



OPEN

Diversity and specificity of molecular functions in cyanobacterial symbionts

Ellen S. Cameron^{1,2}, Santiago Sanchez¹, Nick Goldman¹, Mark L. Blaxter² & Robert D. Finn¹✉

Cyanobacteria are globally occurring photosynthetic bacteria notable for their contribution to primary production and production of toxins which have detrimental ecosystem impacts. Furthermore, cyanobacteria can form mutualistic symbiotic relationships with a diverse set of eukaryotes, including land plants, aquatic plankton and fungi. Nevertheless, not all cyanobacteria are found in symbiotic associations suggesting symbiotic cyanobacteria have evolved specializations that facilitate host-interactions. Photosynthetic capabilities, nitrogen fixation, and the production of complex biochemicals are key functions provided by host-associated cyanobacterial symbionts. To explore if additional specializations are associated with such lifestyles in cyanobacteria, we have conducted comparative phylogenomics of molecular functions and of biosynthetic gene clusters (BGCs) in 984 cyanobacterial genomes. Cyanobacteria with host-associated and symbiotic lifestyles were concentrated in the family Nostocaceae, where eight monophyletic clades correspond to specific host taxa. In agreement with previous studies, symbionts are likely to provide fixed nitrogen to their eukaryotic partners, through multiple different nitrogen fixation pathways. Additionally, our analyses identified chitin metabolising pathways in cyanobacteria associated with specific host groups, while obligate symbionts had fewer BGCs. The conservation of molecular functions and BGCs between closely related symbiotic and free-living cyanobacteria suggests the potential for additional cyanobacteria to form symbiotic relationships than is currently known.

Keywords Mutualism, Facultative symbioses, Host-association, Biosynthetic gene clusters, Phylogenomics

Cyanobacteria, a group of photosynthetic bacteria, have a long evolutionary history with fossil evidence dating back up to 1.9 billion years ago¹. These organisms are found globally in diverse habitats, from aquatic to terrestrial landscapes and from polar to tropical climates^{1–3}. Cyanobacteria have evolved adaptations for survival under numerous types of stressors including desiccation, extreme temperatures, salinity, UV radiation and pathogenic infections⁴. Cyanobacteria play diverse roles in ecosystems including threats from their production of potent toxins when blooms contaminate water sources⁵ to their significant contributions of global primary productivity and biogeochemical cycling in marine systems⁶. As such, much of cyanobacterial research has focused on marine strains and the public health impacts of these organisms⁷. However, the impact of cyanobacteria on ecosystem health extends beyond degrading water quality, as many of the members in this taxonomic group have been found to be critical partners in mutualistic symbiotic associations with a diverse range of eukaryotic hosts.

Keystone examples of cyanobacterial symbioses is that of the endosymbiotic event that occurred some 2.1 billion years ago and led to the development of chloroplasts and photosynthetic eukaryotes⁸, and the evolution of in endosymbiotic *Candidatus Atelocyanobacterium thalassa* into N₂-fixing organelles (nitroplasts) in marine algae⁹. Beyond the endosymbiont origin of chloroplasts and nitroplasts, cyanobacteria are also found in symbiotic associations with diverse hosts such as diatoms, haptophytes, dinoflagellates, metazoans, fungi, macroalgae and land plants in both terrestrial and aquatic environments^{8,10,11}. These symbionts provide hosts with beneficial services including photosynthetic products¹² and fixed nitrogen^{13,14}. The mode of host association is also variable, including epiphytic growth (e.g. on feathermoss¹⁵ or on the diatom *Chaetoceros compressus*¹⁶), endophytic growth (e.g. root colonization of rice (*Oryza sativa*)¹⁷, intra-organismal location in specialised symbiotic structures (e.g. coralloid roots of cycads¹⁸), and intracellular incorporation^{11,13}. These associations can be ancient with examples of cyanobacterial symbionts found in a fossilised lichen from 400 million years

¹European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge CB10 1SD, UK. ²Tree of Life, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. ✉email: rdf@ebi.ac.uk

ago⁸. Such long associations raise the potential for coevolution between the eukaryotic host and specialised cyanobacterial partners¹¹, selecting for symbiotic competence^{19,20}.

Host-symbiont interactions require pathways for communicating and detecting signals²⁰ which may involve secondary metabolites. Secondary metabolites, compounds that are not essential for primary growth and reproduction²¹, are produced by co-localized genes organized as biosynthetic gene clusters (BGC)²². These compounds are often specialized for species interaction and survival in stressful environments, and can include bioactive compounds with antibacterial, antifungal and cytotoxic properties^{4,21,22}. Secondary metabolites have previously been shown to impact symbiotic associations such as diatoms producing compounds to promote growth and attachment of beneficial bacteria²³, or coral microbiomes producing a high diversity of antimicrobial products²⁴. However, secondary metabolites are often produced for a specific physiological or ecological reason and are often taxon specific²⁵, with this specificity potentially being a mechanism for symbiont communication to their potential host²⁰.

Even amongst microbes known for their production of diverse secondary metabolites, cyanobacteria alone are known to produce over 1100 unique secondary metabolites and their genomes frequently contain a high number of BGCs^{26,27}. The majority of cyanobacterial genomes contain polyketide synthase and nonribosomal peptide synthetase pathways that account for up to 5% of their total genome sizes²⁸. The compounds that cyanobacteria produce span diverse roles ranging from UV protection (mycosporines and scytonemin) to grazing deterrents and nutrient scavenging⁷ which may provide additional competitive advantages to hosts²⁹. The compounds may also mediate inter-partner communication in symbioses. For example, the production of nostopeptolide in the cyanobacterium genus, *Nostoc*, is associated with repression of formation of infectious differentiated cells and is down-regulated in the presence of plant hosts^{30,31}. While genome mining approaches have identified many cyanobacterial biosynthetic gene clusters of unknown function^{32,33}, the potential for symbiosis-specific secondary metabolites and their distribution among lineages of cyanobacteria has not been fully explored.

Cross-talk between cyanobacteria and their host-species has been previously reported, ranging from the upregulation of transcription of ammonium and nitrate transporters³⁴ to influencing cell differentiation in the life cycle of *Nostoc*¹³. However, varying reports of host-specificity and phylogenetic clades of symbiont cyanobacteria^{11,13,35,36} requires a phylum-wide study to explore the origins of host-association in this ancient lineage. Uniquely, *Nostoc* has shown broad symbiotic competence with different eukaryotic hosts, yet there still remains questions on molecular drivers of these associations due to the potential of non-host specific responses as isolates from cycads have previously been shown to also enter into symbiotic associations with mosses, fungi and angiosperms (e.g., *Gunnera*)¹⁹. Previous research has identified niche-specific BGCs that have been connected to individual host-specific associations in cyanobacteria³⁷ suggesting the presence of specialized secondary metabolites associated with cyanobacterial symbionts. However, a large-scale analysis of all available cyanobacterial genomes within the context of symbiotic associations has not yet been conducted. In this work, we utilize comparative phylogenomic approaches to identify trends in distribution of (i) molecular functions and (ii) biosynthetic gene clusters which may mediate host-symbiont interactions in this phylum.

Methods

Cyanobacterial genomes, habitat annotation and quality control

Assembled genome sequence data for 1078 species belonging to the phylum Cyanobacteria were downloaded from NCBI RefSeq in January 2023 (Supplementary Table S1). An additional 27 metagenomic assembled genomes (MAGs) taxonomically classified as cyanobacteria from lichen sources³⁸ were included to provide additional examples of host-associated symbionts for a total of 1105 cyanobacterial genomes.

Wherever possible the sampled cyanobacteria were assigned to their source habitat(s) based on available sample metadata, associated publication(s) or metadata describing the original isolation reported by culture collections and as such may include inconsistencies in cases where supporting metadata could not be sourced or lacked sufficient detail. These habitat assignments include aquatic (e.g., freshwater, marine and man-made aquatic sources) and terrestrial (e.g. soils), as well as host-associated environments. Host associations include vascular and non-vascular plants, diatoms, haptophytes, fungi, macroalgae, and marine mammals (epidermal mats). Individual host species were grouped into broad taxonomic categories including bryophytes, cycads, fruit trees (*Garcinia macrophylla*), diatoms, haptophytes, and lichens. Water fern (*Azolla*) cyanobacterial symbionts were placed in their own category. These habitat annotations were also used for grouping the cyanobacteria into two broader lifestyle classifications: free-living and host-associated. Cyanobacterial genomes of which no specific source habitat could be discovered were excluded, leaving 1026 cyanobacterial genomes for comparative analyses.

Quality control filtering was performed using both CheckM³⁹ (v. 1.1.3) and CheckM2⁴⁰ (v. 1.0.1) where a genome had to pass quality filtering (> 90% complete; < 5% contamination) in one method. 984 high-quality (> 90% complete; < 5% contamination; Supplementary Table S2; Supplementary Fig. S1) cyanobacterial genomes were retained for phylogenetic tree reconstruction and downstream analysis. Representatives of Melainabacteria (n = 37), a basal non-photosynthetic lineage of cyanobacteria, were included as an outgroup.

Phylogenetic tree reconstruction

Taxonomic classification of genomes and generation of marker gene alignments was conducted using GTDBtk⁴¹ (v. 2.3.0; Supplementary Table S1). Phylogenetic trees were constructed for the final high-quality set of cyanobacterial genomes using IQ-TREE⁴² (v. 2.2.0). The analysis used the LG + I + G4 model as identified in the IQ tree model finder based on the Bayesian Information Criterion (BIC). A family-level phylogenetic tree for the family Nostocaceae (n = 300), rooted with representatives of the order Elainellales, was constructed using IQ-TREE and the LG + I + G4 model determined by BIC. Phylogenetic trees were visualised using iTOL⁴³ (v.5).

For phylum and family level trees, non-parametric bootstraps ($n = 1000$) were conducted with IQ-TREE to assess the robustness of phylogenetic inferences.

Genome annotation and KEGG completeness estimation

Cyanobacterial genomes were annotated with Prokka⁴⁴ (v.1.14.6) and the resulting gene predictions were functionally annotated with KofamScan⁴⁵ (v.1.3.0) to derive Kyoto Encyclopaedia of Genes and Genomes (KEGG)⁴⁶ ortholog annotations (Supplementary Table S4). KofamScan predictions were used with KEGG-Decoder⁴⁷ (v. 1.3) to generate a table representing molecular function completeness across samples (Supplementary Table S3). KEGG functions were classified as being present using two thresholds, either > 98% complete for a more stringent analysis of distribution and complete function, or > 50% complete for lower stringency examination for the potential presence of molecular functions, herein referred to as indicative functions (Supplementary Fig. S2). Presence/absence matrices generated for KEGG functions were used in a phylogenetic logistic regression⁴⁸ to identify enrichment of molecular functions based on lifestyle classification at the phylum level (Supplementary Tables S5, S6) and enrichment of molecular functions in individual isolation sources in the family Nostocaceae (Supplementary Tables S7, S8). Phylogenetic logistic regressions were conducted using the *phyloglm* function in the R package *phylolm*⁴⁹, using the penalised likelihood with Firth's correction and 100 bootstraps. Responses of lifestyle classification and isolation sources were defined as significant if the p-value was less than 0.05.

Initial assessments of nitrogen fixation capabilities focused on the iron-molybdenum dependent pathway. To explore the presence and distribution of the alternative vanadium-dependent pathway, the presence of KEGG orthologs for both pathways were mapped against a phylogenetic tree of genera of cyanobacteria from the family Nostocaceae which contain host-associated cyanobacterial symbionts.

Biosynthetic gene cluster prediction and classification

BGCs were predicted on cyanobacterial genomes using SanntiS⁵⁰ (v. 0.9.1) due to high performance on both isolate genomes and MAGs, thus providing consistent annotations across all genome types. The predictions were subsequently filtered to remove those occurring at the edges of contigs and those which were less than 3000 bp in length, reflective of the minimum length of BGCs observed in the MiBIG database²². BGCs were initially classified by SanntiS into standard classes such as ribosomally synthesised and post-translationally modified peptides (RiPPs), terpenes, nonribosomal peptides, polyketides, alkaloids, saccharides, and hybrid classes which represent BGCs that cover multiple biochemical classes (Supplementary Table S9). To detect enrichment of total and specific BGC classes in host-associated symbionts, phylogenetic linear regression was conducted at the phylum level (Supplementary Table S10) and in the Family Nostocaceae (Supplementary Table S11). This was performed with the *phylolm* function using 100 bootstraps and a lambda model for covariance.

To expand upon the basic BGC classifications provided by SanntiS and identify diversity in potential products, predicted BGCs in cyanobacteria were clustered with a large, reference set of biosynthetic gene clusters (the 'reference BGC collection termed RefBGC hereafter). RefBGC includes BGC predictions from running SanntiS on MGnify⁵¹ and RefSeq genomes⁵², as well as the BGCs found in MiBIG²², and subsequently refined to only include complete predictions. This clustering enabled the assignment of BGCs to more specific groups based on functional domain composition, utilizing the Louvain community detection method⁵³ and the Sørensen-Dice similarity coefficient⁵⁴. To refine the SanntiS BGC classification assigned to each group antiSmash⁵⁵ (v.7.0.0) predictions were also generated for RefSeq and used to provide more specific natural product annotations, thereby combining the breadth offered by SanntiS and the accurate BGC product assignments provided by antiSMASH. Groups of BGCs containing antiSmash predictions were retained as the final set of BGCs (Supplementary Table S12). The habitat source of each BGC group was used in phylogenetic logistic regression to identify enrichment of specialized biosynthetic gene clusters in cyanobacteria with different lifestyles (Supplementary Tables S13, S14). This was performed with the *phyloglm* function maximizing the penalized likelihood with Firth's correction across 100 bootstraps. Groups found to be significantly enriched at the phylum level were used to assess phylogenetic signal in the family Nostocaceae using the D-statistic⁵⁶ with the *phylo.d* function in the R package *caper*⁵⁷ (v.1.0.2) of lifestyle classification and isolation sources were defined as significant if p-value was less than 0.05.

Results

Enrichment of molecular functions and biosynthetic gene clusters in host-associated cyanobacterial symbionts

Using the taxonomic classifications based on GTDB the cyanobacterial genomes were assigned to 18 taxonomic orders and 42 families, which were monophyletic based on the GTDBtk phylogeny, thus facilitating rigorous interpretation of evolutionary relatedness of these organisms. Of these, Cyanobacteriales ($n = 582$) and PCC-6307 (representative of *Cyanobium gracile*; $n = 262$) comprised over 85% of available genome assemblies (Fig. 1A). Habitat sources were highly skewed, with aquatic environments ($n = 756$) representing > 75% of environmental sources for all genome assemblies. Notably, only 6% ($n = 65$) of assessed cyanobacterial genomes were isolated from host-associated environments including non-vascular and vascular plants, diatoms, haptophytes, seaweeds, metazoan epidermal mats and fungi. Within this, Cyanobacteriales accounted for 93% [5.9% of host-associations in all assessed cyanobacterial genomes; $n = 61$] of all host-associated cyanobacterial symbiont genomes including representatives from all detected habitat source classifications (Fig. 1B). NCBI taxonomy was also considered, however due to challenges with nested, non-monophyletic groupings based on current taxonomic nomenclature, comparisons based on 'taxonomic identity' were not possible. Nevertheless, similar trends were shown with NCBI taxonomy with a high proportion of genomes arising from the orders Synechococcales ($n = 429$) and Nostocales

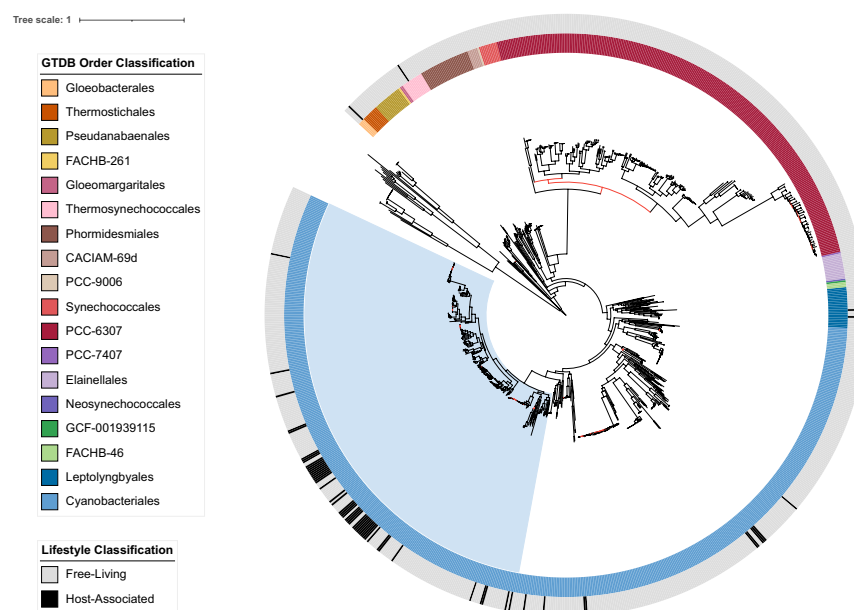
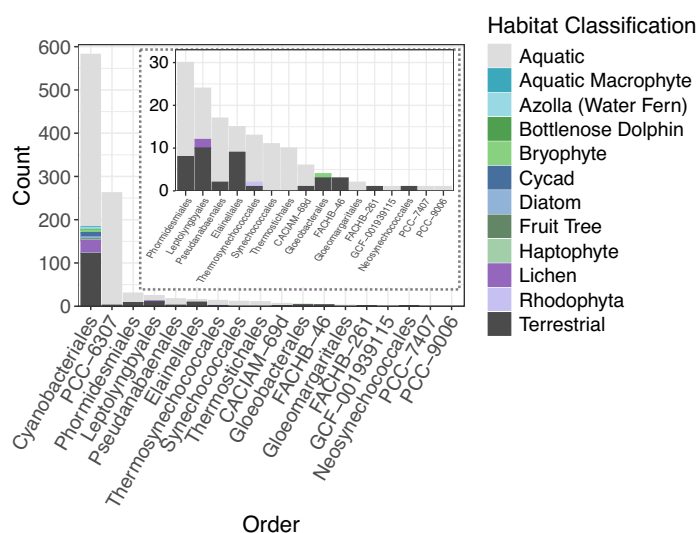
A**B**

Figure 1. Phylogeny and distribution of host-associated lifestyles in the phylum, Cyanobacteria. **(A)** Phylogeny generated using concatenated marker genes from GTDBtk of genome sequences of strains from phylum Cyanobacteria, rooted with representatives of the sister group, Melainabacteria, with 1000 bootstraps. Branches with low bootstrap support (< 50%) are shown in red. The outer annotation track depicts the lifestyle classification to highlight host-associated cyanobacterial symbionts. The inner annotation track depicts the classified taxonomic order assigned by GTDB. Nostocaceae, a family in the Cyanobacteriales, containing the majority of host associations, are shaded in light blue. **(B)** Frequency counts distributed across taxonomic orders for habitat classifications highlighting the different host sources including vascular plants (water fern (*Azolla*), cycad, a fruit tree (*Garcinia macrophylla*), aquatic macrophytes (*Hydrilla verticillata*), non-vascular plants (bryophytes), diatoms, haptophytes, macroalgae (Rhodophyta), fungi, and epidermal mats of aquatic mammals such as dolphins.

(n = 301) comprising nearly 75% of available reference genome assemblies with host-associations concentrated in the Nostocales (Supplementary Fig. S3).

KEGG functional annotations were analysed to identify molecular functions enriched in symbiont genomes by exploring the distribution of complete KEGG functions. In total, 78 complete KEGG functions were variably present across the phylum, of which 21 were significantly associated with lifestyle classification (Fig. 2A; Supplementary Fig. S4). Host-associated lifestyles were found to have a significantly higher level of occurrence

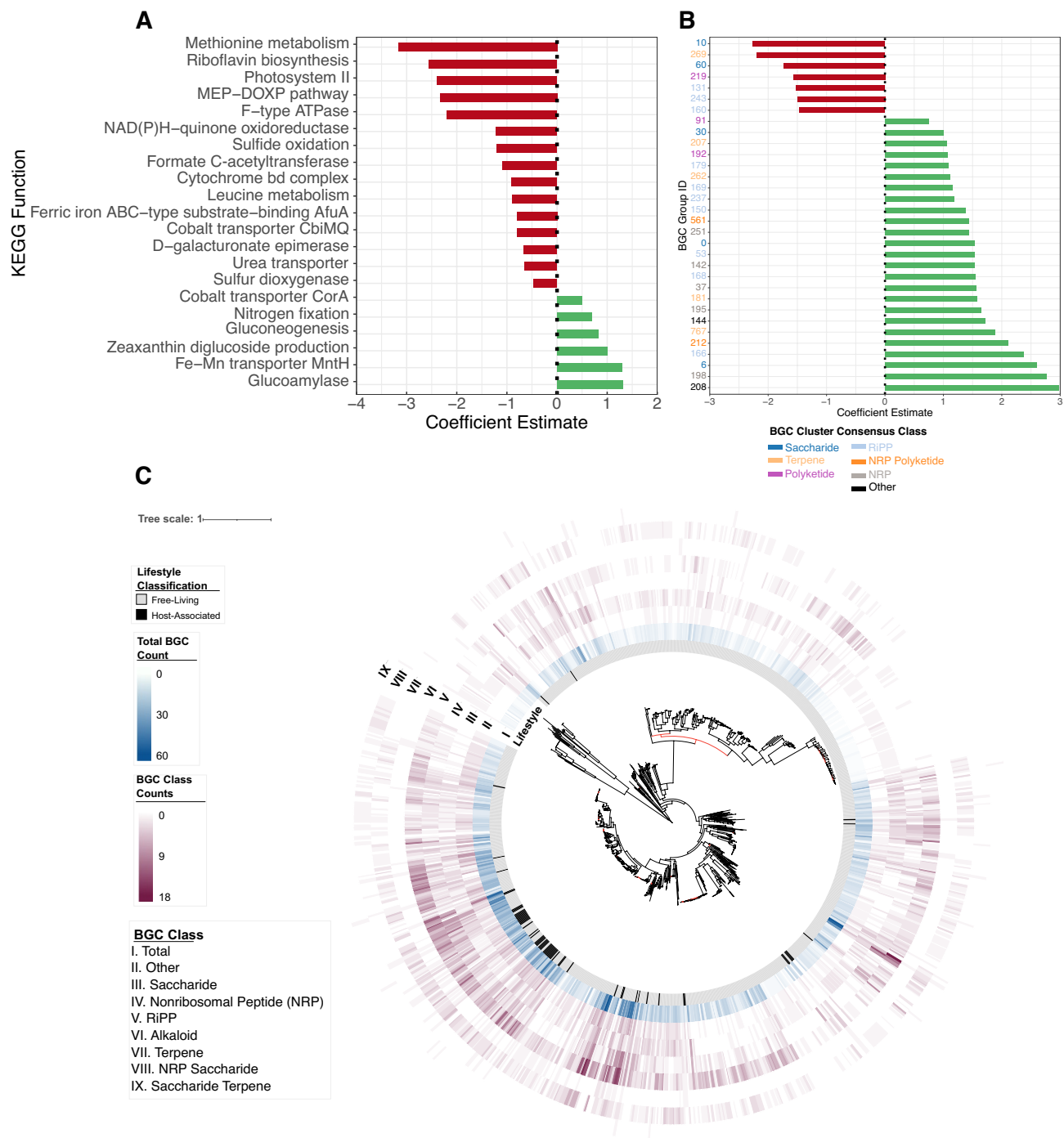


Figure 2. Host-associated enrichment of KEGG pathways and secondary metabolite production potential. **(A)** Coefficient estimates from phylogenetic regressions for KEGG functions shown to be significantly impacted by life-style classification including key functions corresponding to beneficial ecosystem services including nitrogen fixation. Negative coefficient estimates indicate decreased likelihood of detection in host-associated symbionts. Permission for utilizing the KEGG database was obtained from Kanehisa laboratories⁴⁶. **(B)** Coefficient estimates from phylogenetic regressions for unique groups of biosynthetic gene clusters (BGC), denoted by unique numeric identifiers, that are significantly impacted by lifestyle classification. Negative coefficient estimates indicate decreased likelihood of detection in host-associated symbionts. Axis labels for BGC group IDs are coloured according to the consensus of BGC class per group. **(C)** Distribution of counts of total detected biosynthetic gene clusters and classes of biosynthetic gene clusters shown to be significantly impacted by lifestyle classification across the phylum Cyanobacteria.

of functions including those of glucogenesis ($p=0.047$; Est. 0.83), Fe–Mn transporter ($p=6.35e-06$; Est. 1.30), glucoamylase ($p=4.58e-04$; Est. 1.31), zeaxanthin diglucoside production ($p=6.07e-04$; Est. 1.00), cobalt-magnesium transporters ($p=3.73e-02$; Est. 0.50) and nitrogen fixation ($p=1.88e-02$; Est. 0.69). Certain complete functions were also found to have a significantly lower occurrence in host-associated symbionts including photosystem II ($p=3.71e-11$; Est. – 2.39), the MEP-DOXP pathway ($p=8.35e-13$; Est. – 2.33), methionine ($p=8.29e-11$; Est. 3.16) and leucine metabolism ($p=1.11e-03$; Est. – 0.89), F-type ATPase ($p=3.82e-06$; Est. – 2.20), NAD(P) H-quinone oxidoreductase ($p=5.81e-06$; Est. – 1.21), riboflavin biosynthesis ($p=1.63e-13$; Est. – 2.55), sulfide oxidation ($p=2.41e-04$; Est. – 1.20), urea transporters ($p=8.86e-03$; Est. – 0.65), cytochrome bd complex ($p=2.03e-03$; Est. – 0.90), cobalt transport proteins (CbiMQ; $p=5.84e-03$; Est. – 0.79), D-galacturonate epimerase ($p=9.10e-03$; Est. – 0.66), formate dehydrogenase ($p=5.62e-03$; Est. – 1.09), sulfur dioxygenase ($p=0.047$; Est. – 0.47), and iron transport system binding proteins ($p=4.24e-03$; Est. – 0.80). Assessment of indicative functions also revealed a significantly higher likelihood of occurrences for Type I Secretion systems ($p=0.036$; Est. 0.46) and SecSRP secretion pathways ($p=0.02$; Est. 0.62). While not statistically significant, chemotaxis ($p=8.45e-02$; Est. 0.35) showed a higher likelihood of occurrence in host-associated symbionts when assessing indicative functions.

To further explore specializations associated with host-associated relationships in cyanobacteria, 8837 BGCs were identified across 97% ($n=954$) of all cyanobacterial genome assemblies. In total, 19 classes of biosynthetic gene clusters were identified including hybrid classes which span biochemical properties of multiple classifications. Lifestyle classification was found to significantly associate with the number of detected BGCs. Host-associated cyanobacteria were found to have a significantly lower numbers of BGCs in total ($p=4.34e-07$; Est. – 4.26) (Fig. 2C), and this trend was paralleled at the level of individual BGC class. Host-associated cyanobacterial symbionts were found to have significantly lower count of individual biosynthetic gene cluster classes including nonribosomal peptides (NRP; $p=0.011$; Est. – 0.55), RiPPs ($p=9.77e-05$; Est. – 1.23), alkaloids ($p=3.48e-04$; Est. – 0.16), terpenes ($p=1.36e-03$; Est. – 0.46), saccharides ($p=8.7e-03$; Est. – 0.47), saccharide terpenes ($p=0.012$; Est. – 0.036), NRP saccharides ($p=0.043$; Est. – 0.081), and other ($p=2.25e-05$; Est. – 0.76), a class of BGC that does not fit into properties of otherwise described secondary metabolites.

The 8837 biosynthetic gene clusters identified were classified into 124 unique groups representative of BGCs, which are likely to produce similar secondary metabolites based on similarity of the protein domain annotations. Although host-associated symbionts were found to have a significantly lower count of BGCs and classes of BGCs as a whole, individual BGC groups were found to be positively associated with cyanobacterial symbionts. Overall, 60 groups were found to be present in both free-living and host-associated cyanobacteria, 62 groups were found only in free-living cyanobacteria, and only 2 groups were found exclusively in host-associated symbionts corresponding to a terpene in a cycad symbiont and ‘other’ classification in aquatic macrophytes symbionts. Of the 61 groups found in both free-living and host-associated cyanobacteria, 25 were found to have a significantly higher prevalence in host-associated cyanobacteria (Fig. 2B; Supplementary Fig. S5), while only 7 showed a significantly decreased prevalence in host-associated symbionts (Supplementary Table S10).

Host-associated lifestyle appears non-specific with multiple origins in the *Nostocaceae*

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

Ten cyanobacterial genomes were sourced from cycad symbioses but only three of these were found to form a monophyletic clade. *Aulosira*, previously classified as *Nostoc*, comprised monophyletic clade VIII. These symbionts were all from a *Dioon* host supporting previous reports of monophyletic origin of endophytic cyanobacteria with this host species³⁷. Cyanobacteria from other cycad hosts (*Cycas revoluta* ($n=3$), *Macrozamia* ($n=1$), *Zamia pseudoparasitica* ($n=1$), *Encphalartos horridus* ($n=1$), and *Euterpe edulis* ($n=1$)) were distributed across the phylogeny. The genomes sourced from *Cycas revoluta* did not form a monophyletic clade and were distributed across the Nostocaceae tree. The cyanobacterium from the Arecales palm, *Euterpe edulis*, was found in a clade with the cyanobacterium from *Garcinia macrophylla*, a dicot (Malpighiales) fruit tree.

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

Five of the monophyletic clades, denoted II, III, IV, V and VII, contained 66% ($n=16$) of the analysed lichen cyanobionts, and their hosts were all Peltigeraceae fungi. Lichen cyanobionts most distant to the main lichen clades arose from lichens of different family lineages including *Coccocarpiia palmicola* (Coccocarpiaceae) and *Placynthium petersii* (Placynthiaceae) in more basal origins of the Nostocaceae. While all lichens observed in this analysis were of the order Peltigerales, the mycobiont from these lichens are in a different fungal family compared with those in the other analysed cyanolichens (Peltigeraceae), suggesting the potential for genomic diversity in cyanobionts depending on host identity.

Bryophyte cyanobionts did not form host-specific clades, but instead were often found in clades containing lichen cyanobionts or terrestrial isolates. Bryophyte cyanobionts were limited to three host species: *Blasia*

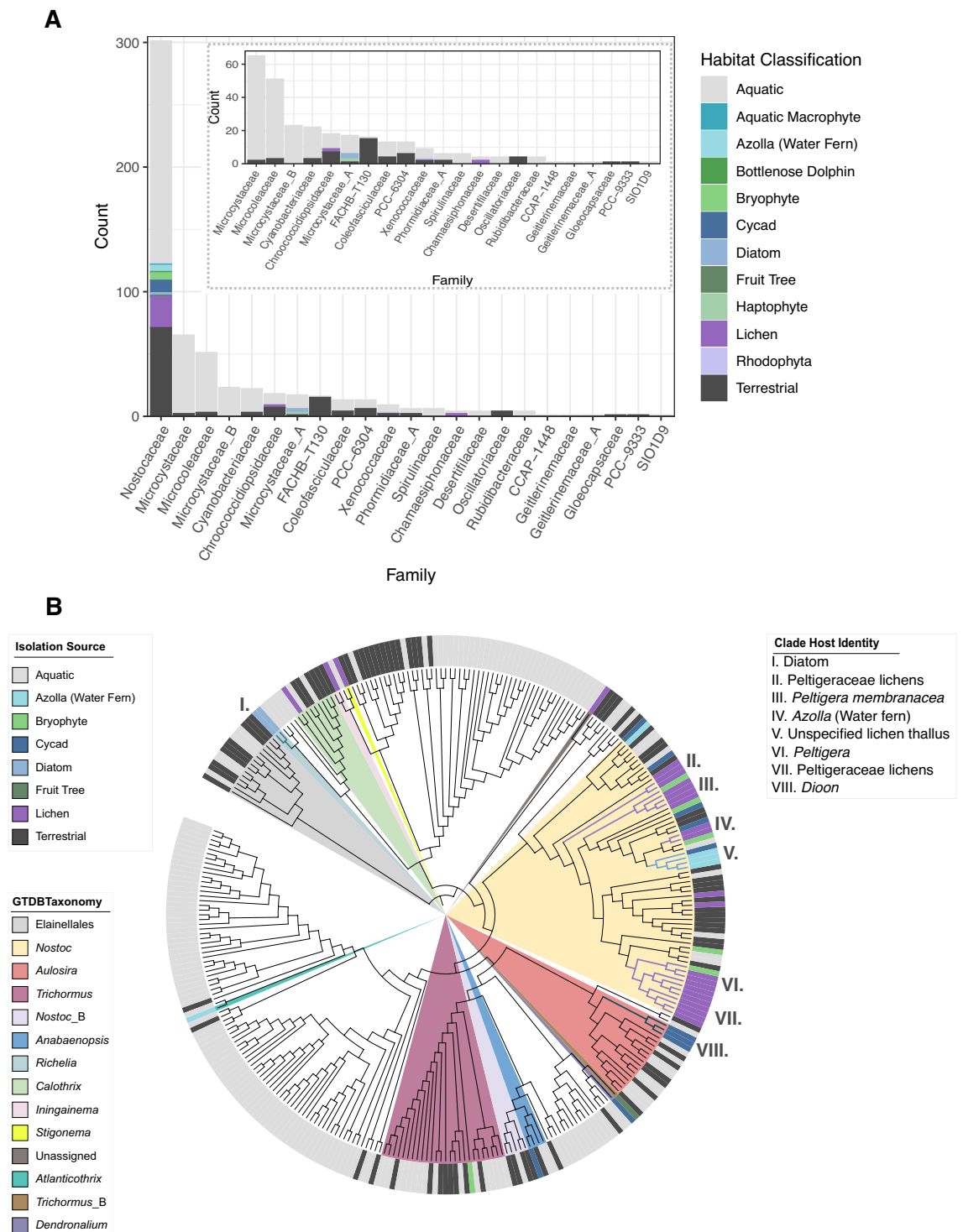


Figure 3. Distribution of host-types in the order Cyanobacteriales and the origin of host associations in Nostocaceae. **(A)** Frequency counts distributed across taxonomic families in the order Cyanobacteriales which includes the majority of host-associated cyanobacterial symbiont genomes spanning a high diversity of eukaryotic hosts in the family, Nostocaceae. Families with low frequency counts are displayed as an inset panel. Family names denoted with an alphabetical suffix represent groups that are not monophyletic in the GTDB reference tree or have unstable placement between releases of GTDB. **(B)** Cladogram of Nostocaceae generated from an alignment of marker genes rooted with the outgroup of Elainellales ($n = 15$) to explore the origin of host-specific association. Note that cyanobacterial symbiont UCYN-A Candidatus *Atelocyanobacterium thalassa* genomes fall within the family Microcystaceae_A according to GTDB and so are outside of subsequent analysis. Genus names denoted with an alphabetical suffix represent groups that are polyphyletic in the GTDB reference tree or have been subdivided in the GTDB reference tree. Genera with host-associations are highlighted, as well as a non-host associated genus of *Nostoc* (*Nostoc_B*). Branches with increased line width and highlighted in colour represent eight monophyletic clades containing symbionts arising from single host classifications.

pusilla (n = 3), *Phaeoceros* (n = 1), and *Leiosporoceros dussi* (n = 3). The multiple isolates from *Blasia pusilla* and *Leiosporoceros dussi* were distributed across the tree but commonly observed in clades with lichen cyanobionts.

Host-specific molecular specialization in *Nostocaceae* symbionts

To identify host specialization of cyanobacterial symbionts in the family *Nostocaceae*, the occurrence of KEGG functions across specific isolation sources was assessed. A total of 72 complete KEGG functions were found across *Nostocaceae* genomes. 19 of these were found in 99% (n = 298; Supplementary Fig. S6) of *Nostocaceae* genomes including functions of amino acid metabolism (cysteine, threonine, alanine, arginine, histidine, tyrosine, glycine, lysine, proline, serine, and tryptophan), nostoxanthin production, retinal biosynthesis, RuBisCo, starch and glycogen synthesis and degradation, copper transporters, and sulfate adenylyltransferase. An additional 17 were found in more than 90% of *Nostocaceae* genomes with functions including additional amino acid metabolism, astaxanthin production, riboflavin biosynthesis and sulfolipid biosynthesis, and Type I secretion systems. Exploration of indicative functions identified the ubiquitous distribution of many additional functions present in 90% of *Nostocaceae* genomes, including nitrogen fixation, Sec-SRP secretion pathways, chemotaxis, and cobalamin and thiamine biosynthesis. Some of these ubiquitous functions had also been observed to be significantly enriched in host-associated genomes at the phylum level. In addition to the ubiquitous distribution of certain molecular functions, specific isolation sources were also found to be associated with the prevalence of certain molecular functions (Fig. 4; Supplementary Tables S7, S8; Supplementary Fig. S7). Sulfur dioxygenase (0.027; Est. – 2.48) had a significantly lower prevalence in symbionts isolated from the water fern, *Azolla*. Lichen cyanobionts were shown to have functions that had either significantly increased or decreased likelihood of occurrence. Phosphonate transporters (p = 5.8e–03; Est. – 1.36), methionine synthesis (p = 0.017; Est. – 1.92) and cytochrome bd complex (p = 0.011; Est. – 1.27) were found to have a significantly lower prevalence in lichen cyanobionts. Conversely, lichen cyanobionts had significantly higher likelihood for Fe–Mn transporters (p = 2.03e–03; Est. 1.85), sulfur assimilation (p = 0.022; Est. 1.83) and photosystem II (p = 1.47e–03; Est. 2.13). Cycad symbionts were found to have significantly higher likelihood for complete pathways for glucoamylase (p = 1.57e–03; Est. 2.42) and chitinase (p = 5.08e–03; Est. 1.91), photosystem II (p = 0.047; Est. 1.84) and Fe–Mn transporters (p = 1.16e–03; Est. 3.14). Bryophyte symbionts were also found to have significantly higher likelihood for complete pathways for Fe–Mn transporters (p = 4.6e–02; Est. 1.71).

Pathways for both iron-molybdenum dependent and vanadium dependent nitrogen fixation were detected in free-living and host-associated cyanobacteria in *Nostocaceae* genera of cyanobacteria that possess host-associated cyanobacterial symbionts (Fig. 5). The Iron-molybdenum dependent nitrogen fixation pathway was much more common across the selected taxa. While the vanadium dependent pathway was more common in lichen and bryophyte symbionts, and the outlier case of an epiphytic *Azolla* symbiont. KEGG orthologs for the vanadium dependent pathway were found in 34% and 17% of free-living cyanobacteria in aquatic and terrestrial environments respectively. In contrast, KEGG orthologs for the vanadium dependent pathway were found in all bryophyte symbionts and over 50% of lichen symbionts. Vanadium dependent nitrogen fixation pathways were not detected in *Azolla*, *Garcinia macrophylla* or diatom symbionts, and were rarely detected in cycads (n = 1). Notably, both fixation pathways were incomplete with the absence of K00531 (*anfG*) in the iron-molybdenum dependent pathway, and K22899 (*vnfH*) in the vanadium dependent pathway.

The distribution and occurrence of classes of BGCs in the family *Nostocaceae* revealed trends correlated with host identity (Supplementary Table S11; Supplementary Fig. S8). The cyanobacterial symbionts of the water fern, *Azolla*, were found to consistently have a significantly lower number of total BGCs (p = 4.02e–05; Est. – 12.40), nonribosomal peptides (p = 1.98e–03; Est. – 2.30), nonribosomal peptide polyketides (p = 4.92e–03; Est. – 1.57), RiPP (p = 7.78e–03; – 2.86), terpenes (p = 7.96e–03; Est. – 1.27), ‘other’ (p = 7.45e–04; – 2.46), a class of BGC that does not fit into properties of otherwise described secondary metabolites. Other symbionts were also found to have a significantly lower number of BGCs including fruit tree symbionts with a significantly lower number of non-ribosomal peptides (p = 0.047; Est. – 3.85), saccharide terpenes in cycad symbionts (p = 0.031; Est. – 0.095) and lichen cyanobionts with a significantly lower number of RiPPs (p = 2.81e–03; Est. – 2.07) and total BGCs (p = 0.040; Est. – 3.86). In addition to reduced counts in some symbionts, others showed significantly increased numbers of specific classes. This included significantly increased numbers of terpenes (p = 0.017; Est. 1.00), alkaloid terpenes (p = 0.010; Est. 0.17), and NRP polyketides (p = 7.10e–03; 1.34) in bryophytes symbionts, and polyketide saccharides (p = 5.18e–29; Est. 1.00) in fruit tree symbionts.

All 32 groups of BGCs that were shown to be significantly impacted by lifestyle classification were detected in the family *Nostocaceae* (Fig. 6; Supplementary Fig. S9). Of these, 28 groups had a significantly non-random evolutionary-distribution, and those which had non-significant phylogenetic signal (groups 169, 198, 208 and 219) were sparsely present within this family. 21 BGC groups were identified to be significantly impacted by specific isolation source with a significantly increased prevalence being observed commonly in multiple terrestrial host-associated environments (e.g., cycad, lichen, bryophytes) alongside free-living terrestrial cyanobacteria.

Discussion

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

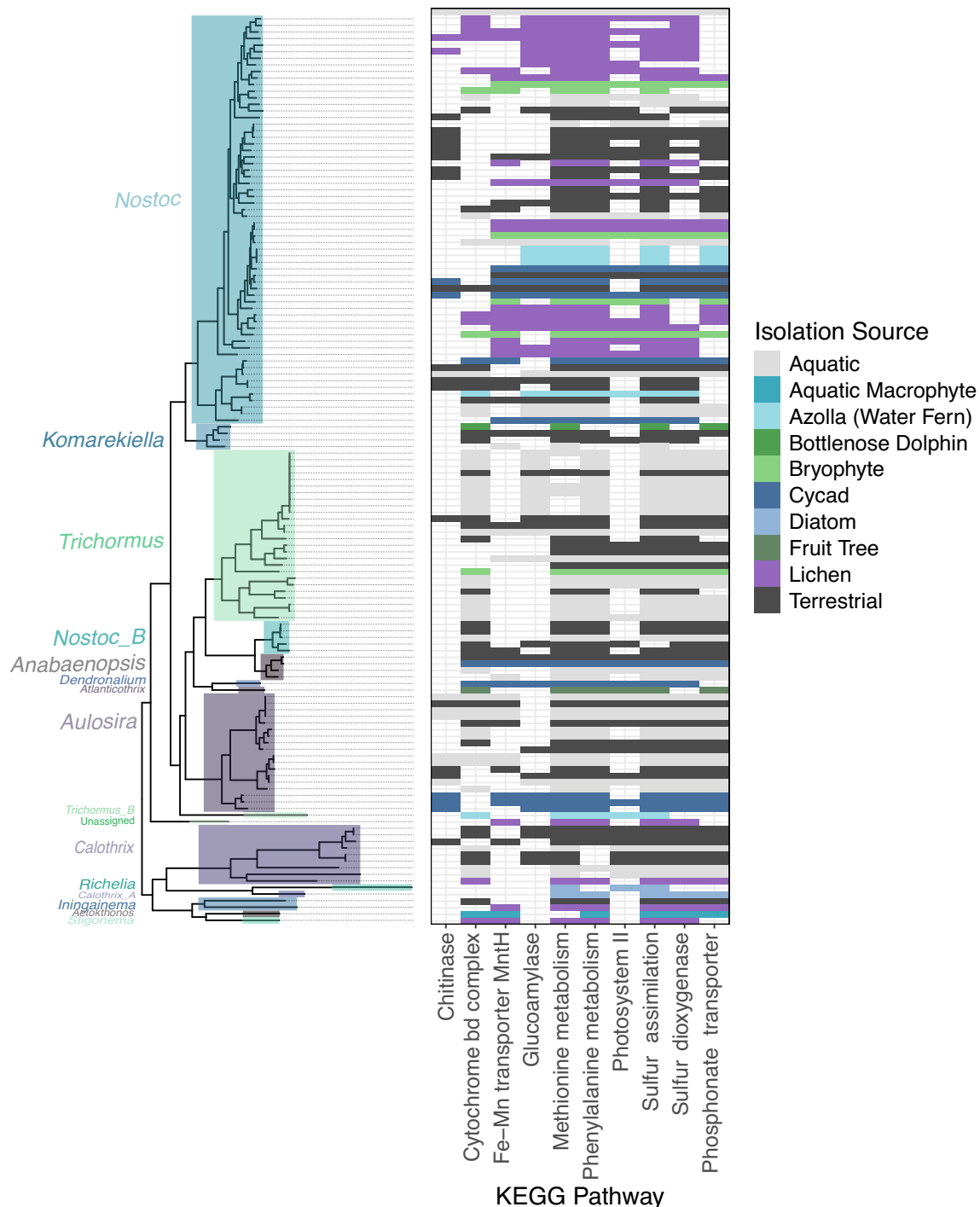


Figure 4. Distribution of significant KEGG functions impacted by isolation source in Nostocaceae genera which include host-associated cyanobacterial symbionts. KEGG functions found to be significantly impacted by specific isolation sources including host-associated symbionts from aquatic macrophytes, epidermal matss of the bottlenose dolphin, bryophytes, cycads, diatoms, lichens, a fruit tree species (*Garcinia macrophylla*) and the water fern, *Azolla*. Permission for utilizing the KEGG database was obtained from Kanehisa laboratories⁴⁶.

The provision of fixed nitrogen to their eukaryotic hosts is one of the key benefits of cyanobacterial symbiosis in plant^{11,13,14}, lichens^{58,59}, and planktonic marine protists⁶⁰ including diatoms⁶¹ and haptophytes⁶². We found enrichment of the iron-molybdenum dependent nitrogen fixation pathways in host-associated cyanobacterial symbionts across the phylum and ubiquitously in the family Nostocaceae, supporting this as one of the key mutualistic beneficial services. Nitrogen fixation in cyanobacteria requires iron⁶³ and has also been shown to require manganese in legume nodule bacterial symbionts^{64,65}, and we demonstrated increased occurrence of Fe-Mn transporters in host-associated cyanobacteria at the phylum level and in cycad and lichen

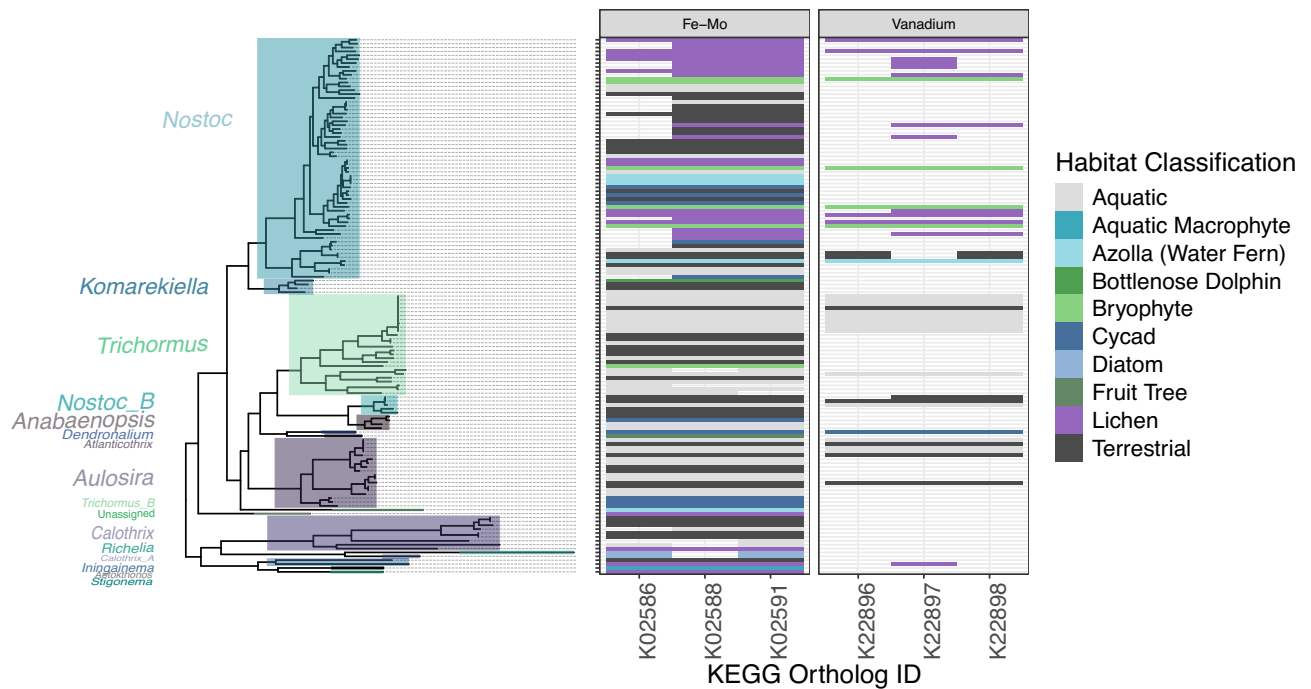


Figure 5. Distribution of KEGG orthologs associated with nitrogen-fixation pathways in Nostocaceae genera which include host-associated cyanobacterial symbionts. Presence of KEGG orthologs associated with iron-molybdenum and vanadium dependent nitrogen-fixation pathways in host-associated Nostocaceae genera. Distribution of KEGG orthologs in both pathways are shown across the phylogenetic tree on the left-hand side. Both pathway types were partially complete with the absence of K00531 (*anfG*) in the iron-molybdenum dependent pathway, and K22899 (*vnfH*) in the vanadium dependent pathway. Permission for utilizing the KEGG database was obtained from Kanehisa laboratories⁴⁶.

symbionts within the family Nostocaceae. Cyanobacterial symbionts were also shown to have a co-occurring alternative vanadium-dependent nitrogen fixation pathway presenting an alternative pathway in conditions of low molybdenum availability. This alternative pathway has previously been reported in other cyanobacterial symbionts arising from rice⁶⁶, bryophytes⁶⁷, and lichens^{68,69} supporting our detection of this pathway in lichen and bryophyte symbionts in this study. We rarely detected vanadium-dependent nitrogen fixation pathways in cycad coralloid root cyanobacterial symbionts suggesting that this pathway is not prevalent in all cyanobacterial symbionts. Similarly, alternative nitrogen fixation pathways have not previously been detected in other root associated bacteria of plants (e.g., *Rhizobia* root nodules of legumes)⁷⁰. Further investigation into the occurrence and impact of alternative nitrogen fixation in cyanobacterial symbionts is still required.

Carbohydrate-active enzymes including chitinase and glucoamylase were found to have a significantly higher prevalence in host-associated cyanobacterial symbionts. While cyanobacteria are known for their photoautotrophic metabolism, they are capable of heterotrophic growth utilizing carbohydrates such as glucose or xylanase⁷¹. The heterotrophic growth of cyanobacteria has been observed in *Nostoc* symbionts from cycads grown in the dark^{72,73}. Thus, the enrichment of carbohydrate active enzymes such as glucoamylase could drive the heterotrophic potential of host-associated cyanobacterial symbionts.

Specificities in carbohydrate utilization associated with host types was observed in the family Nostocaceae with chitinase having a significantly higher prevalence in cycad symbionts. Chitin, a highly abundant polysaccharide, is a key component in the cell walls of fungi^{74,75} and may serve as a source of nitrogen for cyanobacterial and algal growth⁷⁴ or for antifungal activity which may prove advantageous to the host. The presences of carbohydrate utilization genes in bacteria are related to the habitats they are isolated from, with enrichment of carbohydrate metabolism correlated with the carbohydrate composition of the environment⁷⁶. The potential for microbes to target the fungal cell wall to prevent pathogenic fungal infection of plant hosts⁷⁵ suggests a potential additional mutualistic benefit of the cyanobacterial symbionts found in cycads. The relative absence of chitinase activity loci in lichen symbionts demonstrates a potential selection against antifungal activity and a key difference in fungal versus plant-cyanobacterial symbioses. While the other enriched carbohydrate-active enzymes observed at the phylum level were not found to be enriched in specific host types, it will be interesting to explore in more detail the trends in distribution of carbohydrate active enzymes in cyanobacteria to align these results with patterns previously reported across the prokaryotic tree of life⁷⁶ and specificities of carbohydrate enzyme activity for host specific tissues such as the presence of lichenases in *Trebouxia* (chlorophyte algae) in lichen symbionts⁷⁷.

With the exception of the water fern, *Azolla*¹³, the majority of cyanobacterial symbionts are not permanently associated with the host. Thus, cyanobacterial symbionts require the ability to sense and locate hosts. This may be achieved through chemotaxis involving signal transduction pathways in response to chemical

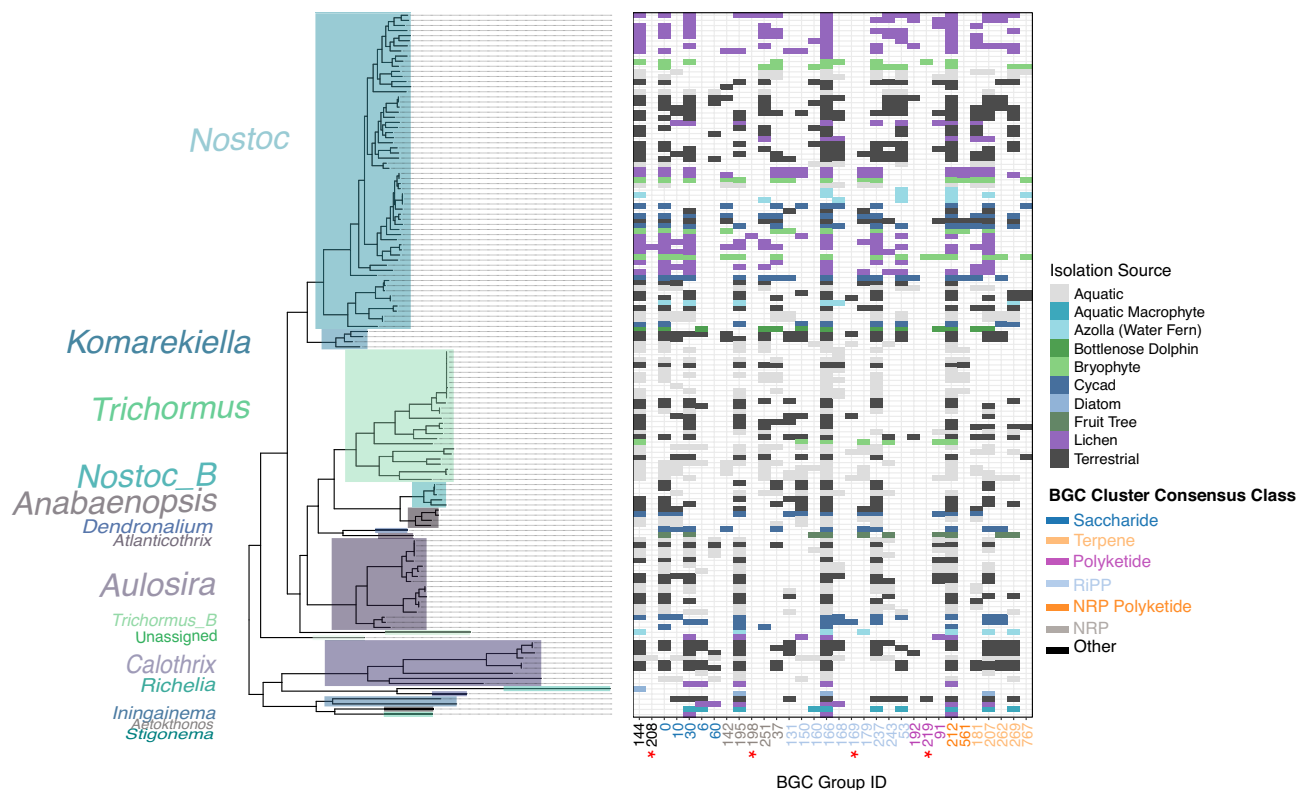


Figure 6. Distribution of biosynthetic gene cluster groups impacted by isolation source in Nostocaceae genera which include host-associated cyanobacterial symbionts. The distribution of 32 BGC groups, denoted by unique numeric IDs, identified as being significantly impacted by lifestyle-classification (i.e. free-living vs. host-associated in genera of Nostocaceae with host-associated cyanobacterial symbionts). Axis labels for BGC group IDs are coloured according to the consensus of BGC class per group. BGC group IDs marked with an asterisk (*) represent groups found to have a non-significant evolutionary distribution.

attractants produced by plants⁷⁸ and the ability to sense chemoattractants has proven to be critical in the formation of plant symbioses^{78,79}. Consideration of partially complete KEGG functions revealed chemotaxis to have a higher prevalence in host-associated cyanobacteria but is not significant ($p = 0.084$). This function was also observed across the Nostocaceae taxa correlating with the occurrence of host-associated symbionts. The enrichment of motility functions has also been previously reported in terrestrial cyanobacteria⁸⁰. As the majority of these symbiotic associations, especially true of those found in terrestrial systems, are facultative for the cyanobacteria^{11,13}, this raises the important question of whether free-living cyanobacteria that possess these characteristics are also potential symbiotic partners and whether the diversity of symbiotically competent cyanobacteria is significantly higher than currently reported.

In addition to the ability to sense and respond to their environment, two secretion systems (Type I secretion systems and Sec-SRP) were also found to have a significantly higher likelihood of occurrence in host-associated symbionts suggesting specialization to release products into the environment. While other secretion systems are known to be used to colonize hosts for pathogenic and symbiotic activity (e.g., Type III secretion systems transporting product directly into a eukaryotic cell)⁸¹, Type I secretion systems are capable of transporting products to the extracellular space in a single step⁸². As observed in bacteria that promote plant growth, the benefit of these microbial partners is often dependent on the secretion systems⁸³. However, in the case of the cyanobacterial symbionts, the questions of what beneficial and symbiotically critical compounds may be produced and released by these organisms and how they vary depending on the eukaryotic host remains unexplored.

One of the most notable patterns in the distribution of classes of biosynthetic gene clusters was observed in Nostocaceae symbionts of the water fern, *Azolla*. These symbionts consistently had a significantly lower number of total BGCs, which was paralleled in specific classes including nonribosomal peptides, nonribosomal peptide polyketides, RiPPs, terpenes, and 'other'. Cyanobacterial symbionts of *Azolla* represent the only currently known permanent obligate symbionts¹³. As secondary metabolites, particularly terpenes, often have roles in mediating complex ecological interactions⁷, so the reduced BGC content in these obligate symbionts may be representative of the reduced complexity of their environment. As *Azolla* symbionts are permanently associated with their host, the requirement for response to environmental stress and to mediate interactions with other organisms is reduced in comparison to cyanobacterial symbionts located in facultative mutualisms where they also need to survive as free-living bacteria.

Reduced numbers of RiPPs were observed in lichen symbionts. RiPPs have very diverse functions ranging from quorum sensing to antifungal and antibacterial properties⁸⁴. Metagenomic sequencing of lichens has forced a reconceptualisation of the symbiosis from a one mycobiont-one photobiont model to one that encompasses additional fungal partners and a diverse microbiome^{38,85}. This diversity may play a critical role in the growth of the lichen³⁸. That lichen cyanobionts have fewer RiPPs may reflect adaptation to coexistence in this diverse community, and is a topic worthy of deeper analysis.

In contrast to overall reduced counts of biosynthetic gene clusters, symbionts in bryophytes and fruit trees were found to have increased numbers of BGCs predicted to produce terpenes, alkaloids, nonribosomal peptides, and polyketide saccharides. These BGC systems may be responsible for important ecological interactions²⁴. Examination of specific unique groups of BGCs in the family Nostocaceae notably revealed that these groups occur in both free-living and host-associated cyanobacteria, and are often not restricted to individual host types. We note that this pattern contrasts previous research suggesting niche specific BGCs only in cycad symbionts³⁷. Cyanobacterial isolates from cycads have also been shown to be capable of forming symbiotic associations in laboratory conditions with mosses, mycorrhizal fungi and *Gunnera* (a flowering plant)¹⁹. The ability for cyanobacterial symbiont isolates to form associations with a broad range of hosts supports our findings of the potential of unspecific host symbiotic competence in secondary metabolite profiles as demonstrated by our large-scale analyses of cyanobacteria and cyanobacterial symbionts.

Previous phylogenetic reconstruction of Cyanobacteria has presented contrasting conclusions concerning the relationships of symbiotic isolates: (i) proposing clades that are comprised of cycad, bryophyte and lichen symbionts⁸⁶; (ii) separation into clades representative of extracellular or intracellular/extracellular symbionts¹¹; (iii) grouping of lichen symbionts⁸⁷; or (iv) grouping of plant-associated symbionts³⁷. We found host-associated cyanobacteria were scattered across the phylogeny, with few monophyletic clades of symbionts, as previously reported for *Nostoc* isolates from lichen symbionts³⁶. Monophyletic clades of cyanobionts involved in symbioses were detected in isolates from diatoms, *Dioon* cycads, sets of Peltigeraceae lichens and the water fern, *Azolla*. In Nostocaceae the basally arising host-associated samples corresponded to lichen symbionts associated with the fungal families Coccocarpiaceae and Placynthiaceae. The other Nostocaceae lichen symbionts analysed were associated with fungal family Peltigeraceae, and were placed intermixed with free-living, *Azolla*-associated and bryophyte-associated isolates. As the lichen fungal partner is known to display a preference in photobiont acquisition^{88,89}, it may be that Coccocarpiaceae and Placynthiaceae fungi have a different range of potential partners than the Peltigeraceae. It will be highly informative to generate genomic data for additional, diverse cyanolichens.

In many cyanobacterial symbioses the symbiont may be found associated with a host or as a free-living form: these life habits are not mutually exclusive. The availability of free-living cyanobacteria in surrounding environments influences the symbiotic partners found in host associations^{13,90} and free-living cyanobacteria closely related to symbiont clades may prove to be potential symbiotic partners. The increased prevalence of specific BGCs observed across both free-living cyanobacteria in terrestrial environments and symbionts found in terrestrial host-associations (e.g., lichens, cycads, bryophytes) further demonstrates this potential for an increased diversity in cyanobacterial symbionts than has currently been observed. Future research focused on generating novel cyanobacterial genomes from additional symbiotic associations will be critical in advancing the understanding of host range and symbiont diversity in the phylum Cyanobacteria.

Data availability

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

Received: 13 April 2024; Accepted: 1 August 2024

Published online: 12 August 2024

References

- Demoulin, C. F. *et al.* Cyanobacteria evolution: Insight from the fossil record. *Free Radic. Biol. Med.* **140**, 206–223 (2019).
- Shestakov, S. V. & Karbysheva, E. A. The origin and evolution of cyanobacteria. *Biol. Bull. Rev.* **7**, 259–272 (2017).
- Moreira, C., Vasconcelos, V. & Antunes, A. Phylogeny and biogeography of cyanobacteria and their produced toxins. *Mar. Drugs* **11**, 4350–4369 (2013).
- Fidor, A., Konkel, R. & Mazur-Marzec, H. Bioactive peptides produced by cyanobacteria of the genus *Nostoc*: A review. *Marine Drugs* **17**, (2019).
- Plaas, H. E. & Paerl, H. W. Toxic cyanobacteria: A growing threat to water and air quality. *Environ. Sci. Technol.* **55**, 44–64 (2021).
- Flombaum, P. *et al.* Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 9824–9829 (2013).
- Leão, P. N., Engene, N., Antunes, A., Gerwick, W. H. & Vasconcelos, V. The chemical ecology of cyanobacteria. *Nat. Prod. Rep.* **29**, 372–391 (2012).
- Usher, K. M., Bergman, B. & Raven, J. A. Exploring cyanobacterial mutualisms. *Annu. Rev. Ecol. Evol. Syst.* **38**, 255–273 (2007).
- Coale, T. H. *et al.* Nitrogen-fixing organelle in a marine alga. *Science* **384**, 217–222 (2024).
- Bergman, B., Matveyev, A. & Rasmussen, U. Chemical signalling in cyanobacterial-plant symbioses. *Trends Plant Sci.* **1**, 191–197 (1996).
- Warshan, D. *et al.* Genomic changes associated with the evolutionary transitions of *Nostoc* to a plant symbiont. *Mol. Biol. Evol.* **35**, 1160–1175 (2018).
- Hyvärinen, M., Härdling, R. & Tuomi, J. Cyanobacterial lichen symbiosis: The fungal partner as an optimal harvester. *Oikos* **98**, 498–504 (2002).
- De Vries, S. & De Vries, J. Evolutionary genomic insights into cyanobacterial symbioses in plants. *Quant. Plant Biol.* **3**, (2022).
- Adams, D. G. & Duggan, P. S. Cyanobacteria-bryophyte symbioses. *J. Exp. Bot.* **59**, 1047–1058 (2008).
- Warshan, D. *et al.* Feathermoss and epiphytic *Nostoc* cooperate differently: Expanding the spectrum of plant-cyanobacteria symbiosis. *ISME J.* **11**, 2821–2833 (2017).

16. Gómez, F., Furuya, K. & Takeda, S. Distribution of the cyanobacterium *Richelia intracellularis* as an epiphyte of the diatom *Chaetoceros compressus* in the western Pacific Ocean. *J. Plankton Res.* **27**, 323–330 (2005).
17. Álvarez, C., Navarro, J. A., Molina-Heredia, F. P. & Mariscal, V. Endophytic colonization of rice (*Oryza sativa* L.) by the symbiotic strain nostoc punctiforme PCC 73102. *MPMI* **33**, 1040–1045 (2020).
18. Chang, A. C. G., Chen, T., Li, N. & Duan, J. Perspectives on endosymbiosis in coralloid roots: Association of CYCADS AND CYANOBACTERIA. *Front. Microbiol.* **10**, 1888 (2019).
19. Meeks, J. C. *et al.* An overview of the genome of nostoc punctiforme, a multicellular, symbiotic cyanobacterium. *Photosynth. Res.* **70**, 85–106 (2001).
20. Hillman, K. & Goodrich-Blair, H. Are you my symbiont? Microbial polymorphic toxins and antimicrobial compounds as honest signals of beneficial symbiotic defensive traits. *Curr. Opin. Microbiol.* **31**, 184–190 (2016).
21. Vining, L. C. Functions of secondary metabolites. *Annu. Rev. Microbiol.* **44**, 395–427 (1990).
22. Terlouw, B. R. *et al.* MIBiG 3.0: A community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **51**, D603–D610 (2023).
23. Shibl, A. A. *et al.* Diatom modulation of select bacteria through use of two unique secondary metabolites. <https://doi.org/10.1073/pnas.2012088117/-/DCSupplemental>.
24. Modolon, F., Barno, A. R., Villela, H. D. M. & Peixoto, R. S. Ecological and biotechnological importance of secondary metabolites produced by coral-associated bacteria. *J. Appl. Microbiol.* **129**, 1441–1457 (2020).
25. O'Brien, J. & Wright, G. D. An ecological perspective of microbial secondary metabolism. *Curr. Opin. Biotechnol.* **22**, 552–558 (2011).
26. Dittmann, E., Fewer, D. P. & Neilan, B. A. Cyanobacterial toxins: Biosynthetic routes and evolutionary roots. *FEMS Microbiol. Rev.* **37**, 23–43 (2013).
27. Calcott, M. J., Ackerley, D. F., Knight, A., Keyzers, R. A. & Owen, J. G. Secondary metabolism in the lichen symbiosis. *Chem. Soc. Rev.* **47**, 1730–1760 (2018).
28. Calteau, A. *et al.* Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. *BMC Genom.* **15**, (2014).
29. Gautam, K., Tripathi, J. K., Pareek, A. & Sharma, D. K. Growth and secretome analysis of possible synergistic interaction between green algae and cyanobacteria. *J. Biosci. Bioeng.* **127**, 213–221 (2019).
30. Laiamera, A. *et al.* Nostopeptolide plays a governing role during cellular differentiation of the symbiotic cyanobacterium nostoc punctiforme. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 1862–1867 (2015).
31. Álvarez, C. *et al.* Symbiosis between cyanobacteria and plants: From molecular studies to agronomic applications. *J. Exp. Bot.* **74**, 6145–6157 (2023).
32. Leikoski, N. *et al.* Genome mining expands the chemical diversity of the cyanobactin family to include highly modified linear peptides. *Chem. Biol.* **20**, 1033–1043 (2013).
33. Wang, H., Fewer, D. P. & Sivonen, K. Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. *PLoS One* **6**, e22384 (2011).
34. Chatterjee, P., Schafran, P., Li, F. W. & Meeks, J. C. Nostoc talks back: Temporal patterns of differential gene expression during establishment of anthoceros-nostoc symbiosis. *Mol. Plant-Microbe Interact.* **35**, 917–932 (2022).
35. Stenroos, S., Högnabba, F., Myllys, L., Hyvönen, J. & Thell, A. High selectivity in symbiotic associations of lichenized ascomycetes and cyanobacteria. *Cladistics* **22**, 230–238 (2006).
36. Rikkinen, J., Oksanen, I. & Lohntander, K. Lichen guilds share related cyanobacterial symbionts. *Science* **297**, 357 (2002).
37. Gutierrez-Garcia, K. *et al.* Cycad coralloid roots contain bacterial communities including cyanobacteria and Caulobacter spp. that encode niche-specific biosynthetic gene clusters. *Genome Biol. Evol.* **11**, 319–334 (2019).
38. Tagirdzhanova, G. *et al.* Evidence for a core set of microbial lichen symbionts from a global survey of metagenomes. *BioRxiv* <https://doi.org/10.1101/2023.02.02.524463> (2023).
39. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
40. Chklovski, A., Parks, D. H., Woodcroft, B. J. & Tyson, G. W. CheckM2: A rapid, scalable and accurate tool for assessing microbial genome quality using machine learning. *Nat. Methods* **20**, 1203–1212 (2023).
41. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. *Bioinformatics* **36**, 1925–1927 (2020).
42. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
43. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v5: An online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
44. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
45. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
46. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
47. Graham, E. D., Heidelberg, J. F. & Tully, B. J. Potential for primary productivity in a globally-distributed bacterial phototroph. *ISME J.* **12**, 1861–1866 (2018).
48. Ives, A. R. & Garland, T. Phylogenetic logistic regression for binary dependent variables. *Syst. Biol.* **59**, 9–26 (2010).
49. Tung Ho, L. S. & Ané, C. A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Syst. Biol.* **63**, 397–408 (2014).
50. Sanchez, S. *et al.* Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. <https://doi.org/10.1101/2023.05.23.540769>.
51. Richardson, L. *et al.* MGnify: The microbiome sequence data analysis resource in 2023. *Nucleic Acids Res.* **51**, D753–D759 (2023).
52. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
53. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, (2008).
54. Sorenson, T. A. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analysis of vegetation on Danish commons. *Kong Dan Vidensk Selsk Biol. Skr.* **5**, 1–5 (1948).
55. Blin, K. *et al.* AntiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).
56. Fritz, S. A. & Purvis, A. Selectivity in mammalian extinction risk and threat types: A new measure of phylogenetic signal strength in binary traits. *Conserv. Biol.* **24**, 1042–1051 (2010).
57. Orme, D. *et al.* caper: Comparative analyses of phylogenetics and evolution in R. (2023).
58. Prieto, M., Montané, N., Aragón, G., Martínez, I. & Rodríguez-Arribas, C. Cyanobacterial variability in lichen cephalodia. *J. Fungi* **9**, 826 (2023).
59. Sanders, W. B. & Masumoto, H. Lichen algae: The photosynthetic partners in lichen symbioses. *Lichenologist* **53**, 347–393 (2021).

60. Decelle, J., Colin, S. & Foster, R. A. Photosymbiosis in marine planktonic protists. In *Marine Protists* (eds. Ohtsuka, S., Suzuki, T., Horiguchi, T., Suzuki, N. & Not, F.) 465–500 (Springer Japan, 2015). https://doi.org/10.1007/978-4-431-55130-0_19.
61. Flores, E., Romanovicz, D. K., Nieves-Mori6n, M., Foster, R. A. & Villareal, T. A. Adaptation to an intracellular lifestyle by a nitrogen-fixing, heterocyst-forming cyanobacterial endosymbiont of a diatom. *Front. Microbiol.* **13**, 799362 (2022).
62. Cornejo-Castillo, F. M. *et al.* Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat. Commun.* **7**, 11071 (2016).
63. Berman-Frank, I., Quigg, A., Finkel, Z. V., Irwin, A. J. & Haramaty, L. Nitrogen-fixation strategies and Fe requirements in cyanobacteria. *Limnol. Oceanogr.* **52**, 2260–2269 (2007).
64. Hood, G., Ramachandran, V., East, A. K., Downie, J. A. & Poole, P. S. Manganese transport is essential for N₂-fixation by *Rhizobium leguminosarum* in bacteroids from galeoid but not phaseoloid nodules. *Environ. Microbiol.* **19**, 2715–2726 (2017).
65. Yoch, D. C. yoch-1979-manganese-an-essential-trace-element-for-n₂-fixation-by-rhodospirillum-rubrum-and-rhodopseudomonas-capsulata. *J. Bacteriol.* **140**, 987–995 (1979).
66. Boison, G., Steingen, C., Stal, L. J. & Bothe, H. The rice field cyanobacteria *Anabaena azotica* and *Anabaena* sp. CH1 express vanadium-dependent nitrogenase. *Arch. Microbiol.* **186**, 367–376 (2006).
67. Nelson, J. M. *et al.* Complete genomes of symbiotic cyanobacteria clarify the evolution of vanadium-nitrogenase. *Genome Biol. Evol.* **11**, 1959–1964 (2019).
68. Hodgkinson, B. P. *et al.* Lichen-symbiotic cyanobacteria associated with Peltigera have an alternative vanadium-dependent nitrogen fixation system. *Eur. J. Phycol.* **49**, 11–19 (2014).
69. Darnajoux, R., Constantin, J., Miadlikowska, J., Lutzoni, F. & Bellenger, J.-P. Is vanadium a biometal for boreal cyanolichens?. *New Phytol.* **202**, 765–771 (2014).
70. Harwood, C. S. Iron-only and vanadium nitrogenases: Fail-safe enzymes or something more?. *Annu. Rev. Microbiol.* **74**, 247–266 (2020).
71. Stebbeg, R., Schmetterer, G. & Rempel, A. Heterotrophy among Cyanobacteria. *ACS Omega* **8**, 33098–33114 (2023).
72. Hoare, D. S., Ingram, L. O., Thurston, E. L. & Walkup, R. Dark heterotrophic growth of an endophytic blue-green alga. *Arch. Mikrobiol.* **78**, 310–321 (1971).
73. Tredici, M. R., Margheri, M. C., Giovannetti, L., Philippis, R. D. & Vincenzini, M. Heterotrophic metabolism and diazotrophic growth of *Nostoc* sp. from *Cycas circinalis*.
74. Blank, C. E. & Hinman, N. W. Cyanobacterial and algal growth on chitin as a source of nitrogen; ecological, evolutionary, and biotechnological implications. *Algal Res.* **15**, 152–163 (2016).
75. S6nchez-Vallet, A., Mesters, J. R. & Thomma, B. P. H. J. The battle for chitin recognition in plant-microbe interactions. *FEMS Microbiol. Rev.* **39**, 171–183 (2015).
76. L6pez-Mond6jar, R., Tl6skal, V., da Rocha, U. N. & Baldrian, P. Global distribution of carbohydrate utilization potential in the prokaryotic tree of life. *mSystems* **7**, (2022).
77. Puginier, C. *et al.* Phylogenomics reveals the evolutionary origins of lichenization in chlorophyte algae. *Nat. Commun.* **15**, 4452 (2024).
78. Nilsson, M., Rasmussen, U. & Bergman, B. Cyanobacterial chemotaxis to extracts of host and nonhost plants. *FEMS Microbiol. Ecol.* **55**, 382–390 (2006).
79. Duggan, P. S., Thiel, T. & Adams, D. G. Symbiosis between the cyanobacterium *Nostoc* and the liverwort *Blasia* requires a CheR-type MCP methyltransferase. *Symbiosis* **59**, 111–120 (2013).
80. Chen, M. Y. *et al.* Comparative genomics reveals insights into cyanobacterial evolution and habitat adaptation. *ISME J.* **15**, 211–227 (2021).
81. Green, E. R. & Mecsas, J. Bacterial secretion systems: An overview. *Microbiol. Spectr.* **4**, (2016).
82. Deleplaire, P. Type I secretion in gram-negative bacteria. *Biochim. Biophys. Acta Mol. Cell Res.* **1694**, 149–161 (2004).
83. Lucke, M., Correa, M. G. & Levy, A. The role of secretion systems, effectors, and secondary metabolites of beneficial rhizobacteria in interactions with plants and microbes. *Front. Plant Sci.* **11**, (2020).
84. Kloosterman, A. M. *et al.* Expansion of RiPP biosynthetic space through integration of pan-genomics and machine learning uncovers a novel class of lantibiotics. *PLoS Biol.* **18**, (2020).
85. Grimm, M. *et al.* The Lichens' microbiota, still a mystery? *Front. Microbiol.* **12**, (2021).
86. Bell-Doyon, P., Laroche, J., Saltonstall, K. & Villarreal Aguilar, J. C. Specialized bacteriome uncovered in the coralloid roots of the epiphytic gymnosperm, *Zamia pseudoparasitica*. *Environ. DNA* **2**, 418–428 (2020).
87. Gagunashvili, A. N. & Andr6sson, 6. S. Distinctive characters of *Nostoc* genomes in cyanolichens. *BMC Genom.* **19**, (2018).
88. Leavitt, S. D. *et al.* Fungal specificity and selectivity for algae play a major role in determining lichen partnerships across diverse ecogeographic regions in the lichen-forming family Parmeliaceae (Ascomycota). *Mol. Ecol.* **24**, 3779–3797 (2015).
89. J6riado, I., Kaasalainen, U., Jylh6, M. & Rikkinen, J. Relationships between mycobiont identity, photobiont specificity and ecological preferences in the lichen genus *Peltigera* (Ascomycota) in Estonia (northeastern Europe). *Fungal Ecol.* **39**, 45–54 (2019).
90. Bouchard, R. *et al.* Contrasting bacteriome of the hornwort *Leiosporoceros dussii* in two nearby sites with emphasis on the hornwort-cyanobacterial symbiosis. *Symbiosis* **81**, 39–52 (2020).

Acknowledgements

This research was supported by EMBL (European Molecular Biology Laboratory) core funds. We would also like to thank the reviewers and editor for their insightful comments and suggestions during the review process, which significantly improved the manuscript.

Author contributions

E.S.C. and R.D.F. designed the study. E.S.C. conducted bioinformatics analysis. S.S. performed clustering of biosynthetic gene clusters predictions. E.S.C., R.D.F., M.L.B., and N.G. drafted and edited manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-69215-8>.

Correspondence and requests for materials should be addressed to R.D.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024