($p=9.54e-04$; -2.43), a class of BGC that does not fit into properties of otherwise
424 described secondary metabolites. Other symbionts were also found to have a
425 significantly lower number of BGCs including fruit tree symbionts with a significantly
426 lower number of non-ribosomal peptides ($p=0.046$; Est. -3.85), saccharide terpenes
427 in cycad symbionts ($p=0.03$; Est. -0.095) and lichen cyanobionts with a significantly
428 lower number of RiPPs ($p=2.84e-03$; Est. -2.09). In addition to reduced counts, some
429 symbionts were found to have significantly increased numbers of terpenes ($p=0.019$;
430 Est. 1.00), alkaloid terpenes ($p=0.016$; Est. 0.16), and NRP polyketides ($p=8.64e-03$;
431 1.33) in bryophytes symbionts, and polyketide saccharides ($p=4.88e-29$; Est. 1.00) in
432 fruit tree fruit tree symbionts.

433 All 32 groups of BGCs that were shown to be significantly impacted by
434 lifestyle classification were detected in the family Nostocaceae (Figure 4B; Figure
435 S7). Of these, 28 groups had a significantly non-random evolutionary-distribution,
436 and those which had non-significant phylogenetic signal (groups 169, 198, 208 and
437 219) were sparsely present within this family. 21 BGC groups were identified to be
438 significantly impacted by specific isolation source with a significantly increased
439 prevalence being observed commonly in multiple terrestrial host-associated
440 environments (e.g., cycad, lichen, bryophytes) alongside free-living terrestrial
441 cyanobacteria.



442

443 *Figure 4: Distribution of significant KEGG functions and groups of biosynthetic gene*
 444 *clusters impacted by isolation source in Nostocaceae genera which include host-*
 445 *associated cyanobacterial symbionts.*

446 (A) KEGG functions found to be significantly impacted by specific isolation sources
 447 including host-associated symbionts from cycads, lichens and the water fern, *Azolla*.

448 (B) The distribution of 32 BGC groups identified as being significantly impacted by
 449 lifestyle-classification (i.e. free-living vs. host-associated in genera of Nostocaceae
 450 with host-associate cyanobacterial symbionts. Group names shown in bolded blue
 451 font face indicate a significantly non-random phylogenetic distribution indicating
 452 shared evolutionary history.

453

454 **4. Discussion**

455 We have compiled and analysed a large dataset of high-quality cyanobacterial
456 genomes to explore the distribution of taxa that are associated with eukaryotic hosts,
457 and to investigate the biochemical diversity and commonalities that distinguish
458 symbionts and free-living isolates. These features could be observed broadly at the
459 phylum level in both molecular functions (as predicted through KEGG orthologs) and
460 BGCs. Broadly, these specialized functions can be summarized into 4 key
461 categories: nitrogen fixation, carbohydrate utilization, environmental communication,
462 and mediation of biotic interactions via secondary metabolite production. We both
463 confirm some of the current understanding of cyanobacterial symbiotic associations
464 and identify novel host specific features in symbiont genomes.

465 The provision of fixed nitrogen to their eukaryotic hosts is one of the key
466 benefits of cyanobacterial symbiosis in both plant^{9,11,12} and lichen systems^{51,52}. We
467 found enrichment of nitrogen fixation in host-associated cyanobacterial symbionts
468 across the phylum and ubiquitously in the family Nostocaceae, supporting this as
469 one of the key mutualistic beneficial services. Nitrogen fixation in cyanobacteria
470 requires iron⁵³ and has also been shown to require manganese in legume nodule
471 bacterial symbionts^{54,55}, and we demonstrated increased occurrence of Fe-Mn
472 transporters in host-associated cyanobacteria at the phylum level and in cycad and
473 lichen symbionts within the family Nostocaceae.

474 Carbohydrate-active enzymes including chitinase, glucoamylase and L-lactate
475 dehydrogenase were found to have a significantly higher prevalence in host-
476 associated cyanobacterial symbionts. Notably, in the family Nostocaceae, chitinase
477 was only found to have a significantly higher prevalence in cycad symbionts. Chitin,
478 a highly abundant polysaccharide, is a key component in the cell walls of fungi^{56,57}
479 and may serve as a source of nitrogen for cyanobacterial and algal growth⁵⁶. The
480 presences of carbohydrate utilization genes in bacteria are related to the habitats
481 they are isolated from, with enrichment of carbohydrate metabolism correlated with
482 the carbohydrate composition of the environment⁵⁸. The potential for microbes to
483 target the fungal cell wall to prevent pathogenic fungal infection of plant hosts⁵⁷
484 suggests a potential additional mutualistic benefit of the cyanobacterial symbionts
485 found in cycads. The relative absence of chitinase activity loci in lichen symbionts

486 demonstrates a potential selection against antifungal activity and a key difference in
487 fungal versus plant-cyanobacterial symbioses. While the other enriched
488 carbohydrate-active enzymes observed at the phylum level were not found to be
489 enriched in specific host types, it will be interesting to explore in more detail the
490 trends in distribution of carbohydrate active enzymes in cyanobacteria to align these
491 results with patterns previously reported across the prokaryotic tree of life⁵⁸.

492 With the exception of diatom endosymbionts and the water fern, *Azolla*¹¹, the
493 majority of cyanobacterial symbionts are not permanently associated with the host.
494 Thus, cyanobacterial symbionts require the ability to sense and locate hosts. This
495 may be achieved through chemotaxis involving signal transduction pathways in
496 response to chemical attractants produced by plants⁵⁹ and the ability to sense
497 chemoattractants has proven to be critical in the formation of plant symbioses^{59,60}.
498 Consideration of partially complete KEGG functions revealed chemotaxis to have a
499 higher prevalence in host-associated cyanobacteria, but is not significant ($p=0.057$,
500 near significant). This function was also observed across the Nostocaceae taxa
501 correlating with the occurrence of host-associated symbionts. The enrichment of
502 motility functions has also been previously reported in terrestrial cyanobacteria⁶¹. As
503 the majority of these symbiotic associations, especially true of those found in
504 terrestrial systems, are facultative for the cyanobacteria^{9,11}, this raises the important
505 question of whether free-living cyanobacteria that possess these characteristics are
506 also potential symbiotic partners and whether the diversity of symbiotically
507 competent cyanobacteria is significantly higher than currently reported.

508 In addition to the ability to sense and respond to their environment, two
509 secretion systems (Type I secretion systems and Sec-SRP) were also found to have
510 a significantly higher likelihood of occurrence in host-associated symbionts
511 suggesting specialization to release products into the environment. While other
512 secretion systems are known to be used to colonize hosts for pathogenic and
513 symbiotic activity (e.g., Type III secretion systems transporting product directly into a
514 eukaryotic cell)⁶², Type I secretion systems are capable of transporting products to
515 the extracellular space in a single step⁶³. As observed in bacteria that promote plant
516 growth, the benefit of these microbial partners is often dependent on the secretion
517 systems⁶⁴. However, in the case of the cyanobacterial symbionts, the questions of
518 what beneficial and symbiotically critical compounds may be produced and released

519 by these organisms and how they vary depending on the eukaryotic host remains
520 unexplored.

521 One of the most notable patterns in the distribution of classes of biosynthetic
522 gene clusters was observed in Nostocaceae symbionts of the water fern, *Azolla*.
523 These symbionts consistently had a significantly lower number of total BGCs, which
524 was paralleled in specific classes including nonribosomal peptides, nonribosomal
525 peptide polyketides, RiPPs, terpenes, and 'other'. Cyanobacterial symbionts of *Azolla*
526 represent the only currently known permanent obligate symbionts¹¹. As secondary
527 metabolites, particularly terpenes, often have roles in mediating complex ecological
528 interactions⁶, so the reduced BGC content in these obligate symbionts may be
529 representative of the reduced complexity of their environment. As *Azolla* symbionts
530 are permanently associated with their host, the requirement for response to
531 environmental stress and to mediate interactions with other organisms is reduced in
532 comparison to cyanobacterial symbionts located in facultative mutualisms where
533 they also need to survive as free-living bacteria.

534 Reduced numbers of RiPPs were observed in lichen symbionts. RiPPs have very
535 diverse functions ranging from quorum sensing to antifungal and antibacterial
536 properties⁶⁵. Metagenomic sequencing of lichens has forced a reconceptualisation of
537 the symbiosis from a one mycobiont-one photobiont model to one that encompasses
538 additional fungal partners and a diverse microbiome^{34,66}. This diversity may play a
539 critical role in the growth of the lichen³⁴. That lichen cyanobionts have fewer RiPPs
540 may reflect adaptation to coexistence in this diverse community, and is a topic
541 worthy of deeper analysis.

542 In contrast to overall reduced counts of biosynthetic gene clusters, symbionts
543 in bryophytes and fruit trees were found to have increased numbers of BGCs
544 predicted to produce terpenes, alkaloids, nonribosomal peptides, and polyketide
545 saccharides. These BGC systems may be responsible for important ecological
546 interactions¹⁸. Examination of specific unique groups of BGCs in the family
547 Nostocaceae notably revealed that these groups occur in both free-living and host-
548 associated cyanobacteria, and are often not restricted to individual host types. We
549 note that this pattern contrasts previous research suggesting niche specific BGCs
550 only in cycad symbionts³³. Cyanobacterial isolates from cycads have also been

551 shown to be capable of forming symbiotic associations in laboratory conditions with
552 mosses, mycorrhizal fungi and *Gunnera* (a flowering plant)¹³. This supports our
553 findings of the potential of unspecific host symbiotic competence in secondary
554 metabolite profiles as demonstrated by our large-scale analyses of cyanobacteria
555 and cyanobacterial symbionts.

556 Previous phylogenetic reconstruction of Cyanobacteria has presented
557 contrasting conclusions concerning the relationships of symbiotic isolates: (i)
558 proposing clades that are comprised of cycad, bryophyte and lichen symbionts³²;
559 (ii) separation into clades representative of extracellular or intracellular/extracellular
560 symbionts⁹; (iii) grouping of lichen symbionts⁶⁷; or (iv) grouping of plant-associated
561 symbionts³³. We found host-associated cyanobacteria were scattered across the
562 phylogeny, with few monophyletic clades of symbionts, as previously reported for
563 *Nostoc* isolates from lichen symbionts³¹. Monophyletic clades of cyanobionts
564 involved in symbioses were detected in isolates from diatom endosymbionts, *Dioon*
565 cycads, sets of Peltigeraceae lichens and the water fern, *Azolla*. In Nostocaceae the
566 basally arising host-associated samples corresponded to lichen symbionts
567 associated with the fungal families Coccocarpiaceae and Placynthiaceae. The other
568 Nostocaceae lichen symbionts analysed were associated with fungal family
569 Peltigeraceae, and were placed intermixed with free-living, *Azolla*-associated and
570 bryophyte-associated isolates. As the lichen fungal partner is known to display a
571 preference in photobiont acquisition^{68,69}, it may be that Coccocarpiaceae and
572 Placynthiaceae fungi have a different range of potential partners than the
573 Peltigeraceae. It will be highly informative to generate genomic data for additional,
574 diverse cyanolichens.

575 In many cyanobacterial symbioses the symbiont may be found in a host
576 association or as a free-living form: these life habits are not mutually exclusive. The
577 availability of free-living cyanobacteria in surrounding environments influences the
578 symbiotic partners found in host associations^{11,70} and free-living cyanobacteria
579 closely related to symbiont clades may prove to be potential symbiotic partners. The
580 increased prevalence of specific BGCs observed across both free-living
581 cyanobacteria in terrestrial environments and symbionts found in terrestrial host-
582 associations (e.g., lichens, cycads, bryophytes) further demonstrates this potential
583 for an increased diversity in cyanobacterial symbionts than has currently been

584 observed. Future research focused on generating novel cyanobacterial genomes
585 from additional symbiotic associations will be critical in advancing the understanding
586 of host range and symbiont diversity in the phylum Cyanobacteria.

587 5. References

- 588 1. Demoulin, C. F. *et al.* Cyanobacteria evolution: Insight from the fossil record.
589 *Free Radic Biol Med* **140**, 206–223 (2019).
- 590 2. Shestakov, S. V. & Karbysheva, E. A. The origin and evolution of
591 cyanobacteria. *Biology Bulletin Reviews* **7**, 259–272 (2017).
- 592 3. Moreira, C., Vasconcelos, V. & Antunes, A. Phylogeny and biogeography of
593 cyanobacteria and their produced toxins. *Marine Drugs* vol. 11 4350–4369
594 (2013).
- 595 4. Fidor, A., Konkol, R. & Mazur-Marzec, H. Bioactive peptides produced by
596 cyanobacteria of the genus nostoc: A review. *Marine Drugs* vol. 17 (2019).
- 597 5. Plaas, H. E. & Paerl, H. W. Toxic Cyanobacteria: A Growing Threat to Water
598 and Air Quality. *Environmental Science and Technology* vol. 55 44–64 (2021).
- 599 6. Leão, P. N., Engene, N., Antunes, A., Gerwick, W. H. & Vasconcelos, V. The
600 chemical ecology of cyanobacteria. *Natural Product Reports* vol. 29 372–391
601 (2012).
- 602 7. Usher, K. M., Bergman, B. & Raven, J. A. Exploring cyanobacterial
603 mutualisms. *Annual Review of Ecology, Evolution, and Systematics* vol. 38
604 255–273 (2007).
- 605 8. Bergman, B., Matveyev, A. & Rasmussen, U. Chemical signalling in
606 cyanobacterial-plant symbioses. *Trends in Plant Science* vol. 1 191–197
607 (1996).
- 608 9. Warshan, D. *et al.* Genomic changes associated with the evolutionary
609 transitions of nostoc to a plant symbiont. *Mol Biol Evol* **35**, 1160–1175 (2018).
- 610 10. Hyvärinen, M., Härdling, R. & Tuomi, J. Cyanobacterial lichen symbiosis: The
611 fungal partner as an optimal harvester. *Oikos* **98**, 498–504 (2002).
- 612 11. De Vries, S. & De Vries, J. Evolutionary genomic insights into cyanobacterial
613 symbioses in plants. *Quantitative Plant Biology* vol. 3 (2022).

- 614 12. Adams, D. G. & Duggan, P. S. Cyanobacteria-bryophyte symbioses. *Journal of*
615 *Experimental Botany* vol. 59 1047–1058 (2008).
- 616 13. Meeks, J. C. *et al.* An Overview of the Genome of Nostoc Punctiforme, a
617 Multicellular, Symbiotic Cyanobacterium. *Photosynthesis Research* vol. 70
618 (2001).
- 619 14. Hillman, K. & Goodrich-Blair, H. Are you my symbiont? Microbial polymorphic
620 toxins and antimicrobial compounds as honest signals of beneficial symbiotic
621 defensive traits. *Current Opinion in Microbiology* vol. 31 184–190 (2016).
- 622 15. Vining, L. C. Functions OF Secondary Metabolites. *Annu Rev Microbiol* **44**,
623 395–427 (1990).
- 624 16. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate
625 experimentally validated biosynthetic gene clusters. *Nucleic Acids Res* **51**,
626 D603–D610 (2023).
- 627 17. Shibl, A. A. *et al.* Diatom modulation of select bacteria through use of two
628 unique secondary metabolites. *Proc Natl Acad Sci* 117 (2020).
- 629 18. Modolon, F., Barno, A. R., Villela, H. D. M. & Peixoto, R. S. Ecological and
630 biotechnological importance of secondary metabolites produced by coral-
631 associated bacteria. *Journal of Applied Microbiology* vol. 129 1441–1457
632 (2020).
- 633 19. O'Brien, J. & Wright, G. D. An ecological perspective of microbial secondary
634 metabolism. *Current Opinion in Biotechnology* vol. 22 552–558 (2011).
- 635 20. Hillman, K. & Goodrich-Blair, H. Are you my symbiont? Microbial polymorphic
636 toxins and antimicrobial compounds as honest signals of beneficial symbiotic
637 defensive traits. *Current Opinion in Microbiology* vol. 31 184–190 (2016).
- 638 21. Dittmann, E., Fewer, D. P. & Neilan, B. A. Cyanobacterial toxins: Biosynthetic
639 routes and evolutionary roots. *FEMS Microbiology Reviews* vol. 37 23–43
640 (2013).

- 641 22. Calcott, M. J., Ackerley, D. F., Knight, A., Keyzers, R. A. & Owen, J. G.
642 Secondary metabolism in the lichen symbiosis. *Chemical Society Reviews* vol.
643 47 1730–1760 (2018).
- 644 23. Calteau, A. *et al.* Phylum-wide comparative genomics unravel the diversity of
645 secondary metabolism in Cyanobacteria. *BMC Genomics* **15**, (2014).
- 646 24. Gautam, K., Tripathi, J. K., Pareek, A. & Sharma, D. K. Growth and secretome
647 analysis of possible synergistic interaction between green algae and
648 cyanobacteria. *J Biosci Bioeng* **127**, 213–221 (2019).
- 649 25. Liaimera, A. *et al.* Nostopeptolide plays a governing role during cellular
650 differentiation of the symbiotic cyanobacterium *Nostoc punctiforme*. *Proc Natl*
651 *Acad Sci U S A* **112**, 1862–1867 (2015).
- 652 26. Álvarez, C. *et al.* Symbiosis between cyanobacteria and plants: from molecular
653 studies to agronomic applications. *Journal of Experimental Botany* vol. 74
654 6145–6157 (2023).
- 655 27. Leikoski, N. *et al.* Genome mining expands the chemical diversity of the
656 cyanobactin family to include highly modified linear peptides. *Chem Biol* **20**,
657 1033–1043 (2013).
- 658 28. Wang, H., Fewer, D. P. & Sivonen, K. Genome mining demonstrates the
659 widespread occurrence of gene clusters encoding bacteriocins in
660 cyanobacteria. *PLoS One* **6**, (2011).
- 661 29. Chatterjee, P., Schafran, P., Li, F. W. & Meeks, J. C. Nostoc Talks Back:
662 Temporal Patterns of Differential Gene Expression during Establishment of
663 Anthoceros-Nostoc Symbiosis. *Molecular Plant-Microbe Interactions* **35**, 917–
664 932 (2022).
- 665 30. Stenroos, S., Högnabba, F., Myllys, L., Hyvönen, J. & Thell, A. High selectivity
666 in symbiotic associations of lichenized ascomycetes and cyanobacteria.
667 *Cladistics* **22**, 230–238 (2006).
- 668 31. Rikkinen, J., Oksanen, I. & Lohtander, K. Lichen guilds share related
669 cyanobacterial symbionts. *Science* vol. 297 357 (2002).

- 670 32. Bell-Doyon, P., Laroche, J., Saltonstall, K. & Villarreal Aguilar, J. C.
671 Specialized bacteriome uncovered in the coralloid roots of the epiphytic
672 gymnosperm, *Zamia pseudoparasitica*. *Environmental DNA* **2**, 418–428
673 (2020).
- 674 33. Gutierrez-Garcia, K. *et al.* Cycad coralloid roots contain bacterial communities
675 including cyanobacteria and *Caulobacter* spp. That encode niche-specific
676 biosynthetic gene clusters. *Genome Biol Evol* **11**, 319–334 (2019).
- 677 34. Tagirdzhanova, G. *et al.* Evidence for a core set of microbial lichen symbionts
678 from a global survey of metagenomes. *bioRxiv* 2023.02.02.524463 (2023)
679 doi:10.1101/2023.02.02.524463.
- 680 35. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W.
681 CheckM: Assessing the quality of microbial genomes recovered from isolates,
682 single cells, and metagenomes. *Genome Res* **25**, 1043–1055 (2015).
- 683 36. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A
684 toolkit to classify genomes with the genome taxonomy database.
685 *Bioinformatics* **36**, 1925–1927 (2020).
- 686 37. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast
687 and effective stochastic algorithm for estimating maximum-likelihood
688 phylogenies. *Mol Biol Evol* **32**, 268–274 (2015).
- 689 38. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for
690 phylogenetic tree display and annotation. *Nucleic Acids Res* **49**, W293–W296
691 (2021).
- 692 39. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**,
693 2068–2069 (2014).
- 694 40. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity
695 in a globally-distributed bacterial phototroph. *ISME Journal* **12**, 1861–1866
696 (2018).
- 697 41. Ives, A. R. & Garland, T. Phylogenetic logistic regression for binary dependent
698 variables. *Syst Biol* **59**, 9–26 (2010).

- 699 42. Tung Ho, L. S. & Ané, C. A linear-time algorithm for gaussian and non-
700 gaussian trait evolution models. *Syst Biol* **63**, 397–408 (2014).
- 701 43. Sanchez, S. *et al.* Expansion of novel biosynthetic gene clusters from diverse
702 environments using SanntiS. doi:10.1101/2023.05.23.540769.
- 703 44. Richardson, L. *et al.* MGnify: the microbiome sequence data analysis resource
704 in 2023. *Nucleic Acids Res* **51**, D753–D759 (2023).
- 705 45. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current
706 status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**,
707 D733–D745 (2016).
- 708 46. Blin, K. *et al.* AntiSMASH 7.0: New and improved predictions for detection,
709 regulation, chemical structures and visualisation. *Nucleic Acids Res* **51**, W46–
710 W50 (2023).
- 711 47. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of
712 communities in large networks. *Journal of Statistical Mechanics: Theory and*
713 *Experiment* **2008**, (2008).
- 714 48. Sørensen, T. A method of establishing groups of equal amplitude in plant
715 sociology based on similarity of species content, and its application to analysis
716 of vegetation on Danish commons. *Kong Dan Vidensk Selsk Biol Skr* **5**, 1–5
717 (1948).
- 718 49. Fritz, S. A. & Purvis, A. Selectivity in mammalian extinction risk and threat
719 types: A new measure of phylogenetic signal strength in binary traits.
720 *Conservation Biology* **24**, 1042–1051 (2010).
- 721 50. Orme, D. *et al.* caper: Comparative Analyses of Phylogenetics and Evolution in
722 R. (2023).
- 723 51. Prieto, M., Montané, N., Aragón, G., Martínez, I. & Rodríguez-Arribas, C.
724 Cyanobacterial Variability in Lichen Cephalodia. *Journal of Fungi* **9**, (2023).
- 725 52. Sanders, W. B. & Masumoto, H. Lichen algae: The photosynthetic partners in
726 lichen symbioses. *Lichenologist* vol. 53 347–393 (2021).

- 727 53. Berman-Frank, I., Quigg, A., Finkel, Z. V., Irwin, A. J. & Haramaty, L. Nitrogen-
728 fixation strategies and Fe requirements in cyanobacteria. *Limnol Oceanogr* **52**,
729 2260–2269 (2007).
- 730 54. Hood, G., Ramachandran, V., East, A. K., Downie, J. A. & Poole, P. S.
731 Manganese transport is essential for N₂-fixation by *Rhizobium leguminosarum*
732 in bacteroids from galeoid but not phaseoloid nodules. *Environ Microbiol* **19**,
733 2715–2726 (2017).
- 734 55. Yoch, D. C. yoch-1979-manganese-an-essential-trace-element-for-n2-fixation-
735 by-rhodospirillum-rubrum-and-rhodopseudomonas-capsulata. *J Bacteriol* 987–
736 995 (1979).
- 737 56. Blank, C. E. & Hinman, N. W. Cyanobacterial and algal growth on chitin as a
738 source of nitrogen; ecological, evolutionary, and biotechnological implications.
739 *Algal Res* **15**, 152–163 (2016).
- 740 57. Sánchez-Vallet, A., Mesters, J. R. & Thomma, B. P. H. J. The battle for chitin
741 recognition in plant-microbe interactions. *FEMS Microbiology Reviews* vol. 39
742 171–183, (2015).
- 743 58. López-Mondéjar, R., Tláskal, V., da Rocha, U. N. & Baldrian, P. Global
744 Distribution of Carbohydrate Utilization Potential in the Prokaryotic Tree of Life.
745 *mSystems* **7**, (2022).
- 746 59. Nilsson, M., Rasmussen, U. & Bergman, B. Cyanobacterial chemotaxis to
747 extracts of host and nonhost plants. *FEMS Microbiol Ecol* **55**, 382–390 (2006).
- 748 60. Duggan, P. S., Thiel, T. & Adams, D. G. Symbiosis between the
749 cyanobacterium *Nostoc* and the liverwort *Blasia* requires a CheR-type MCP
750 methyltransferase. *Symbiosis* **59**, 111–120 (2013).
- 751 61. Chen, M. Y. *et al.* Comparative genomics reveals insights into cyanobacterial
752 evolution and habitat adaptation. *ISME Journal* **15**, 211–227 (2021).
- 753 62. Green, E. R. & Meccas, J. Bacterial Secretion Systems: An Overview.
754 *Microbiol Spectr* **4**, (2016).

- 755 63. Delepelaire, P. Type I secretion in gram-negative bacteria. *Biochimica et*
756 *Biophysica Acta - Molecular Cell Research* vol. 1694 149–161 (2004).
- 757 64. Lucke, M., Correa, M. G. & Levy, A. The Role of Secretion Systems, Effectors,
758 and Secondary Metabolites of Beneficial Rhizobacteria in Interactions With
759 Plants and Microbes. *Frontiers in Plant Science* vol. 11 (2020).
- 760 65. Kloosterman, A. M. *et al.* Expansion of RiPP biosynthetic space through
761 integration of pan-genomics and machine learning uncovers a novel class of
762 lantibiotics. *PLoS Biol* **18**, (2020).
- 763 66. Grimm, M. *et al.* The Lichens' Microbiota, Still a Mystery? *Frontiers in*
764 *Microbiology* vol. 12, (2021).
- 765 67. Gagunashvili, A. N. & Andrésón, Ó. S. Distinctive characters of Nostoc
766 genomes in cyanolichens. *BMC Genomics* **19**, (2018).
- 767 68. Jüriado, I., Kaasalainen, U., Jylhä, M. & Rikkinen, J. Relationships between
768 mycobiont identity, photobiont specificity and ecological preferences in the
769 lichen genus *Peltigera* (Ascomycota) in Estonia (northeastern Europe). *Fungal*
770 *Ecol* **39**, 45–54 (2019).
- 771 69. Leavitt, S. D. *et al.* Fungal specificity and selectivity for algae play a major role
772 in determining lichen partnerships across diverse ecogeographic regions in the
773 lichen-forming family Parmeliaceae (Ascomycota). *Mol Ecol* **24**, 3779–3797
774 (2015).
- 775 70. Bouchard, R. *et al.* Contrasting bacteriome of the hornwort *Leiosporoceros*
776 *dussii* in two nearby sites with emphasis on the hornwort-cyanobacterial
777 symbiosis. *Symbiosis* **81**, 39–52 (2020).

778

779 **Competing Interests**

780 The authors declare no competing interests.

781 **Data Availability**

782 The data analysed during in this study are available from RefSeq and the European
783 Nucleotide Archive (ENA) repositories with accession numbers provided in
784 Supplementary Table S1.

785 **Acknowledgements**

786 This research was supported by EMBL (European Molecular Biology Laboratory)
787 core funds.