Model cyanobacterial consortia reveal a consistent core microbiome independent of inoculation source or cyanobacterial host species

Andreja Kust[1,2], Jackie Zorz[1,2], Catalina Cruañas Paniker[2,#], Keith Bouma-Gregson[3], Netravathi Krishnappa[2], Jillian F. Banfield[1,2,4,5], Spencer Diamond[1,2,*]

[1]Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, CA, USA.
[2]Innovative Genomics Institute, University of California, Berkeley, CA, USA.
[3]U.S. Geological Survey, California Water Science Center, Sacramento, CA, USA.
[4]Department of Environmental Science, Policy and Management, University of California, Berkeley, CA, USA.
[5]School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia.

[#]Current Address: Department of Life Sciences, Imperial College London, London, United Kingdom.

***Corresponding author:**
Spencer Diamond
Innovative Genomics Institute
University of California - Berkeley
Innovative Genomics Institute Building
2151 Berkeley Way, Berkeley, CA 94720
Phone: (818) 268-5776
Email: sdiamond@berkeley.edu

Abstract

Cyanobacteria are central to biogeochemical cycling, climate change, and eutrophication. While they readily develop associations with environmental microorganisms, the question of whether they consistently recruit specific microbiomes remains unresolved. Here, we established *in vitro* cyanobacterial consortia by inoculating five different cyanobacterial strains with microbiomes from three freshwater environments to determine if similar non-cyanobacterial organisms were recruited into stable cyanobacteria-based communities. Using amplicon and genome-resolved metagenomic methods we determined that stable cyanobacterial communities largely share the same taxa, regardless of the host cyanobacterial species or inoculation source. Specifically, we identified a 25 species core microbiome, which was significantly enriched in functions for micronutrient biosynthesis, metabolite transport/exchange, and anoxygenic photosynthesis. Furthermore, core species were enriched in putative mobile genetic elements, including two large (> 100 kb) circularized and curated plasmids, encoding functions that may support cyanobacteria-heterotroph symbioses. The consistency observed across these systems indicate that cyanobacterial consortia maintained under similar conditions converge to a predictable structure, and support a core microbiome with taxonomic and functional similarity to core microbiomes of other phototroph-heterotroph symbiotic assemblages.

Introduction

Cyanobacteria, widespread bacteria that perform oxygenic photosynthesis, play globally vital roles in biogeochemical cycling[1], climate change[2], and eutrophication[3]. Their prolific primary productivity accounts for 20-30% of CO_2[4] and up to 50% of nitrogen fixation globally[5,65,6]. While they are critically important primary producers, cyanobacteria can also exert negative impacts on freshwater ecosystems by forming dense seasonal blooms that are currently increasing in both intensity and duration[3,7]. Yet, cyanobacteria do not exist in isolation. They are part of complex communities, living alongside a diverse array of microorganisms, and the interactions within these communities are crucial in shaping the overall environmental impact of cyanobacterial assemblages.

While the provision of fixed carbon is a well understood community trophic interaction of cyanobacteria, other interactions within their communities are less well understood and likely far more complex[8]. The microenvironment surrounding cyanobacterial cells, dubbed “the phycosphere'", is a term that also includes the consortia surrounding eukaryotic algae and is analogous to the area of enhanced microbial activity in the plant rhizosphere[9–11]. Phycosphere microorganisms likely act as significant fixed carbon sinks. They may participate in cooperative interactions with the cyanobacterial host, which involve activities like supplying macronutrients through remineralization, delivering regenerated nitrogen and phosphorus under low-nutrient conditions, increasing the bioavailability of micronutrients (e.g., Fe), and aiding in oxidative stress reduction. Conversely, phycosphere microorganisms can also engage in competitive interactions, such as competing for inorganic nutrients with the cyanobacterial host[11,1213,14].

Recent work has generated considerable insight into the phylogenetic diversity and function of cyanobacteria and their co-occurring microorganisms through cultivation-independent analysis[15–18]. Despite these advancements, a comprehensive understanding of the dynamics of cyanobacterial community assembly, the factors influencing these dynamics, and the precise identity and roles of core members within these communities is still evolving[19 20; 21 22–24]. Accordingly, it is increasingly important to conduct systematic and controlled *in vitro* studies focused on cyanobacteria in a community setting. Such research is vital for closing knowledge

gaps in how these communities form, defining their interactions with environmental microorganisms, and determining whether there is a consistent core microbiome across cyanobacterial communities that may be linked to conserved metabolic interactions and functionalities.

In this study we addressed these knowledge gaps in cyanobacterial community biology by conducting a systematic *in vitro* cultivation experiment where five diverse model cyanobacterial strains were exposed to bacterial inocula sourced from three distinct natural freshwater ecosystems. Through successive passaging of these *in vitro* systems, we established stable and ecologically relevant cyanobacterial consortia. Interrogation of these systems by 16S rRNA sequencing during their development, and genome-resolved metagenomics at the stable end point showed that despite the use of species rich and phylogenetically distinct source inocula, the communities that ultimately formed were markedly reduced in diversity, converged on a predictable shared structure with a notable degree of similarity, and were enriched in microbial taxa common to both freshwater cyanobacterial communities and other phototropic symbiotic assemblages[11,25 26;20]**.** We identified a set of 25 microbial species that constituted a core microbiome across our *in vitro* systems. Core microbiome taxonomy was highly enriched in bacterial lineages also known to be present in the plant rhizosphere[27–31], and we provide evidence that core microbiome species were independently and reproducibly selected into our communities from geographically distinct source inocula. Functional analysis of core microbiome genomes identified a significant enrichment of genes for cofactor metabolism, membrane transport, and anoxygenic photosynthesis relative to non-core microbiome species. We also observed that core microbiome members were significantly enriched in putative mobile genetic elements (MGEs) relative to non-core members, and that these elements encoded many of the functions enriched in the core microbiome as well as functions that have previously been identified to support phototroph-heterotroph symbioses[32].

Results

Inoculation of model cyanobacterial strains with bacteria from diverse freshwater sources established representative and stable *in vitro* consortia.

Water samples were collected from three freshwater bodies that are typical cyanobacterial habitats located in Discovery Bay, California ("Discovery Bay"), Branscomb, California ("Eel River"), and Berkeley, California ("Strawberry Creek") (**Fig. 1A and Table S1**). Metagenomic analysis showed that each freshwater body had an independent and predominantly consistent taxonomic composition (**Fig. S1**), represented by common freshwater bacterial phyla [33,34]. The concentrated microbial fractions from the individual habitats were used to inoculate four axenic model cyanobacterial host species (five distinct strains) in triplicates: *Synechococcus elongatus* PCC 7942 (hereinafter "7942"), *Synechococcus elongatus* UTEX 3055 (hereinafter "3055"), *Anabaena sp.* PCC 7120 (hereinafter "A7120"), *Synechocystis sp.* PCC 6803 (hereinafter "6803"), and *Leptolyngbya sp.* BL0902 (hereinafter "L0902") resulting in 180 initial *in vitro* consortia (See Methods; **Fig. 1A,** and **Table S2**). Following regular culture passaging (weekly and bi-weekly) for 12-weeks 120 consortia survived.

We found that the species of host cyanobacterium (p-value = $3.412e^{-9}$; logistic regression), the source location for microbial inocula (p-value = $1.236e^{-4}$; logistic regression), and the interaction between host strain and passage rate (p-value = 0.0122; logistic regression), but not passage rate alone (p-value = 0.203; logistic regression), were predictive of culture survival through the 12-week passaging period. Overall, we observed that 6803, L0902, and A7120 had the highest survival rates while 3055 and 7942 had the lowest, and that sourcing microbial inocula from the

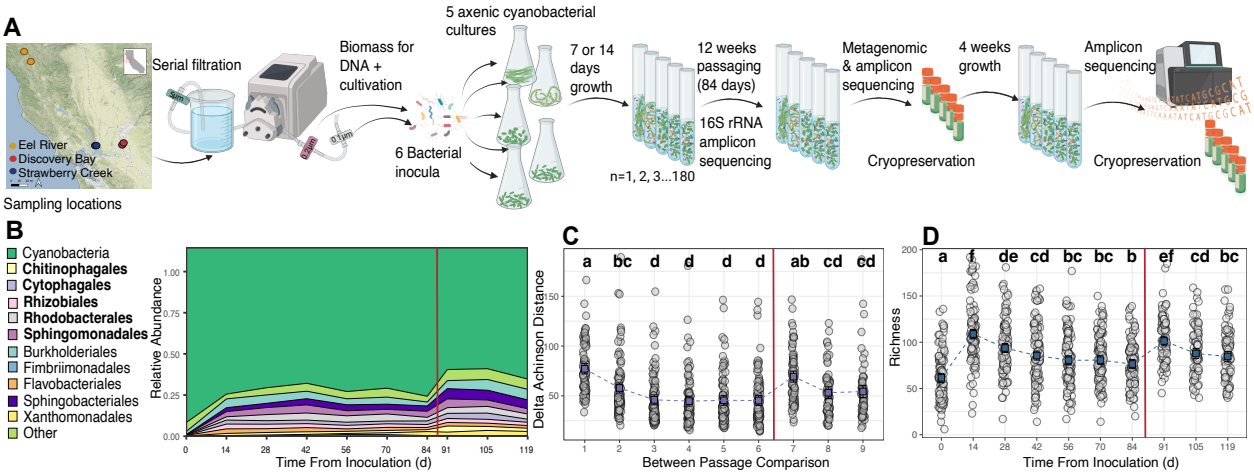

Figure 1 | Assessment of *in vitro* community stability over the 119 d passaging experiment. (A) Experimental design for cyanobacterial *in vitro* communities cultivation. Enrichments were passaged every 7 or 14 days over 12 weeks, totaling 6 to 12 passages. At each passage, samples were taken for microscopic analysis, optical density (OD) measurements, and amplicon sequencing. After the final 12-week passage, samples were additionally collected for metagenomic sequencing and cryopreservation. The frozen consortia were then regrown, passaged, and harvested as before. (Figure designed using BioRender). **(B)** Order-level community composition based on 16S rRNA amplicon sequencing at 14 d intervals over 119 d. Taxa found in the core microbiome are indicated in bold. In all plots the vertical red line separates pre-cryopreservation passaging (left) and post-cryopreservation passaging (right). **(C)** Change in Aitchinson distance between previous passage time points for *in vitro* communities at 14 d intervals over 119 d. Gray dots give change for each community compared to itself in the preceding time point, purple squares indicate estimated marginal means, and error bars represent 95% CI of marginal mean. Statistical significance between comparisons estimated using linear mixed effects models with comparisons not sharing the same letter being significantly different (FDR ≤ 0.05). **(D)** Observed richness of *in vitro* communities at 14 d intervals over 119 d. Gray dots give values for each community at a given time point, blue squares indicate estimated marginal means, and error bars represent 95% CI of marginal mean. Statistical significance between time points estimated using linear mixed effects models with time points not sharing the same letter being significantly different (FDR ≤ 0.05).

Eel River yielded higher survival rates than Discovery Bay or Strawberry Creek (**Fig. S2 and Table S2**).

We evaluated changes in community diversity, taxonomic composition, and stability through longitudinal analysis of 16S rRNA sequencing data collected from each passaging time point. We found significant changes in alpha diversity, beta diversity, and community order-level taxonomy within the initial fourteen day period of the experiment. Subsequently, communities stabilized, and after 28 days almost no significant changes were observed in diversity metrics or taxonomic composition (**Fig. 1B-D, Fig. S3, and Table S3-S6**). Additionally, we found that cryopreservation and subsequent revival and passaging of communities had minimal significant impact on community diversity metrics or composition (**Fig. 1B-D and Fig. S3, Table S3-S6**). Ultimately, the communities that formed and stabilized were highly similar in both taxonomic representation and relative abundance to cyanobacterial communities that have been sampled and isolated from natural environments[18,35,36].

Genomic resolution enabled comprehensive characterization of *in vitro* consortia

Cultures that survived the initial 12-week passaging were submitted for whole-genome shotgun sequencing **(Table S2**). In total 134 samples were sequenced (n = 14 Source samples; n = 120 *in vitro* communities) yielding approximately 1.6 terabase pairs (Tbp) of total sequence data. Assembly, genome binning, and genome de-replication at 95% average nucleotide identity (ANI) yielded a final non-redundant set of 537 species-level genomes (See Methods; **Table S7**), with 321 (59.8 %) satisfying MIMAG criteria as high-quality draft genomes and 216 (40.2 %) satisfying MIMAG criteria as medium-quality draft genomes [37]. The recovered genomes spanned 17 distinct phyla (**Fig. 2A**) with prominent representation of Bacteroidetes (n = 149 genomes) and Proteobacteria (n = 133 Alphaproteobacteria and n = 106 Gammaproteobacteria). The majority of these genomes (84.2 %) have no species representative in genomic databases, and 59 and 8 of these genomes were novel at the genus and family level respectively (**Table S7**). These genomes effectively represented our *in vitro* communities with 86.2 % of reads from each sample mapping back to genomes on average (IQR = 83.3 - 90.3 %) (**Table S2**). The representation of environmental source genomes in our set was comparatively lower with 9.3 % of reads from each sample mapping back to genomes on average (IQR = 4.5 - 12.8 %**)**.

Selection of specific taxa during stabilization of *in vitro* consortia

Due to lower genomic recovery from source samples we employed assembly independent marker gene analysis to directly compare environmental source samples to *in vitro* community samples (**Table S8**, See Methods). We found that *in vitro* consortia had significantly lower levels of alpha-diversity relative to environmental source samples supporting that a substantial selective event occurred during *in vitro* consortia formation (**Fig. S4A and S4B**). Additionally, the alpha-diversity of environmental samples was not predictive of the alpha-diversity observed for *in vitro* consortia (**Fig. S4C-F**). Thus, we hypothesized that specific taxa are more likely than others to be selected into cyanobacterial communities and assessed this using a permutation-based analysis (See Methods). Out of 428 order-level taxonomic groups detected by marker gene analysis, 22 orders were significantly underrepresented and 59 orders were significantly overrepresented in our *in vitro* communities (**Fig. 2B and Table S9**). Orders within Proteobacteria and Bacteroidetes were largely overrepresented in communities and orders within CPR (Patescibacteria) and Omnitrophica were underrepresented. This data supports the hypothesis that a common set of specific taxa are selected into cyanobacterial communities despite the use of dissimilar source inocula and cyanobacterial host strains. Furthermore, this provides an explanation for why species rich environments, such as Strawberry Creek, do not ultimately form the richest communities as CPR contributed to a larger portion of that environment's richness (**Fig. S1 and S4C-D**).

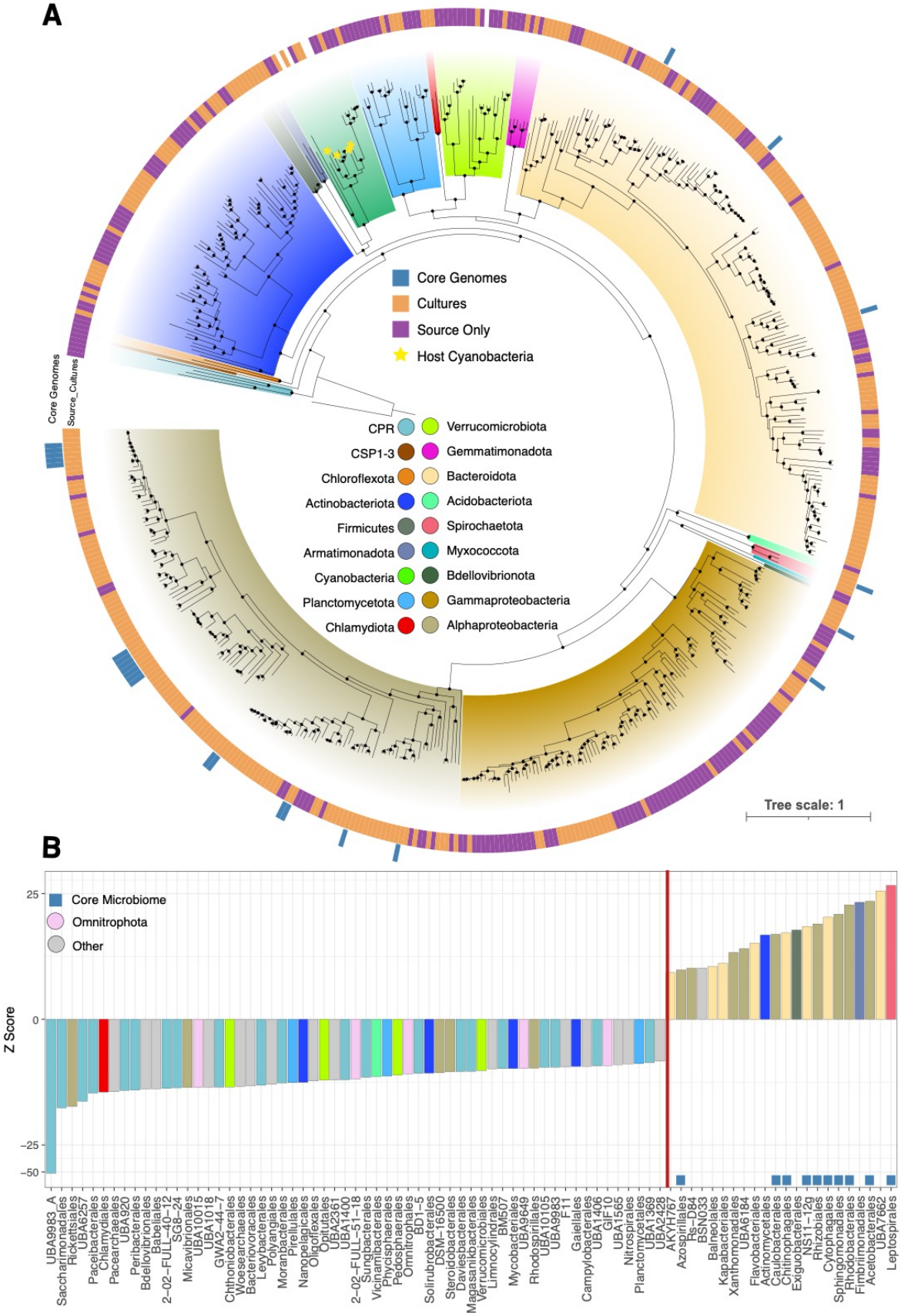
A
Core Genomes
Cultures
Source Only
Host Cyanobacteria
Core Genomes
Source_Cultures
CPR
CSP1-3
Chloroflexota
Actinobacteriota
Firmicutes
Armatimonadota
Cyanobacteria
Planctomycetota
Chlamydiota
Verrucomicrobiota
Gemmatimonadota
Bacteroidota
Acidobacteriota
Spirochaetota
Myxococcota
Bdellovibrionota
Gammaproteobacteria
Alphaproteobacteria
Tree scale: 1
B
Core Microbiome
Omnitrophota
Other
Z Score
25
0
−25
−50
UBA9983_A
Saccharimonadales
Rickettsiales
UBA6257
Paceibacterales
Chlamydiales
Pacearchaeales
UBA920
Peribacterales
Bdellovibrionales
Babeliales
2−02−FULL−40−12
SG8−24
Micavibrionales
UBA10015
UBA1018
GWA2−44−7
Chthoniobacterales
Woesearchaeales
Bacteriovoracales
Levybacterales
Polyangiales
Moranbacterales
Pirellulales
Nanopelagicales
Oligoflexales
Opitutales
UBA2361
UBA1400
2−02−FULL−51−18
Sungbacterales
Vicinamibacterales
Phycisphaerales
Pedosphaerales
Omnitrophales
BD1−5
Solirubrobacterales
DSM−16500
Steroidobacterales
Daviesbacterales
Magasanikbacterales
Verrucomicrobiales
Limnocylindrales
BM507
Mycobacteriales
UBA9649
Rhodospirillales
UBA10105
UBA9983
F11
Gaiellales
Campylobacterales
UBA1406
GIF10
UBA1565
Nitrospirales
Planctomycetales
UBA1369
UBA2428
AKYH767
Azospirillales
Rs−D84
BSN033
Balneolales
Kapabacteriales
Xanthomonadales
UBA6184
Flavobacteriales
Actinomycetales
Caulobacterales
Chitinophagales
Exiguobacterales
NS11−12g
Rhizobiales
Cytophagales
Sphingomonadales
Rhodobacterales
Fimbriimonadales
Acetobacterales
UBA7662
Leptospirales

Figure 2 | Maximum Likelihood Phylogenetic Tree and Enrichment of Microbial Orders. **(A)** The phylogenetic tree was constructed using a concatenated alignment of 120 bacterial-specific marker genes. An archaeal reference genome (GCF_000213215.1) served as an outgroup for rooting the tree. Nodes marked with black dots represent a high bootstrap support (≥ 95%, calculated using ufboot with n = 1000 iterations). The outer ring highlights core microbiome genomes (in blue), while the inner ring differentiates genomes observed in *in vitro* cultures (orange) and those exclusive to source samples (purple). Branch tips with cyanobacterial host species are denoted by yellow stars, and have no outer ring colors. **(B)** Microbial orders showing significant over or under enrichment when comparing ribosomal protein L6 marker gene OTU (rpL6 OTU) diversity between true *in vitro* community compositions and randomly sampled *in vitro* communities (n = 10,000 permutations) where all observed rpL6 OTUs were sampled. The Z-score indicates the level of enrichment of an order in observed communities relative to communities generated from random sampling. Only orders with statistically significant enrichment values are shown (FDR ≤ 0.05; permutation test). Orders positioned on the left of the red line are underrepresented in true consortia, while those on the right of the line are over-represented in true consortia. Blue squares identify taxa found in the core microbiome. Bars are colored by phylum as indicated in the key of Fig. 2A with additional colors added and noted in the key of Fig. 2B for minor phyla (Other) and Omnitrophica. Also see **Tables S8 and S9.**

Inoculation source and cyanobacterial host strain impact community composition
Genome resolved analysis showed that the final communities obtained after 84 d of growth and passaging were dominated by a cyanobacterial host strain with co-occurring Proteobacteria and Bacteroidetes species, as well as the appearance of other rarer phyla including Spirochaetota, Actinobacteriota, and Armatimonadota **(Fig. 3A, 3B, and Table S10**). When the compositional dissimilarity of our communities was evaluated in isolation (e.g. without source samples) we found that the inoculum source ($R^2 = 0.230$, $p < 0.0001$; PERMANOVA), host cyanobacterial strain ($R^2 = 0.114$, $p < 0.0001$; PERMANOVA), and the interaction between the inoculation source and host cyanobacterial strain ($R^2 = 0.075$, $p < 0.0001$; PERMANOVA) were statistically associated with variation in community composition (**Fig. S5**). Despite the discernible impact of these factors on *in vitro* community composition we note that *in vitro* communities broadly converge to a taxonomically similar state with the enrichment of specific major taxonomic groups relative to the source environment samples (**Fig. 2B, 3A, and 3B**). In direct comparisons of compositional dissimilarity between source environment and *in vitro* communities we observe strong separation between source and *in vitro* community samples on the primary axis of variation, and clustering of *in vitro* community samples together with limited separation on the secondary axis of variation (**Fig. S5, Table S11**).

A core set of species was identified across model cyanobacterial communities.
The shared taxonomic composition and developmental trajectories (**Fig. 1B, 3A, 3B, and Fig. S3**) of our *in vitro* cyanobacterial communities encouraged us to identify microbial species that potentially formed a core microbiome. As there is no consensus definition of a core microbiome [38,39]we employed an operational definition and identified 25 microbial species that satisfied our criteria for core microbiome organisms (See Methods). Hereinafter we refer to non-core species detected in cultures as auxiliary species and species only detected in environmental samples as source species (**Fig. 2A, 3C-E, and Table S7**). Species from the core microbiome on average comprised 1.0 ± 1.3 % (Range: 0 - 7.3 %) of relative community abundance. However, their average relative abundance when considering only the non-cyanobacterial fraction of communities was 14.7 ± 19.4 % (Range: 0 - 82.4%) (**Fig. 3C**), with individual core species being detectable in 18.5 % - 68.5 % of all samples (**Table S7**). In support of our hypothesis, core microbiome organisms were almost exclusively (24/25 species) part of taxonomic orders that showed significant enrichment across all cyanobacterial communities (**Fig. 2B**) with the most numerous being Rhizobiales (n = 7 species), Rhodobacterales (n = 5 species) and Sphingomonadales (n = 5 species) **(Fig. 3D)**. We also observed that core microbiome species were closely taxonomically clustered within their respective orders (**Fig. 2A**). Core microbiome species had significantly lower phylogenetic distances between their genomes on average relative to auxiliary or source inoculum species (**Fig. 3E**). For 18 core microbiome species no species-representative genome exists in public databases (**Table S7**), and thus the recovery of these genomes represents a significant expansion of genomic and functional information for organisms that may occupy a generalized core niche in cyanobacterial communities.

Identical species were selected into communities from geographically distinct sites
To investigate whether the same core and auxiliary species across our in vitro communities were naturally selected from different environmental sources rather than being spread through cross-contamination, we analyzed population level single nucleotide variants (SNVs) for individual genomes between samples. We employed inStrain (See Methods) to compare these SNVs and conducted 3088 two-sample comparisons across 133 genomes, which included comparisons of 13 core microbiome genomes. No pairwise average nucleotide identity (popANI) comparisons indicated identical strain sharing among samples with different source inocula (n = 0/176 comparisons). Additionally, we observed that the SNV profiles were significantly influenced by the

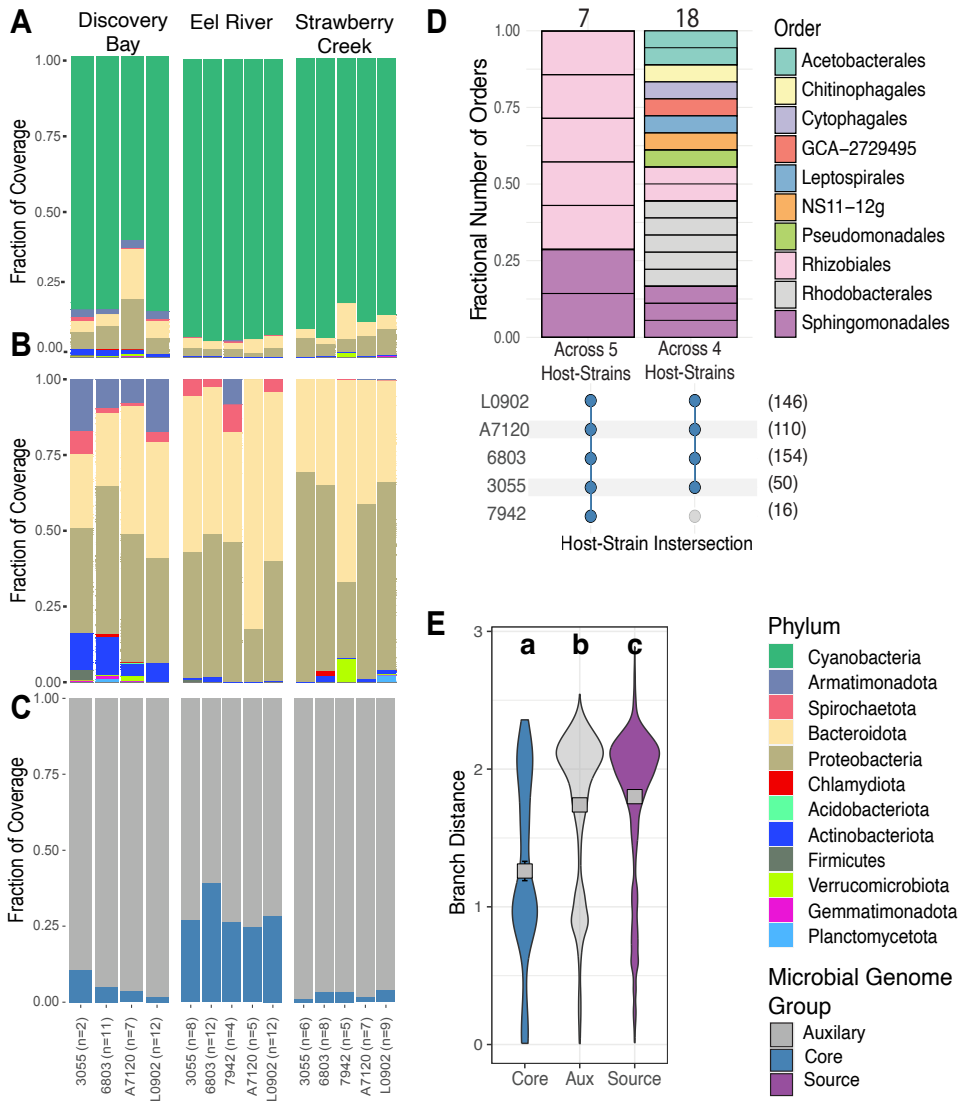

Figure 3 | Microbial community composition and core microbiome. A. Phylum level relative abundance of *in vitro* communities based on mapping samples to a database of 537 species-level genomes. Sample compositions are averaged within cyanobacterial strain for each environmental inoculum. **B.** Phylum level taxonomic composition of i*n vitro* communities with cyanobacteria removed. **C.** Composition of *in vitro* communities with cyanobacteria removed depicting the relative fraction assigned to core and auxiliary microbiome genomes. **D.** Cyanobacterial host association and order level taxonomy of the 25 genomes identified as the core microbiome members. Number above colored bars indicate the total number of genomes. Colored bars indicate the order level taxonomy of genomes. Blue upset plots below each bar shows the set of cyanobacterial host strains where genomes were detected in at least 3 cultures where that strain was the host. The numbers in parentheses indicate the total number of species that were observed in at least 3 cultures with each cyanobacterial host. All core microbiome members were detected in at least 3 cultures from at least 2 independent source inocula. **E.** Distributions of pairwise all-vs-all phylogenetic branch distance comparisons between genomes within core (n = 300 comparisons), auxiliary (n = 43,017 comparisons), and source (n = 19,900 comparisons) groupings. Gray squares in the plot indicate the mean, and error bars indicate the 95% confidence interval. Statistical differences between groups are indicated by letters, and groups not sharing the same letter are significantly different (FDR ≤ 0.05; Pairwise Wilcoxon test).

identity of cyanobacterial host and source inoculum (**Fig. S6 and Tables S12-S13**). These findings support the hypothesis that specific selection pressures inherent to species, rather than cross-contamination, are responsible for the incorporation of identical species into communities from geographically disparate water sources.

Enriched metabolic functions in the core cyanobacterial microbiome

To identify metabolic functions associated with members of the cyanobacterial core microbiome we quantified the enrichment of KEGG orthology groups (KOs) identified in core microbiome genomes (n = 25) relative to their frequency in non-core genomes (n = 512). To account for a strong taxonomic bias in the core microbiome we also repeated our analysis using a set of non-core genomes (n = 366) that was constrained to the three phyla in the core microbiome (Bacteroidota, Proteobacteria, and Spirocheaotota). We identified **1**073 and 1065 KOs with significant differential enrichment (FDR ≤ 0.05; Fisher test) in core microbiome genomes relative to the full and constrained non-core genomes, respectively (**Tables S14-S16**). Functional category analysis of these differentially enriched KO sets identified a significant over representation of KOs involved in cofactor metabolism (Fisher test; OR_{All} = 1.6, FDR_{All} = 0.007; $OR_{Constrained}$ = 1.9, $FDR_{Constrained}$ = 0.0002) and membrane transport (Fisher test; OR_{All} = 1.7, FDR_{All} = 1.9 e-6; $OR_{Constrained}$ = 1.7, $FDR_{Constrained}$ = 1.1e-5) in core microbiome genomes relative to non-core genomes (**Fig. 4A**). The results of the constrained analysis mirrored the full analysis and additionally identified KOs involved in signal transduction (Fisher test; $OR_{Constrained}$ = 1.7, $FDR_{Constrained}$ = 0.006**)** as significantly overrepresented in the core microbiome (**Fig. 4A**).

Within cofactor metabolism, 39 KOs in the porphyrin metabolic pathway were identified as enriched in core microbiome members (**Fig. 4B**). These pathways drive the energy intensive biosynthesis of several compounds including vitamin B12, chlorophylls, and hemes. We observed that core microbiome genomes were enriched in functions both at the initiation and terminal steps of these pathways. This included enrichment of 5-aminolevulinate synthase (K00643; Odds $Ratio_{All}$ = 9.9, FDR_{All} = 6.5 e-6; Fisher test), enabling the synthesis of 5-Aminolevulinate (ALA) from glycine. In contrast, non-core genomes were significantly more likely to have the alternate three-step pathway for ALA biosynthesis, starting from glutamate (K02492; Fisher test; OR_{All} = 0.08, FDR_{All} = 0.0001). The gene cobC (K02225), involved in vitamin B12 synthesis, was the KEGG ortholog with the strongest core microbiome association (Fisher test; OR_{All} = 24.6, FDR_{All} = 3.3 e-8), and most other enzymes in the vitamin B12 and bacteriochlorophyll synthesis pathways were also enriched in core microbiome genomes (**Fig. 4B and Table S15**).

We found that KOs encoding anoxygenic photosystem components were enriched in core microbiome species. Of the core microbiome members, 80% (20/25) contained anoxygenic photosystems in their genomes **(**K08929/K08928/K13991; **Fig. 5, Table S15-S17**, suggesting that photosynthesis is an important metabolic strategy of cyanobacteria-associated bacteria. Based on the predicted functions encoded in core microbiome genomes, the electron donors for these anoxygenic phototrophs likely include sulfide, hydrogen, or organic substrates. Only one member of the core microbiome, a *Georhizobium* species, encoded RuBisCO for carbon fixation.

Genes involved in the degradation (K06162-6) and transport (K02041-4) of phosphonate were found statistically more often in the core microbiome, providing an alternate source of phosphorus to the community. Core microbiome genomes were also enriched in KOs required for methylamine (K15228-9), nitrile (K01721/K20807), and urea (K01428-30) utilization, putatively expanding the sources of carbon and nitrogen available to cyanobacterial communities (**Fig. 5 and Table S15-S17**). Finally, many of the core microbiome species with the potential for methylamine degradation to formaldehyde, also encoded a downstream pathway to further oxidize formaldehyde to formate and CO_2.

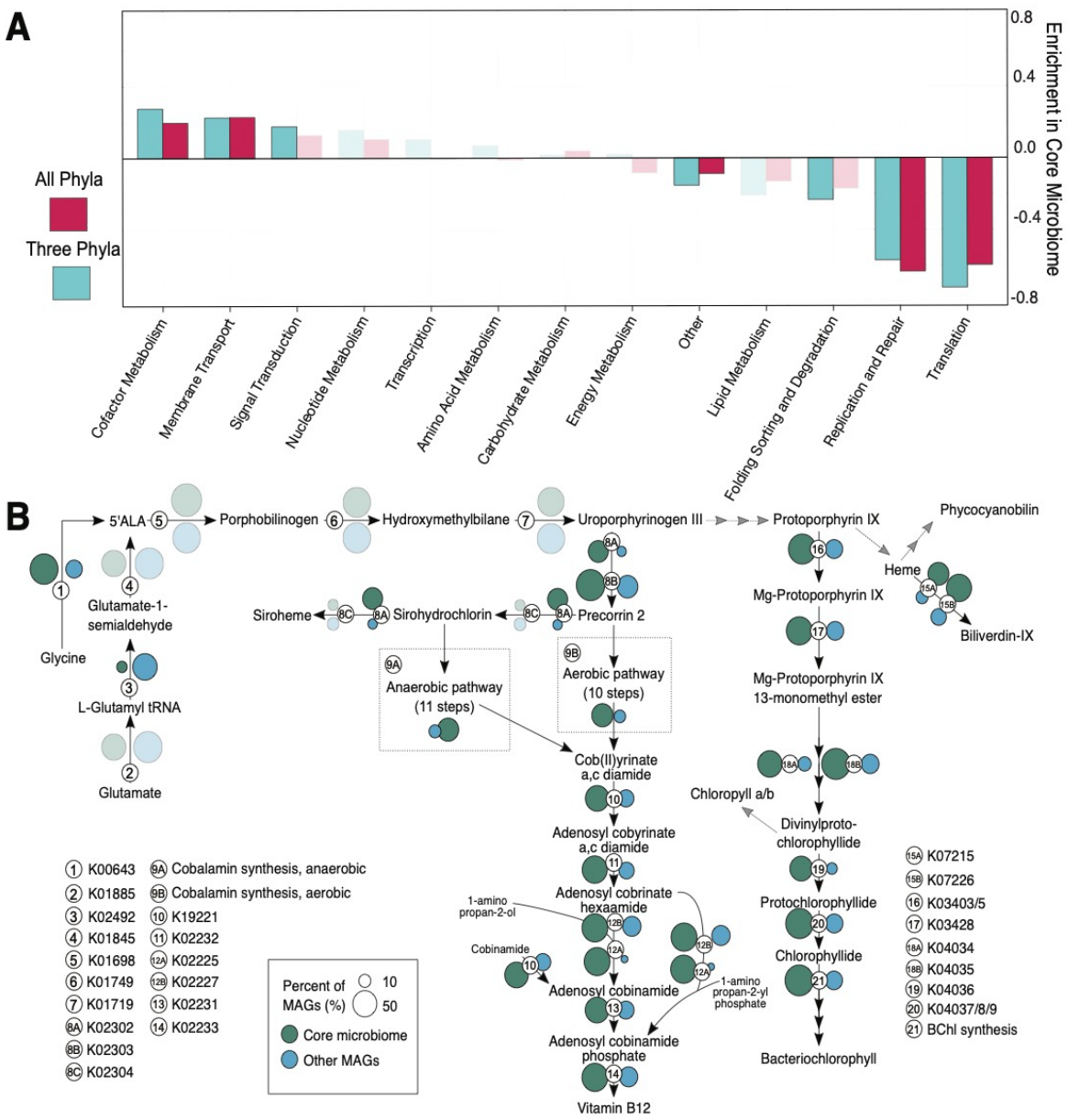

Figure 4 | Metabolic functions enriched in the core microbiome. A. Functional categories of statistically over or under enriched KEGG functions in the core microbiome. The enrichment odds-ratio for each functional category is shown for the set of KEGG functions with differential enrichment when the core microbiome genomes were compared to all 512 non-core genomes (red) and the 366 non-core genomes in the phyla Bacteroidota, Proteobacteria, and Spirocheaotota (blue). Bars are solid colors for categories with statistically significant over or under representation (FDR ≤ 0.05; Fisher test) and transparent if not significant. **B.** Metabolic reaction map for porphyrin compound biosynthetic pathways. Numbers indicate the KEGG KO associated with a reaction. This map shows enrichment data from the comparison of core genomes to all 512 non-core genomes. The size of colored circles indicates the percentage of MAGs in core (green; n = 25) or non-core (blue; n = 512) sets that encode the respective reaction. Circles are solid colors if a KEGG KO was detected as significantly over or under enriched in the core microbiome (FDR ≤ 0.05; Fisher test), and transparent if not significant. Also see **Tables S14-S16.**

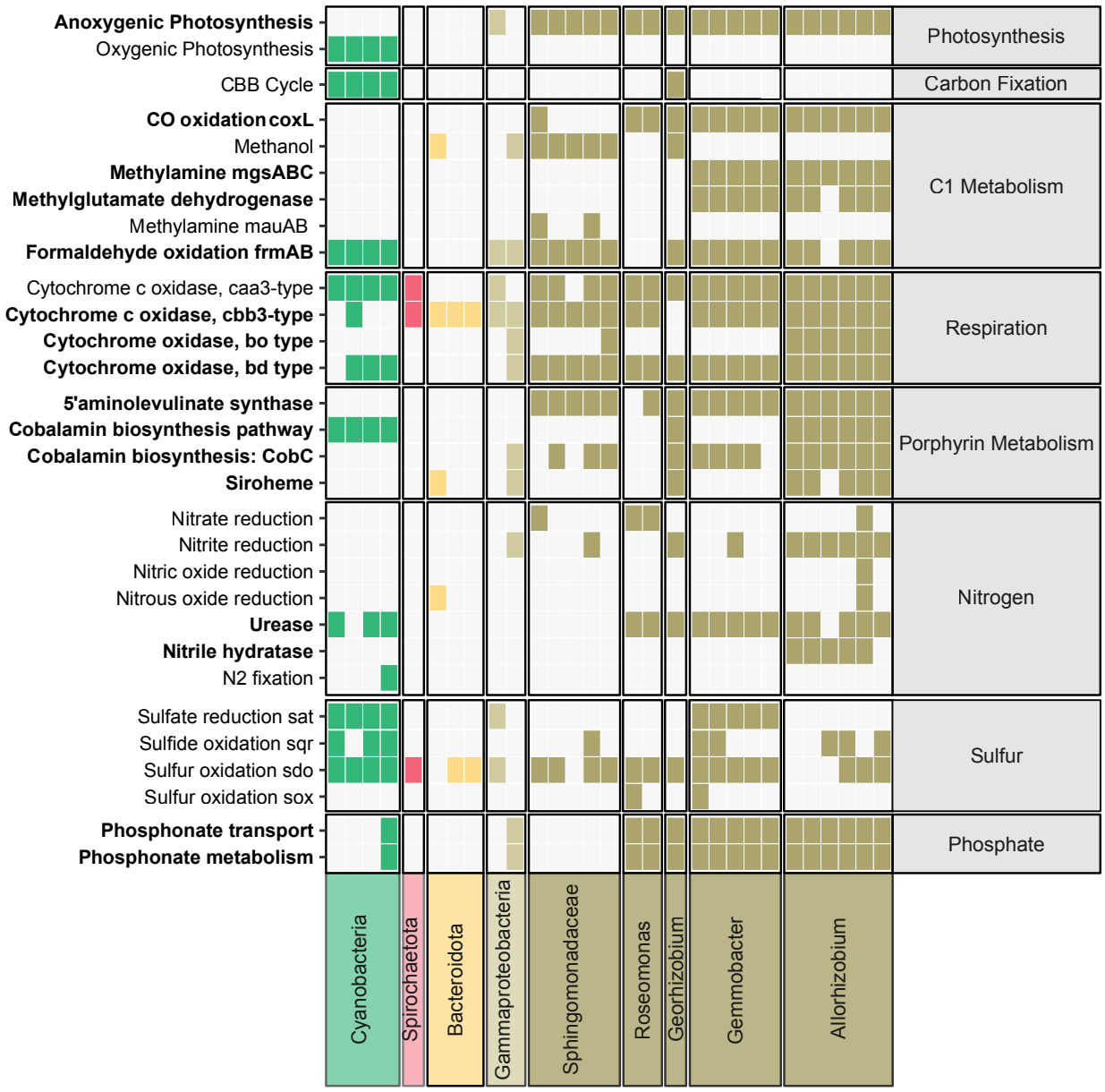

Figure 5 | Comparison of select metabolic functions in cyanobacterial hosts and core microbiome species. Encoded metabolic potential for specific biogeochemical turnover reactions identified in cyanobacterial host species genomes and the 25 core microbiome genomes. Specific metabolic functions (left) are separated by higher level metabolic categories (right). Functions highlighted in bold had at least one associated KO with significant over enrichment in the core microbiome relative to all 512 non-core genomes. Each column represents an individual genome in the core microbiome, genomes are grouped at different taxonomic levels (phylum, class, and order), and colored by phylum level taxonomy: Cyanobacteria (green), Spirochaetota (red), Gammaproteobacteria (light brown), and Alphaproteobacteria (brown). Identification of the potential for a specific function in each genome is indicated by a colored box, and boxes share phylum color. Criteria for identifying specific functions can be found in **Table S17**.

Core microbiome genomes have significantly more predicted MGEs than non-core microbiome genomes.
We searched for MGEs within our dataset to test the hypothesis that mobile genetic elements (MGEs) facilitate symbiotic relationships in Cyanobacterial communities, which has been previously observed for bacteria participating in plant-microbial symbioses [32]. Utilizing methods that assessed sequence composition and gene content, we identified 850 contigs across all genome bins as likely plasmid elements (**Table S18**). Most were linear fragments (98%), prompting us to treat them as possible MGEs that may or may not be integrated into bacterial chromosomes. 17 of the 25 core microbiome genomes yielded 198 contigs predicted to be MGEs, including one 12 kb circular element (**Table S18**). Core microbiome bins had a significantly higher proportion of contigs predicted to be MGEs than bins for auxiliary organisms (Dunn Test = 3.7, p = 0.0003) and genomes detected only in source environment samples (Dunn Test = 6.5, p = 0) (**Fig 6A**). When we aligned the predicted MGEs from the core microbiome to the PLSDB database, we discovered that 6 contigs from 3 core microbiome bins for Rhizobiales and Sphingomonadales were closely related (mash dist ≤ 0.1) to known plasmid sequences from Rhizobiales and Sphingomonadales isolates (**Table S19**).

Recognizing the challenge in determining the completeness or genomic context of plasmid fragments without circularization, we searched for, identified, and curated two additonal large (>130 kb) circular elements each linked to a *Gemmobacter* genome within the core microbiome (**Fig. 6B, Fig. S7**). These elements encoded gene functions previously recognized as enriched within core microbiome genomes (**Table S15 and S20**). Notably, they included gene clusters responsible for the metabolism of amides and aromatics, as well as the transport of amino acids and other small organic molecules. One of these circular elements harbored a substantial hydrogenase gene cluster, a feature rarely found on plasmids (**Fig. 6B**). The discovery of these complete and circular elements implies that some of the MGEs associated with core microbiome genomes may be authentic extrachromosomal plasmids.

We repeated our functional enrichment analysis comparing core microbiome genomes to non-core genomes and removed contigs predicted to be MGEs from all genomes and found 196 KOs were no longer significantly enriched in the core microbiome (**Fig. 6C and Table S15 and S16**). We infer that these 196 KOs were primarily MGE encoded. Functional category analysis of this KO set identified signal transduction, membrane transport, and carbohydrate metabolism as MGE enriched (**Fig. 6D**). Specific KEGG functions associated with predicted MGE contigs included nitrate transporters (NasST), cytochrome c complex IV genes (cyoABCD), the proton translocating transhydrogenase (pntAB), and genes involved in exo-polysaccharide biosynthesis (exoM). Other functions specific to plasmids included several subunits of the trb Type IV secretion and plasmid conjugation system (**FIg. 6D and Table S16**).

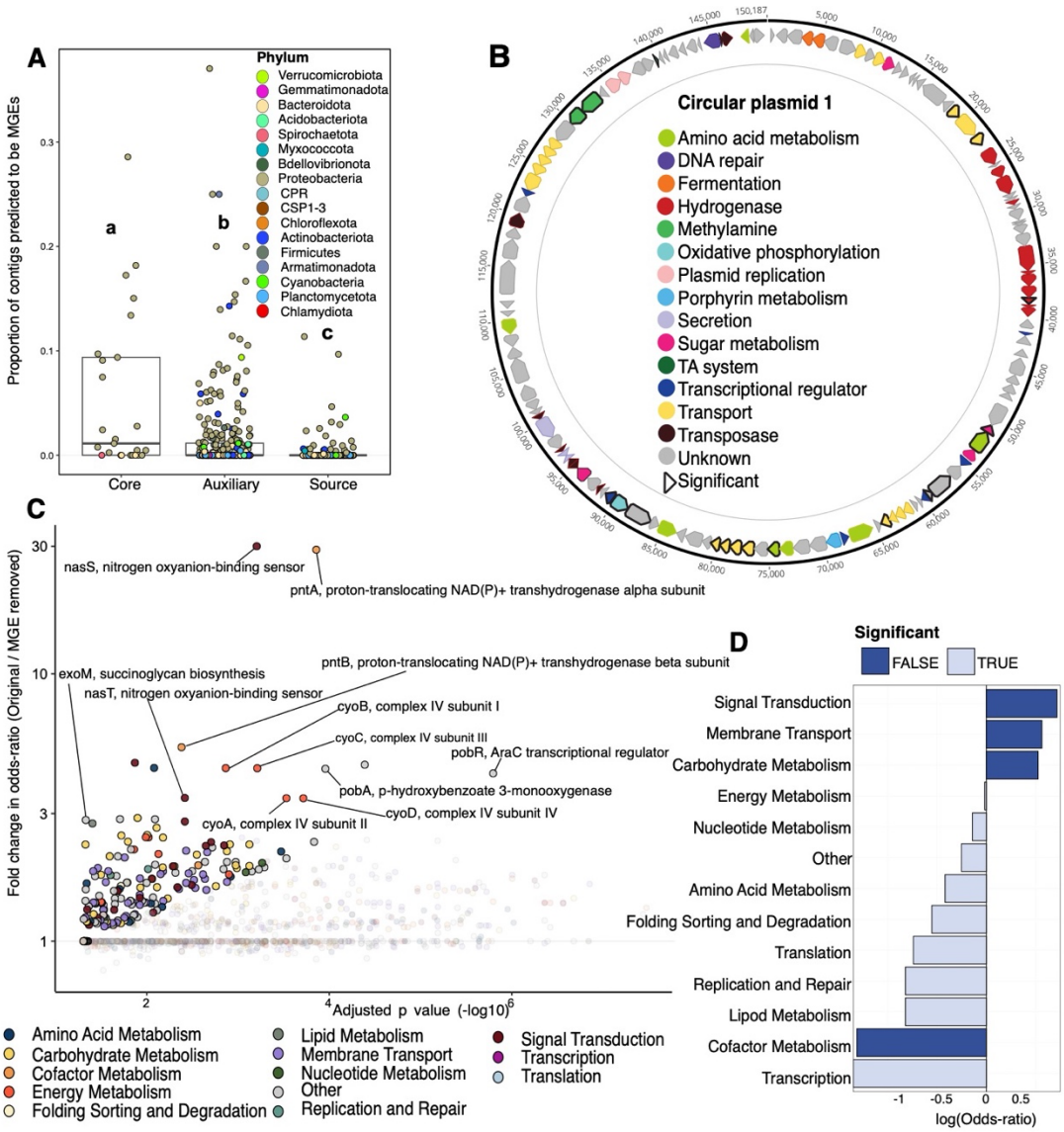

Figure 6 | Predicted mobile genetic element distribution and functional potential. (A) Proportion of contigs predicted to be mobile genetic elements in each genome of the core, auxiliary, and source microbiome. Points are colored by phylum level taxonomy of the genome. Box plots behind points indicate median, first, and third quartiles of data. Statistical differences between groups are indicated by letters, and groups not sharing the same letter are significantly different ($Statistic_{Core\text{-}v\text{-}Aux}$ = 3.7, FDR < 0.001; $Statistic_{Core\text{-}v\text{-}Source}$ = 6.5, FDR < 0.0001; $Statistic_{Aux\text{-}v\text{-}Source}$ = 6.6, FDR < 0.0001; Dunn Test). **(B)** One of two curated circular plasmid sequences co-binned with a Gemmobacter genome from the core microbiome (unicom2021_Dbay_Inner_A7120_BW_A_idba_concoct_16). Genes are colored by functional category and genes with a black outline are assigned to KEGG KOs that are statistically enriched in the core microbiome relative to all 512 non-core genomes. **(C)** Individual KEGG KOs that become non-significant when the enrichment analysis for core microbiome functions is repeated with predicted MGE contigs removed from all genomes. The x-axis indicates the $-\log_{10}$(FDR) for each KO in the original enrichment analysis. The y-axis indicates the fold change in the odds-ratio for a KO between the original analysis and the analysis conducted with all MGE contigs removed. Solid colored points indicate KOs that became non-significant in this analysis while transparent points indicate KOs that showed no change between the analyses. Points are colored by functional category. **(D)** Functional category analysis for the set of KEGG KOs that becomes non-significant when all MGE contigs are removed from genomes. Functional categories that are significantly over or underrepresented in this set of KOs are colored accordingly (FDR ≤ 0.05; Fisher test). Also see **Tables S15, S16 and S20.**

Discussion

In this study, we offer a detailed exploration of cyanobacterial community assembly and function in a controlled setting. Our systematic approach has allowed for a robust statistical analysis of the effects that a cyanobacterial host and disparate environmental microbiome exposures exert on community assembly dynamics, final composition, and function. Despite the discernible impact of the inoculum source and the cyanobacterial host species on the community composition, the resulting communities tended to stabilize into a taxonomically similar structure that closely resembles natural cyanobacterial consortia [15,16,40–42]. Furthermore, the microbial orders that were enriched during the community formation process encompassed taxa that are common in other known phototroph-heterotroph symbiotic assemblages including Rhizobiales, Sphingomonadales, and Rhodobacterales species [9,26,43]. The consistent presence of these taxa in both artificially constructed and natural cyanobacterial communities, as well as in symbiotic systems involving eukaryotic phototrophs, suggest that taxa and metabolic interdependencies are commonly shared in these associations.

We achieved statistically robust confirmation that core microbiome genomes are enriched in functions that facilitate metabolic collaboration. Notably, the enrichment of anoxygenic photosynthesis, which utilizes light wavelengths not absorbed by cyanobacteria [44], suggests a competitive edge for these organisms over taxonomically related organisms that do not have this capability. By absorbing light energy, the anoxygenic phototrophs in the core microbiome may offer photoprotection[44–46]**.** They also may contribute to the synthesis and distribution of energetically expensive porphyrin-based compounds like chlorophylls, hemes, and vitamin B12[47]. Exchanges of this type are supported by studies showing enhanced growth of cyanobacteria when supplemented with the porphyrin precursor 5-aminolevulinate (5-ALA) [48]. The core microbiome's enrichment in 5-ALA synthase, which enables a more energetically efficient route for 5-ALA production from glycine, may be an advantage in communities where porphyrin compounds are critical resources.

Core microbiome genomes show a significant enrichment in genes for accessing alternative phosphorous and nitrogen sources, which are limiting resources for cyanobacteria in natural systems [49–52]. This includes the breakdown and transport of phosphonates, which can make up to 25% of the dissolved organic phosphate in aquatic environments [53], and the degradation of methylamine, a byproduct of protein decomposition [54]. Recent studies have observed these interactions in phototroph-heterotroph assemblages, as bacteria capable of phosphonate metabolism can support Microcystis growth in environments where methylphosphonate is the sole phosphorus source [55]. Similarly, cross-feeding between Rhodobacteraceae, capable of methylamine degradation, and diatoms has been demonstrated in marine ecosystems [55,56]. Generally our findings support that porphyrin precursor sharing and enhancement of micronutrient bioavailability may be key functions of the cyanobacterial-associated core microbiome.

We found that the cyanobacterial core microbiome was enriched in MGEs, and by manually circularizing and curating two of these elements, we confirmed that some are indeed true extrachromosomal plasmids. While these MGEs lacked certain functional genes prevalent in the core microbiome, such as those in porphyrin metabolism, they were significantly enriched in genes involved in signal transduction and carbohydrate metabolism. This includes genes regulating nitrogen assimilation (nasST), exo-polysacharide biosynthesis (exoM), and plasmid conjugal transfer. Notably, plasmid encoded exo-polysacharide biosynthetic gene clusters are essential for establishing plant-bacteria associations in Rhizobiales, and exo-polysacharides mediating symbioses are regulated by environmental nitrogen and phosphorus sensing [57]**.** Thus,

the findings resonate with results of prior studies reporting that genes mediating plant-bacterial symbioses are encoded on MGEs including integrative conjugative elements (ICEs) and plasmids [32,58,59]. Our observation of a similar pattern of genomic organization in the cyanobacterial core microbiome, where potential symbiotic genes are located on MGEs, highlights a potentially universal strategy in symbiotic genome evolution.

Our findings offer insights into microbial community assembly dynamics using a system that is easy to establish, stable, environmentally relevant, and taxonomically realistic. The commonality between our communities and other higher phototropic symbioses suggests they can serve as proxies for studying broader ecological interactions of phototrophic oriented microbial assemblages. This work lays the foundation for a broader understanding of cyanobacterial microbiomes and underscores the importance of studying cyanobacteria as part of a consortium rather than in isolation. Finally, our establishment and biobanking of these well-defined and stable model systems provides future opportunities to decipher the intricacies of cyanobacteria-specific community interactions.

Materials and Methods

Environmental water collection

A set of three freshwater bodies that are typical cyanobacterial habitats with different physicochemical properties (**Fig. 1A** and **Table S1**) were selected for sampling: Discovery Bay (Discovery Bay, CA), characterized by low flow, eutrophication, and yearly cyanobacterial bloom occurrence, the Eel River (Branscomb, CA), an unpolluted river system with seasonal benthic cyanobacterial growth[1560], and Strawberry Creek (Berkeley, CA), a slow-flowing creek and typical cyanobacteria habitat. Roughly 4 L of water was collected from two sub-locations at each of the sampling sites, and pH, dissolved oxygen, and temperature of the water were analyzed directly on site using a YSI probe. All additional water chemical properties for each sampling site were determined by the UC Davis analytical laboratory from a frozen 250ml sample (**Table S1**).

Cyanobacterial inoculation, growth, and passaging

Five axenic cyanobacterial strains: *Synechococcus elongatus* PCC 7942, *Synechococcus elongatus* UTEX 3055, *Anabaena sp.* PCC 7120, *Synechocystis sp.* PCC 6803, and *Leptolyngbya sp.* BL0902 were grown in BG-11 medium[61] at 24°C in a 12:12 light:dark cycle and their axenicity prior the experiment was confirmed by fluorescence microscopy after DAPI staining and by incubation of samples taken from cultures on LB agar plates in the dark at 30°C.

Prior to inoculation of each cyanobacterial host with bacteria from freshwater samples the axenic cyanobacterial cultures were split into 15ml glass tubes and adjusted to a starting optical density of OD_{680} = 0.2. Bacteria from each freshwater sub-location sample were independently concentrated using a multi-step filtration method. Water was pre-filtered through 20 and 5 µm filters and subsequently passed through 0.2 and 0.1 µm polyethersulfone filters until saturation, indicating sufficient bacterial biomass had been collected. Approximately 2 L of water was necessary to saturate filters from all three locations (n = 6 sub-locations). Using sterile technique the 0.2 and 0.1 µm filters were cut in half, and the 0.2 and 0.1 µm filters in the respective halves were combined. One set of combined filter halves was immediately frozen at -80°C for later DNA extraction, while the other set of combined filter halves was sterilely transferred to 30 mL of BG11 media and vortexed at maximum speed for 30 minutes to dissociate bacteria. Subsequently, the optical density of the resulting bacterial suspension was determined at 600 nm (OD_{600}). Using sterile technique a volume of bacterial suspension was then added to the 15ml tubes containing the axenic cyanobacterial hosts to achieve a normalized bacterial inoculation density of OD_{600} =

0.01. This filtering and inoculation process was repeated three times for each of the six sub-locations where water was collected, resulting in three biological inoculation replicates for each condition we evaluated (N = 180 total starting cultures: n = 5 cyanobacterial hosts, n = 6 sub-locations, n = 2 passaging intervals).

Following inoculation, the *in vitro* cyanobacterial communities were grown at 24°C in a 12:12 light:dark cycle and passaged into fresh BG-11 medium at either weekly or bi-weekly intervals. At the start and end of each passage interval, using sterile technique, a sample was collected to determine the optical density of the co-culture at 680, 700, and 750 nm. A volume of each co-culture from the previous passage was then inoculated into fresh BG-11 such that OD_{680} = 0.2. Additionally at each passage interval biomass was collected for DNA extraction (2 × 2 mL), and fixation for microscopy (1 mL). Biomass for DNA extraction was centrifuged for 20 min at 11,000xg at room temperature, frozen, and stored at -80°C until DNA extraction. Culture survival throughout the experiment was evaluated using visual inspection, optical density measurements, light microscopy, and fluorescence microscopy after DAPI staining. Cultures that had visible loss of cyanobacterial pigmentation, did not increase in optical density, or had visible eukaryotic or non-host cyanobacterial contamination under microscopic examination at the end of a passaging interval were removed from the experiment.

DNA extraction for shotgun metagenomics and sequencing

DNA for shotgun metagenomic sequencing was extracted from the frozen biomass of all co-culture samples that were not removed (See above) from the experiment (n = 120 samples), the frozen 5 µm polyethersulfone filter generated from the first biological replicate filtration of both Discovery Bay sub-locations (n = 2 samples), the frozen combined 0.2 and 0.1 µm polyethersulfone filter set generated from the first biological replicate filtration of each environmental sub-location (n = 6 samples), and the first biological replicate of the bacterial suspension generated for each sub-location directly used to inoculate axenic cyanobacterial cultures (n = 6 samples). DNA extraction was performed with the Qiagen DNeasy PowerSoil Pro Kit (Qiagen, USA) with modification. An additional initial lysis step was performed where samples were first heated in CD1 solution at 65°C for 30 minutes prior to the subsequent bead beating step. The DNA concentrations of the samples were quantified using a Quibit fluorometer (Thermo Fisher Scientific). DNA extracted from co-cultures (n = 120 samples) and from source material samples (filters and bacterial inocula ; n = 14 samples) were subsequently sent for sequencing using two different sequencing strategies. Raw DNA from source material samples was sent to the QB3 Genomics Laboratory at the University of California, Berkeley for library preparation and paired-end sequencing on an Illumina NovaSeq instrument using a 250 bp X 2 read length at an approximate depth of 10 Gbp per sample. Raw DNA from co-culture samples were sent to Novogene (Guangzhou, China) for library preparation and paired-end sequencing on an Illumina NovaSeq instrument using a 150 bp X 2 read length at an approximate depth of 10 Gbp per sample.

DNA extraction for 16S rRNA amplicon based metagenomics and sequencing

Genomic DNA was extracted from the biomass of cyanobacterial culture samples using the method 'Extraction of Chromosomal DNA From S. elongatus' as previously described by Clerico et al. (2007). A total of 1,935 samples were processed for sequencing in 10 batches. Samples were amplified for the V4–V5 hyper-variable regions of the 16S rRNA gene using primers 515F (5′-GCTCTTCCGATCTGTGYCAGCMGCCGCGGTAA-3') and 926R (5'-GCTCTTCCGATCTCCGYCAATTYMTTTRAGTTT-3'), both containing Illumina adapter sequences. Each PCR reaction included 10 ng of template DNA, 0.5 ng of the ZymoBIOMICS® Microbial Community Standard (ZymoResearch), and the Phusion® High-Fidelity PCR Master Mix (Thermo Fisher Scientific).

To assist with the identification of standards in 16S rRNA amplicon data, we constructed and included a dilution curve of the standard in each amplicon sequencing batch. Briefly, DNA extracted from a single cyanobacterial community sample was randomly selected from each sequencing batch and was spiked with varying concentrations of the ZymoBIOMICS® Microbial Community Standard (2, 1, 0.5, 0.25, 0.1 ng). This spike-in was prepared in duplicate resulting in 10 spike-in samples per sequencing batch and a total of 90 spike-in samples across all sequencing runs. Additionally, 1 sample per run included only 0.5 ng of the standard without any cyanobacterial community DNA, and 3 wells contained only the PCR reaction mix without any template. For additional information on samples and standards used in this analysis see **Table S3**

The PCR products were purified using SPRI AMPure XP beads (Beckman Coulter) to select for a fragment size of approximately 400 bp. The purified PCR products were then quantified using the Quant-iT High-Sensitivity dsDNA Assay Kit (Thermo Fisher Scientific) in a Tecan Spark® Multimode Microplate reader (Tecan). Following purification and quantification, the PCR products were sent to the Innovative Genomic Institute's Next Generation Sequencing Core. Here, they were sequenced on an Illumina MiSeq instrument using a read length of 2 x 300 bp and an expected read depth of 25,000 reads per sample.

16S rRNA data processing and downstream statistical analysis

Amplicon sequence variants (ASVs) and a count table of ASVs per sample were generated using the combined set of fastq files from all samples submitted for V4–V5 amplicon sequencing (n = 1935 samples; **Table S3**) using the USEARCH-UNOISE3 pipeline[62,63] with modifications. Briefly, we first assessed expected error rates and read orientation of raw reads using the `usearch -fastx_info` and `usearch -search_oligodb` commands with default parameters. Forward and reverse reads from each sample were then merged and aggregated using following command: `usearch -fastq_mergepairs *_R1_*.fastq -relabel @ -fastqout merged.fq -fastq_maxdiffs 10 -fastq_pctid 90`. Merged reads were then placed into the same orientation by searching against the SILVA_v138 SSU reference database [64] using the following command: `usearch -orient merged.fq -db SILVA_138_SSURef_NR99_tax_silva_trunc.fasta -fastqout oriented.fq -tabbedout orient.txt`. We subsequently removed reads from the analysis where the forward and reverse V4–V5 primers did not fully and identically match forward and reverse reads respectively by first searching for these primer sequences in our merged reads using the following command: `usearch -search_oligodb merged.fq -db primers_V4-V5.fa -strand both -threads 24 -userout merged_primer_hits.txt -userfields query+target+qlo+qhi+qstrand+diffs+trowdots+pairs`. We subsequently identified and filtered these problematic merged reads from the dataset using a custom R script: `Rscript filter_usearch_oligo_match.R -i merged.fq -H merged_primer_hits.txt`. Primers were then hard trimmed from the merged reads using the following command: `usearch -fastx_truncate merged_filtered.fq -stripleft 19 -stripright 20 -fastqout stripped.fq`. Merged reads with bases called as N or with an expected error rate of > 1 were then removed from the data with the following command: `usearch -fastq_filter stripped.fq -fastq_maxee 1.0 -fastq_maxns 0 -fastaout filtered.fa`. Unique sequences and their total count in the full dataset were then identified using the following command: `usearch -fastx_uniques filtered.fa -fastaout uniques.fa -strand both -sizeout -relabel Uniq -minuniquesize 1 -uc uniques.uc`. We then identified ASVs in our dataset and required an ASV to have at least 100 observed counts across

all samples (0.005% of total sequenced reads) using the following command: `usearch -unoise3 uniques.fa -zotus zotus.fa -tabbedout unoise3.txt -minsize 100`. An ASV count table was generated by mapping all sequences following primer removal to ASV sequences using the following command: `usearch -otutab stripped.fq -zotus zotus.fa -otutabout zotutab.txt -mapout zmap.txt`. In total we identified 2129 ASVs and were able to assign 87 % of sequenced amplicons generated in our study to this ASV set. Taxonomy was assigned to ASVs by first training a custom naive-bayes classifier using the SILVA_v138 database [64] truncated to the SSU V4-V5 region using the QIIME package [64,65]. Subsequently ASVs were taxonomically classified using the trained classifier. Finally, we identified ASVs associated with spike-in standards using a custom R script that performs two steps: (1) The sequence identity of ASVs to a database of 16S rRNA sequences from known spike-in organisms is calculated using ublast; (2) Robust regression analysis is applied to evaluate the slope of ASV counts relative to the known mass of spike-in standards across sets of spike-in samples included in each amplicon sequencing run. ASVs with a sequence identity ≥ 90% to known spike-in 16S sequences and a statistically significant positive slope (p-val of regression ≤ 0.05) were identified as standards. See **Tables S3-S5**, Data Availability, and Code Availability for additional information.

Amplicon samples were subset and filtered for downstream analysis to achieve three criteria: (i) to make amplicon and genome-resolved analyses congruent; (ii) to allow for direct comparison between covariates of weekly and bi-weekly passaging regimes; (iii) to remove select samples of low quality. Briefly, we first subset amplicon samples to the same set of cultures (e.g. combinations of source environment, cyanobacterial host strain, biological replicate, and passage rate) used in genome resolved metagenomic analysis (n = 108 individual cultures; n = 1240 amplicon samples; See Below). Genome resolved metagenomic analysis identified cultures that visually survived passaging, but did not pass community composition quality control criteria (See Below). Subsequently, we only retained and analyzed amplicon samples collected across these 108 cultures at bi-weekly intervals (0, 14, 28, 42, 56, 70, 84, 91, 105, and 119 d) to allow for direct comparisons between weekly and bi-weekly passaging regimes (n = 108 individual cultures; n = 991 amplicon samples). Finally, we removed individual amplicon samples with ≤ 1000 total ASV counts (n = 1 amplicon sample) and samples where the most abundant ASV was not the expected host cyanobacterium (n = 18 amplicon samples). Following sample selection and filtering amplicon analysis was performed on a final set of 972 samples (**Table S3 - S5**).

Statistical analysis of amplicon data was conducted in R v4.2.3 (See Code Availability). The relative abundance of order level taxa in cultures was calculated as the fraction of total counts assigned to ASVs grouped at the order taxonomic level per sample after ASVs identified as internal standards were removed. Relative abundance was averaged across all samples displayed in respective plots. Area plots were constructed using the ggplot package [66] in R.

Longitudinal alpha-diversity analysis was conducted on rarified ASV counts with ASVs identified as internal standards removed. Briefly, ASV matrices were ratified using the GUniFrac package in R [67,68] to the depth of the sample with the lowest number of counts (4746 and 642 counts for all ASVs and ASVs with cyanobacteria removed respectively). Richness and Shannon diversity metrics were calculated using the microbiome package in R [67]. The impact of inoculation source, cyanobacterial host, passage rate, and cultivation time on Richness and Shannon diversity were evaluated using linear mixed effects models implemented with the lme4 package in R [69]. All models took the following form: `[alpha-diversity metric] ~ General Site + Strain + Passage Rate + Total Days + (1|Tube Number)`. Total cultivation time in days was modeled as a factor rather than a continuous variable to allow for comparisons between time

points. Variation for each cyanobacterial culture (e.g. tube number; n = 108 cultures) was included as a random effect to account for the individual variance in each culture over longitudinal sampling. Evaluation of covariate significance in models was performed using Type III Analysis of Variance with the anova function in R. Pseudo-R-squared values for models were calculated with the MuMIn package [70,71] in R. Estimated marginal means for alpha-diversity values at each time point were generated and statistically significant differences between all pairwise marginal means were evaluated using the emmeans package [72] in R. Compact letter displays for significantly different groups were generated using the multcomp package in R.

Lagged beta-diversity analysis was conducted on additive log ratio (ALR) transformed ASV count data with and without ASVs taxonomically classified as cyanobacteria. Briefly, we first removed ASVs with positive counts in < 2 samples. We subsequently performed zero imputation on count matrices using the zCompositions package [73] in R. ALR transformation was then applied using a custom R function (See Code Availability) that utilizes the counts of ASVs classified as standards in each sample to calculate the geometric means for each sample. Aitchison distance between samples was then calculated as the euclidean distance between ALR transformed samples using the vegan package [70] in R. For each individual cyanobacterial culture (n = 108 cultures) we captured the atchison distance between each sample and its previous time point (e.g. 0 d ← 14 d, 14 d ← 28 d). This resulted in a maximum of 9 time point comparisons across the time course for each cyanobacterial culture. The impact of inoculation source, cyanobacterial host, passage rate, and cultivation time on beta-diversity between preceding time points were evaluated using linear mixed effects models implemented with the lme4 package in R [69]. Subsequent statistical analysis was conducted identically to alpha-diversity analysis above with cyanobacterial culture number included as a random effect in the model. For all diversity modeling statistical outputs see **Table S13**.

Differential abundance of ASVs between the 0 d → 84 d and the 84 d → 119 d time points was evaluated using the Maaslin2 package [69,74] in R. Briefly, ASVs with < 10 positive counts across all samples were first removed (n = 1682 remaining ASVs). The Maaslin2 function was then run on the resulting count matrix with the following parameters: `min_abundance = 0, min_prevalence = 0, min_variance = 0, analysis_method = "LM", max_significance = 0.05, fixed_effects = c("General_Site", "Strain", "Passage_Rate", "Total_Days_Fac"), reference = c("Total_Days_Fac,84"), random_effects = c("Tube_No"), correction ="BH", cores = 12, plot_heatmap = F, plot_scatter = F`. ASVs with significant differences were required to have a FDR ≤ 0.05. For all differential abundance statistical outputs see **Table S6.**

Assembly independent marker gene analysis of shotgun metagenomic data

Due to lower recovery of genomic bins from source environment samples relative to cyanobacterial cultures, we chose to apply an assembly independent marker gene analysis method to more fairly compare diversity and taxonomic composition between environmental samples and cyanobacterial community cultures (**Table S8 and S9**). This approach allows the identification of operational taxonomic units (OTUs) directly from shotgun metagenomic read data and circumvents the expected data losses that occur during short read assembly and genomic binning of samples with high species complexity. Briefly, we used SingleM v0.13.2 (https://github.com/wwood/singlem) to search metagenomic reads from each sample against the singleM database v3.0.5 (https://zenodo.org/records/7130825) containing 59 conserved single copy marker genes from bacteria and archaea. Search was implemented using the `singlem pipe` command with default parameters. Identified marker genes were labeled with GTDB taxonomy using the pplacer algorithm for phylogenetic placement. Subsequently OTUs were

generated by clustering marker sequences identified across all samples at 95 % nucleotide identity using the following command: `singlem summarise --input-otu-tables [All singleM pipe outputs] --cluster --cluster-id 0.95 --wide-format-otu-table [output]`.

Statistical analysis of singleM single copy marker gene OTUs was conducted in R v4.2.3 (See Code Availability). We initially evaluated three potential marker genes with demonstrated potential for species level identification and clustering[75] to conduct the subsequent diversity and taxonomic analysis: ribosomal protein L6 (rpL6), ribosomal protein S3 (rpS3), and ribosomal protein S9 (rpS9). All genes evaluated exhibited a similar level of OTU recovery and predicted similar taxonomic composition across all samples, thus we subsequently chose rpL6 to conduct the full analysis. Samples used in our analysis were subset to a set of 108 cyanobacterial community cultures and 13 source environment samples that were used in genome resolved metagenomic analysis (See Below). Following sample subsetting OTUs with no counts in any remaining sample were removed resulting in a total final set of 8,156 rpL6 OTUs (**Table S8**). Relative abundance of phylum level taxa was calculated as the fraction of total counts assigned to OTUs grouped at the phylum taxonomic level per sample. Phyla representing less than 1% of the total were collectively categorized under 'Other'. For clarity, specific phylum names were modified, with 'Patescibacteria' being renamed to 'CPR'. Relative abundance was averaged across all samples displayed in respective plots. Plots were constructed using the ggplot package [66] in R.

Alpha-diversity analysis was conducted on rarefied OTU counts. Briefly, OTU counts across all samples (n = 121 samples) were ratified using the GUniFrac package in R [76] to the depth of the sample with the lowest number of counts (965 counts). Richness and Shannon diversity metrics were calculated using the microbiome package in R [77] The means of richness and Shannon diversity were compared between all source samples grouped by sampling location using analysis of variance (ANOVA) followed by pairwise comparisons using Tukey's multiple comparisons of means implemented in R with comparisons reporting a FDR ≤ 0.05 being significant. The means of richness and Shannon diversity were compared between all source and all culture samples using Welch's two Sample t-test implemented in R with comparisons reporting a p-value ≤ 0.05 being significant. The means of richness and Shannon diversity were compared between all cyanobacterial culture samples grouped by sampled location using analysis of variance (ANOVA) followed by pairwise comparisons using Tukey's multiple comparisons of means implemented in R with comparisons reporting a FDR ≤ 0.05 being significant.

To explicitly assess the dissimilarity of the non-cyanobacterial community composition between environmental source samples and cultures we removed rpL6 OTUs from the dataset that were taxonomically classified in the phylum cyanobacteria and subsequently assessed the beta-diversity of samples using Aitchison distance. Briefly, we first removed rpL6 OTUs classified as cyanobacteria and those with positive counts in < 2 samples (n = 1468 OTUs remaining). We subsequently performed zero imputation on the filtered rpL6 OTU count data using the zCompositions package [73] in R. Centered log-ratio (CLR) transformation was then applied using a custom R function (See Code Availability). Aitchison distance between samples was then calculated as the euclidean distance between CLR transformed samples using the vegan package [70] in R. The marginal influence of sample type (cyanobacterial culture or source environment) and location (Discovery Bay, Eel River, or Strawberry Creek) were estimated using permutational analysis of variance implemented in the vegan package in R [70] using a model of the form: Distance_Matrix ~ sample_type + location. Model terms were assessed marginally using 999 permutations. Principal component analysis was performed in R on centered and scaled CLR transformed rpL6 OTU count data to display community dissimilarities. PCA was chosen for this

data display to better preserve the true distances between samples, specifically the large distances between the environmental source environment and cyanobacterial culture samples.

To evaluate the over/under enrichment of bacterial taxonomic orders in cultured cyanobacterial communities relative to source environmental samples we applied a permutation based method where the observed frequency of taxa in cyanobacterial communities was compared to the frequency of taxa in communities generated through random sampling of all possible taxa identified in both cyanobacterial and source communities. Briefly, we first assessed the number of unique rpL6 OTUs identified in each of the 108 cyanobacterial community samples used in this analysis (**Table S8**). We then created a set of rpL6 OTUs that were non-cyanobacterial and detected in at least one sample across the set of 121 cyanobacterial community and source environment samples (7,773 rpL6 OTUs). We then generated 10,000 permuted data sets of the 108 cyanobacterial community samples where for each sample in a permuted dataset we selected a number OTUs from the total set of 7,773 randomly without replacement equal to the original number of unique rpL6 OTUs identified in that sample (e.g. if a sample had 20 unique OTUs we randomly selected 20 OTUs into its permuted version). This was done such that permuted sample sets had the same OTU richness as the original cyanobacterial community samples. We then aggregated OTUs at the order level and calculated the number of times each order was observed in the set of original 108 samples as well as in each set of the 10,000 permuted samples. Then for each taxonomic order we generated two p-values, one for over enrichment and one for under enrichment, by quantifying the number of times an order appeared more often or less often in the set of permuted samples relative to its observed frequency in the original set of 108 cyanobacterial communities divided by the number of permutations. A combined, two-sided, p-value was generated by taking the minimum of the two p-values and multiplying by 2. A pseudo z-score for each order was generated by subtracting the mean number of times an order was observed across all permuted data sets from the number of times it was observed in the original set of 108 cyanobacterial communities divided by the standard deviation of the number of times an order was observed across all permuted data sets. The p-values for each order were corrected for multiple testing using FDR and comparisons reporting a FDR ≤ 0.05 were considered significant. Also see Code Availability and **Tables S8 and S9**.

Metagenome assembly, binning, bin de-replication

Raw sequence reads underwent processing to remove Illumina adaptors and phiX sequences using BBduk (https://sourceforge.net/projects/bbmap/). Quality trimming of the reads was performed using Sickle[78]. Each sample was individually assembled using idba_ud [79], or when idba_ud failed metaspades [79,80] to generate DNA scaffolds. Scaffolds shorter than 1000 bp were excluded from further analysis. Scaffolds larger than 5000 bp were then subjected to overlap based reassembly using COBRA v50[81,82]. Contigs extended and validated by COBRA were non-redundantly re-incorporated into the assembly. Gene prediction was conducted using prodigal v2.6.3 [83] in metagenome mode, and subsequent annotation was performed through a blast search against the uniprot and kegg[78 84,85] databases. Genome binning was performed on each sample individually by grouping scaffolds longer than 2,500 bp based on sequence compositionality and differential coverage. Cross-mapping of reads from all samples was carried out using bowtie2 [86] to generate differential coverage information for binning. Samples were then independently binned using four algorithms: MetaBat2, maxbin2, CONCOCT, and vamb [87–90]. The best bins across the four methods for each sample were selected using DasTool [91]. Bins were then evaluated for completeness and contamination using checkM [91,92], which utilizes a single copy gene inventory method. Only bins with an estimated completeness of ≥60 % and estimated contamination of ≤5 % were retained. Bins of sufficient quality across all samples were de-replicated at the species level and a representative species bin was slected using dRep [93]. Bins were considered the same species if ANI was ≥ 95 % across ≥ 10 % of the genome length.

Taxonomic placement of genomes was performed using the GTDB-tk [94] classify workflow against the GTDB-R214 taxonomy. Gene prediction was conducted again on non-redundant species-representative genomes using prodigal v2.6.3 [83] in single genome mode. This resulted in a final non-redundant genome set of 537 genomes representing all samples in the study. Also see **Tables S7.**

Phylogenetic analysis

GToTree v1.6.34 [95] was employed to detect, align, trim, and concatenate 120 universal single copy genes [96] from genomes within the non-redundant genome set as well as an archaeal genome outgroup *Acidianus hospitalis* W1 (GCF_000213215.1_2). Genomes were excluded from phylogenetic analysis if they had < 10 % of the markers identified. A Maximum-likelihood phylogenetic reconstruction of the aligned sequences was conducted with IQ-TREE multicore version 1.6.12 [97]. The best-fit models of evolution were determined for each marker protein partition using ModelFinder Plus with the -m MFP option, considering all available models and their possible combinations with rate heterogeneity and base frequency variations. The robustness of the resulting tree topology was assessed using ultrafast bootstrap analysis with 1,000 replicates (-bb 1000) and the Shimodaira-Hasegawa approximate likelihood ratio test (SH-aLRT) with 1,000 replicates (-alrt 1000). To improve the accuracy of bootstrap values, more thorough searches were employed during the bootstrap phase using the -bnni option. The resulting phylogenetic tree was displayed using iTol v6.8.1 [98].

Genome Abundance Mapping and Diversity Analysis

Sequencing reads from each sample were mapped to a concatenated FASTA file of all 537 dereplicated genomic bins using Bowtie2 [86]. Subsequently, CoverM v0.6.1 [99] was used to quantify the number of reads mapped, the coverage, and the breadth of coverage for each genome. Breadth is defined as the proportion of a genomic bases in a genome covered by at least one mapped read. Mapped reads were discarded from this analysis if < 50 bp of a read mapped, the read mapping identity was < 97 %, or if both reads in a pair did not map in proper orientation and within the estimated sequencing insert length. Finally, for a genome to be considered as present in a sample we required that the genome have at least 100 properly mapped reads from that sample and that $\log_2$(Observed Breadth / Expected Breadth) ≥ -1. Expected breath for a genome in a sample was calculated as: Expected Breath = $1 - e^{-0.883 * \text{coverage}}$. If a genome did not pass these filtering criteria its counts and coverage were set to 0 for a sample. These filtering criteria were applied as they significantly reduce the frequency at which genomes that are not truly in a sample are identified in a sample due read recruitment and mapping to homologous regions shared by multiple genomes.

Following mapping of sample reads to genomes we initially identified and filtered out samples from subsequent analysis that did not meet the following criteria: culture samples where ≥ 50 % of reads did not map to assembled genomes (n = 1 sample), culture samples where the expected host cyanobacterium did not comprise ≥ 90 % of the cyanobacterial community fraction (n = 11 samples), culture samples that were determined to be mis-labeled following sequencing (n = 1 sample). Of the 134 samples submitted for sequencing 13 source environment samples and 108 cyanobacterial culture samples were utilized for downstream analysis (n = 121 total samples). This set of filtered samples was also used in all other analyses conducted in this study.

Relative abundance of phylum level taxa was calculated as the fraction of total counts assigned to genomes grouped at the phylum taxonomic level per sample. Phyla representing less than 1% of the total were collectively categorized under 'Other'. For clarity, specific phylum names were modified, with 'Patescibacteria' being renamed to 'CPR'. Relative abundance was averaged

across all samples displayed in respective plots. Plots were constructed using the ggplot package [66] in R.

To explicitly assess the dissimilarity of the non-cyanobacterial community composition between environmental source samples and cultures (n = 121 samples) we removed the four genomes known to be from our host cyanobacterial species and subsequently assessed the beta-diversity of samples using Aitchison distance. Briefly, genomes assigned to our host cyanobacterial species and genomes with positive counts in < 2 samples were removed (n = 501 genomes remaining). Zeros in filtered genome read count data were imputed using the zCompositions package [73] in R. CLR transformation was then applied to the data using a custom R function (See Code Availability). Aitchison distance between samples was then calculated as the euclidean distance between CLR transformed samples using the vegan package [70] in R. The marginal influence of sample type (cyanobacterial culture or source environment) and location (Discovery Bay, Eel River, or Strawberry Creek) were estimated using permutational analysis of variance implemented in the vegan package in R [70] using a model of the form: `Distance_Matrix ~ sample_type + location`. Model terms were assessed marginally using 999 permutations. Principal component analysis was performed in R on centered and scaled CLR transformed genome coverage data to display community dissimilarities. PCA was chosen for this data display to better preserve the true distances between samples, specifically the large distances between the environmental source samples and cyanobacterial culture samples.

To explicitly assess the dissimilarity existing between cyanobacterial community cultures we assessed beta-diversity of these samples by themselves (n = 108 samples) using Aitchison distance. Briefly, genomes with positive counts in < 2 samples were removed (n = 291 genomes remaining). Zeros in filtered genome read count data were imputed using the zCompositions package [73] in R. CLR transformation was then applied to the data using a custom R function (See Code Availability). Aitchison distance between samples was then calculated as the euclidean distance between CLR transformed samples using the vegan package [70] in R. The influence of cyanobacterial host strain, inoculation source location, passage rate, and the interactions of these covariates on community dissimilarity were estimated using permutational analysis of variance implemented in the vegan package in R using a model of the form: `Distance_Matrix ~ host_strain * location * passage_rate`. Model terms and interactions were assessed sequentially using 9,999 permutations. Uniform manifold approximation (UMAP) implemented with the umap package[100] in R was performed on CLR transformed count data to display relationships between samples. UMAP was chosen for this purpose as it provides better recovery of sample clusters and sub-cluster structure, allowing the nested nature of covariate influences to be visible.

Core microbiome criteria

The definition of a core microbiome in this study was driven by two primary criteria: (i) A core species of the cyanobacterial community microbiome will reproducibly and consistently co-occur with multiple cyanobacterial host species; (ii) A core species will be recruited into a cyanobacterial community reproducibly from multiple source environments. As an explicit definition for a core microbiome is lacking we developed an operational definition to fulfill these dual criterion of host-diversity and environmental representation. This enabled focus on species with a consistent and ecologically significant presence in cyanobacterial communities. Additionally, by balancing stringency with ecological realism, it is more likely that our findings reflect genuine biological interactions rather than sporadic occurrences or artifacts of sampling. To fulfill the first criteria, we required that a species was present (had non-zero metagenomic read counts) with at least 4 cyanobacterial host species in at least 3 independent communities for each host (totalling a

minimum of 12 samples). This made it more likely the observed relationships were not host-specific but rather indicative of a broader ecological interaction across the diverse cyanobacterial hosts evaluated. To fulfill the second criteria, a core species was also required to be present (had non-zero metagenomic read counts) in cyanobacterial communities inoculated using at least 2 distinct source environmental locations in at least 3 independent communities from each distinct source inoculum (totalling a minimum of 6 samples). This made it more likely that the observed relationships were not environment-specific but rather indicative of a broader environment agnostic selective pressure that enabled these species to enter communities despite the disparate microbial diversity observed in the source material.

Population level single nucleotide variant analysis

We evaluated the strain-level similarity of genomes between cyanobacterial cultures to: (i) Estimate the frequency at which identical strain sharing occurred between cultures, which may indicate cross-contamination; (ii) Evaluate the impact of the inoculum source and cyanobacterial host strain on the microdiversity of individual genomes across our cyanobacterial communities. To evaluate between-culture strain-level similarity the population level single nucleotide variants (SNVs) were compared for each genome between samples using inStrain v1.6.4 [101]. Briefly, we first profiled the SNVs of genomes in each of the 108 samples using inStrain profile as follows: `inStrain profile [sample_bam_file] [genome_scaffolds] -o [inStrain_output] -p 64 -s [genome_scaffold2bin_file] --min_snp 9999 --skip_mm_profiling`. We applied inStrain defaults using a minimum coverage of 5X to conclusively identify an SNV in a sample. Subsequently we used inStrain compare in database mode to produce an all-vs-all comparison of the genomes across the 108 SNV profiles generated as follows: `inStrain compare -i [all_108_instrain_profile_outputs] -o [inStrain_compare_output] -p 64 -s [genome_scaffold2bin_file] --database_mode`. We note that this analysis has significantly higher genomic coverage requirements relative to genomic bin recovery and quantitative community compositional analysis. Despite these technical issues we evaluated at least 1 inStrain two-sample comparison for 133 non-host cyanobacterial genomes in our study (n = 3088 total comparisons). This set included 13 core microbiome species genomes (n = 555 total comparisons). Subsequent statistical analysis was conducted in R v4.2.3 (See Code Availability). Also see **Table S12.**

To evaluate the frequency at which identical strain sharing occurred between cultures, and potential cross-contamination, we considered a genome to be identical between samples if its population ANI (popANI) between two samples was ≥ 99.999% based on previously published metrics [101]. We specifically evaluated identical strain sharing occurring between cultures inoculated from distinct environmental sources, as identical strains occurring in cultures inoculated using geographically distinct source material would be probabilistically more indicative of cross-contamination. We found that across a total of 176 two-sample comparisons (n = 11 genomes; n = 6 core microbiome genomes) between samples inoculated from different source environments, 0 comparisons were above the popANI threshold to indicate identical strain sharing. Thus, we concluded that while identical species were recovered from geographically distinct source inocula, the species recovered were in fact different strains that were not likely directly transferred between cultures by cross-contamination.

The impact of sharing the same cyanobacterial host strain or source environment on pairwise genome popANI was modeled using beta regression. Briefly, popANI values for comparisons involving any of the 4 host cyanobacterial species genomes were removed from the dataset (n = 3,088 comparisons evaluated). As some popANI values were equal to 1 (n = 20 comparisons) a small constant (0.0000001) was subtracted from all popANI values to enable beta regression modeling. popANI was modeled as a function of the inoculation source location being the same

or different and the cyanobacterial host strain being the same or different for all 3,088 two sample comparisons using beta regression implemented using the betareg package [102] in R with the following model form: `popANI ~ same_location + same_host`. Estimated marginal mean popANI values were calculated using the emmeans package in R [72,101]. Statistical significance for pairwise comparisons between estimated marginal means was evaluated using the multcomp package[103] in R. Tukey's method was applied for comparing multiple means with adjusted p-values ≤ 0.05 considered significant. Also see Code Availability.

Functional Annotation and Core Microbiome Functional Enrichment Analysis

Predicted proteins in all 537 genomes were annotated using kofamscan [103,104] with KOfam HMM set r02_18_2020. Only HMM hits exceeding score thresholds set individually for each HMM were used for downstream analysis. In cases where a protein had more than one significant hit to a Kofam HMM, the annotation with the highest score was chosen. In addition, the pipeline METABOLIC [105] was used to annotate predicted proteins using the function METABOLIC-G.pl and default settings.

To determine metabolic functions enriched in genomes of the core microbiome, Fisher tests were conducted to evaluate the over/under enrichment of each identified KO in core microbiome genomes (n = 25) relative to all remaining genomes (n = 512 genomes) using the *fisher.test* function in R v 4.2.1. The p-values were adjusted to account for multiple testing using false discovery rate with the *p.adjust* function in R, with FDR ≤ 0.05 considered significant. KOs were counted using presence/absence per genome. Two sets of tests were run on KOs. The first set of tests used all genomes (n = 537 genomes) and all KOs that appeared in more than 10 MAGs (n = 4264 KOs). The second set of tests was constrained to genomes in the phyla Proteobacteria, Bacteroidota, and Spirochaetota (n = 391 genomes), and KOs that were found in over 10 MAGs (n = 3749 KOs). As the core microbiome had a highly constrained taxonomic composition, the constrained test was used to evaluate if differential enrichment of KOs was an artifact of taxonomic bias (e.g. was due to functions encoded in taxa not present in the core microbiome) or if the same functions were consistently evaluated as differentially enriched in the core microbiome even when taxonomy was constrained in the analysis. KOs were assigned to broader functional categories as in Diamond et al. [106]. Differential over/under representation analysis of functional categories was performed using the *fisher.test* function in R v 4.2.1. The p-values were adjusted to account for multiple testing using false discovery rate with the *p.adjust* function in R, with FDR ≤ 0.05 considered significant. See **Tables S14-S16** for additional information.

Identification, curation, and analysis of putative mobile genetic elements

We identified contigs predicted to be MGEs within the 537 genomes assembled in this study using geNomad[81] with the parameters: `--end-to-end, --cleanup, --enable-score-calibration, and --splits 8`. Contigs predicted to be plasmids by geNomad were filtered prior to downstream analysis to only include contigs with a FDR ≤ 0.05, and a length over 12 kb. A minimum contig length of 12 kb was chosen as the plasmid prediction accuracy of geNomad has been reported to be ≥ 90.0% for contigs above this length[81]. Predicted plasmids were assigned to genomes using a guilt-by-association approach where the genomic bin a plasmid contig was found in was considered the putative host. The proportion of contigs predicted to be plasmids per genome was compared between groups of genomes assigned to the core, auxiliary, and source microbiome groups using Dunn's test implemented in the dunn.test package in R. The proportion of plasmid contigs per genome was chosen as the comparison metric to control for differences in total contig number per genome across all genomes analyzed. All predicted plasmid contigs associated with core microbiome genomes (n = 198 contigs) were searched against PLSDB v2021_06_23_v2 [107] using the PLSDB search web server (https://ccb-microbe.cs.uni-

saarland.de/plsdb/plasmids/search_form/seq/) with the mash dist strategy, min p-value ≤ 0.1, and min mash distance ≤ 0.5. For additional information see **Tables S18-S20**.

Additional circular element search, curation, and analysis was performed on all core microbiome genomes (n = 25 genomes). Briefly, the *de novo* assembly function of Geneious Prime 2023.0.3 (Biomatters Ltd) with default parameters was used to perform overlap assembly on all contigs from each core microbiome genome. We subsequently evaluated the circularity of assembled contigs in each genome by searching for overlaps between the first and last 500 bp of each assembled contig larger than 10 kb using BLAST v2.10.1. This effort identified a single putatively circular *de novo* assembled contig (Plasmid_1) in a *Gemmobacter* genome of the core microbiome (unicom2021_Dbay_Inner_A7120_BW_A_idba_concoct_16). Circularity and *de novo* assembly joins of this contig were verified by mapping of sample reads to the contig using Bowtie2[86] and visually determining that mapped paired reads spanned all assembly joins and the ends of the contig. Following identification of a circular plasmid in a *Gemmobacter* genome of the core microbiome we attempted to manually identify additional circular plasmids in the 4 non-representative genome bins that were recovered during our assembly and binning of samples (See Above), that shared species identity with *Gemmobacter* genomes of the core microbiome, but were not chosen as the representative genome bins for those species by dRep. While the representative genome bins for a species demonstrate the highest overall quality across a set of metrics, alternative genome bins recovered for the same species may still have individual contigs with better assembly. By applying the same *de novo* assembly, overlap identification, and read mapping procedure to the 4 additional *Gemmobacter* bins not in our species representative bin set, we identified one additional circular plasmid (Plasmid_2) in a *Gemmobacter* genome of a different species than Plasmid_1. Genes on these two plasmids were annotated with prodigal v2.6.3 [83], and functional annotations were predicted using kofamscan [104] with KOfam HMM set r02_18_2020.

Assessment of functions putatively enriched on MGEs associated with genomes of the core microbiome was conducted by repeating KEGG KO enrichment analysis on all 537 genomes assembled in this study after removing contigs identified as plasmids by genomad. Briefly, we first removed all 850 contigs classified as predicted plasmids by genomad from genomes . We then re-quantified the presence/absence of KEGG KOs in each genome with predicted plasmid contigs removed. All KOs used in the original full enrichment analysis (See Above; n = 537 genomes) were re-evaluated for statistical over/under enrichment in core microbiome genomes relative to non-core microbiome genomes using the *fisher.test* function in R v 4.2.1. The p-values were adjusted to account for multiple testing using false discovery rate with the *p.adjust* function in R, with FDR ≤ 0.05 considered significant. We then identified the set of KOs that were significantly enriched in the original core microbiome analysis, but were no longer significantly enriched in the core microbiome analysis with MGE contigs removed. A relative change between the two analyses for each KO was calculated as: `Fold Change in Odds Ratio = KO_Odds_Ratio_Original / KO_Odds_Ratio_MGE_Removed`.

Data Availability

Genomic data including non-redundant species representative genomic bins and raw sequencing read data generated as part of this project are available under the NCBI BioProject accession number **[pending]**.

Code Availability

Code used in the analysis for this paper are available at the following GitHub repository: https://github.com/SDmetagenomics/Kust_Cyano_2023

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Acknowledgements

We would like to thank Prof. Susan Golden, Prof. James Golden, and Dr. Arnaud Taton for providing the cyanobacterial strains used in this study. Eel River sampling occurred at the UC Angelo Coast Range Reserve. Funding for this project was provided in part by the Gordon and Betty Moore Foundation (GBMF9321). Funding for this project was also provided in part by the Shurl and Kay Curci Foundation.

Author Contributions

AK, JFB, and SD conceived overall experimental concepts and design. AK and SD designed the sampling and cultivation strategy. AK, KBG, and SD collected water samples. AK, CCP, and SD performed cultivation and sampling. AK, CCP, NK, and SD performed sequencing and sequence processing. AK and SD performed sequence assembly and genomic binning. AK, JZ, and SD performed annotation and statistical analysis on amplicon and metagenomic datasets. AK, JZ, and SD wrote the manuscript. AK, JZ, JFB, and SD edited the manuscript

Competing Interests

The authors declare no competing interests.

Corresponding Author

All correspondence should be addressed to Spencer Diamond (sdiamond@berkeley.edu)

References

1. Lewis M. Ward, Birger Rasmussen, Woodward W. Fischer. Primary Productivity Was Limited by Electron Donors Prior to the Advent of Oxygenic Photosynthesis. Journal of Geophysical Research: Biogeosciences 124, 211–226 (2018).
2. Chorus, I., Fastner, J. & Welker, M. Cyanobacteria and Cyanotoxins in a Changing Environment: Concepts, Controversies, Challenges. Water 13, 2463 (2021).
3. Huisman, J. et al. Cyanobacterial blooms. Nature Reviews Microbiology 2018 16:8 16, 471–483 (6 2018).
4. 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).
5. Karl, D. et al. Dinitrogen fixation in the world's oceans. The Nitrogen Cycle at Regional to Global Scales 47–98 (2002).

6. Stal, L. J. Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature? Environ. Microbiol. 11, 1632–1645 (2009).
7. Paerl, H. W. & Huisman, J. Blooms Like It Hot. Science (2008) doi:10.1126/science.1155398.
8. Paerl, H. W. & Pinckney, J. L. A mini-review of microbial consortia: Their roles in aquatic production and biogeochemical cycling. Microb. Ecol. 31, 225–247 (1996).
9. Durán, P. et al. Shared features and reciprocal complementation of the *Chlamydomonas* and *Arabidopsis* microbiota. Nat. Commun. 13, 1–14 (2022).
10. Bell, W. & Mitchell, R. Chemotactic and growth responses of marine bacteria to algal extracellular products. Biol. Bull. 143, 265–277 (9 1972).
11. Seymour, J. R., Amin, S. A., Raina, J.-B. & Stocker, R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. Nature Microbiology 2, 1–12 (2017).
12. Jackrel, S. L., Schmidt, K. C., Cardinale, B. J. & Denef, V. J. Microbiomes Reduce Their Host's Sensitivity to Interspecific Interactions. MBio (2020).
13. Kim, M., Shin, B., Lee, J., Park, H. Y. & Park, W. Culture-independent and culture-dependent analyses of the bacterial community in the phycosphere of cyanobloom-forming *Microcystis aeruginosa*. Sci. Rep. 9, 20416 (2019).
14. Dick, G. J. et al. The genetic and ecophysiological diversity of *Microcystis*. Environ. Microbiol. 23, 7278–7313 (12 2021).
15. Bouma-Gregson, K. et al. Impacts of microbial assemblage and environmental conditions on the distribution of anatoxin-a producing cyanobacteria within a river network. ISME J. 13, 1618–1634 (2019).
16. Pérez-Carrascal, O. M. et al. Single-colony sequencing reveals microbe-by-microbiome phylosymbiosis between the cyanobacterium *Microcystis* and its associated bacteria. Microbiome vol. 9 (12 2021).
17. Zhu, C.-M. et al. Alternate succession of aggregate-forming cyanobacterial genera correlated with their attached bacteria by co-pathways. Sci. Total Environ. 688, 867–879 (2019).
18. Cook, K. V. et al. The global interactome. Limnol. Oceanogr. 65, S194–S207 (2020).
19. Halary, S. et al. Intra-population genomic diversity of the bloom-forming cyanobacterium, *Aphanizomenon gracile*, at low spatial scale. ISME Commun 3, 57 (2023).
20. Duxbury, S. J. N. et al. Reproducible spatial structure formation and stable community composition in the cyanosphere predicts metabolic interactions. bioRxiv (2022) doi:10.1101/2022.12.13.520286.
21. Haines, M., Vadlamani, A., Daniel Loty Richardson, W. & Strous, M. Pilot-scale outdoor trial of a cyanobacterial consortium at pH 11 in a photobioreactor at high latitude. Bioresour. Technol. 354, 127173 (2022).
22. Nelson, C., Giraldo-Silva, A. & Garcia-Pichel, F. A symbiotic nutrient exchange within the cyanosphere microbiome of the biocrust cyanobacterium, *Microcoleus vaginatus*. ISME J. 15, 282–292 (2021).
23. Zhu, L. et al. Bacterial Communities Associated with Four Cyanobacterial Genera Display Structural and Functional Differences: Evidence from an Experimental Approach. Front. Microbiol. 7, 1662 (2016).
24. Cole, J. K. et al. Phototrophic biofilm assembly in microbial-mat-derived unicyanobacterial consortia: model systems for the study of autotroph-heterotroph interactions. Front. Microbiol. 5, 109 (2014).
25. Rishiram Ramanan, Zion Kang, Byung-Hyuk Kim, Dae-Hyun Cho, Long Jin, Hee-Mock Oh. Phycosphere bacterial diversity in green algae reveals an apparent similarity across habitats. Algal Research 8, 140–144 (2015).
26. Steinrücken, P., Jackson, S., Müller, O., Puntervoll, P. & Kleinegris, D. M. M. A closer look into the microbiome of microalgal cultures. Front. Microbiol. 14, 1108018 (2023).

27. Ling, N., Wang, T. & Kuzyakov, Y. Rhizosphere bacteriome structure and functions. Nat. Commun. 13, 1–13 (2022).
28. Trivedi, P., Leach, J. E., Tringe, S. G., Sa, T. & Singh, B. K. Plant–microbiome interactions: from community assembly to plant health. Nat. Rev. Microbiol. 18, 607–621 (2020).
29. Xu, J. et al. The structure and function of the global citrus rhizosphere microbiome. Nat. Commun. 9, 1–10 (2018).
30. Li, Y. et al. Rhizobacterial communities of five co-occurring desert halophytes. PeerJ 6, (2018).
31. Stopnisek, N. & Shade, A. Persistent microbiome members in the common bean rhizosphere: an integrated analysis of space, time, and plant genotype. ISME J. 15, 2708–2722 (2021).
32. Wardell, G. E., Hynes, M. F., Young, P. J. & Harrison, E. Why are rhizobial symbiosis genes mobile? Philos. Trans. R. Soc. Lond. B Biol. Sci. 377, 20200471 (2022).
33. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A guide to the natural history of freshwater lake bacteria. Microbiol. Mol. Biol. Rev. 75, 14–49 (2011).
34. Chiriac, M.-C., Haber, M. & Salcher, M. M. Adaptive genetic traits in pelagic freshwater microbes. Environ. Microbiol. 25, 606–641 (2023).
35. Kim, M., Shin, B., Lee, J., Park, H. Y. & Park, W. Culture-independent and culture-dependent analyses of the bacterial community in the phycosphere of cyanobloom-forming *Microcystis aeruginosa*. Sci. Rep. 9, 20416 (2019).
36. Smith, D. J. et al. Individual *Microcystis* colonies harbour distinct bacterial communities that differ by *Microcystis* oligotype and with time. Environ. Microbiol. 23, 3020–3036 (2021).
37. Bowers, R. M. et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. 35, 725 (2017).
38. Neu, A. T., Allen, E. E. & Roy, K. Defining and quantifying the core microbiome: Challenges and prospects. Proc. Natl. Acad. Sci. U. S. A. 118, (2021).
39. Chunbo Dong, Qiuyu Shao, Qingqing Zhang, Ting Yao, Jianzhong Huang, et al. Preferences for core microbiome composition and function by different definition methods: Evidence for the core microbiome of *Eucommia ulmoides* bark. Sci. Total Environ. 790, 148091 (2021).
40. Woodhouse, J. N., Ziegler, J., Grossart, H.-P. & Neilan, B. A. Cyanobacterial Community Composition and Bacteria–Bacteria Interactions Promote the Stable Occurrence of Particle-Associated Bacteria. Front. Microbiol. 9, 329181 (2018).
41. Cook, K. V. et al. The global interactome. Limnol. Oceanogr. 65, S194–S207 (2020).
42. Smith, D. J. et al. Individual *Microcystis* colonies harbour distinct bacterial communities that differ by Microcystis oligotype and with time. Environ. Microbiol. 23, 3020–3036 (2021).
43. Jackrel, S. L., Yang, J. W., Schmidt, K. C. & Denef, V. J. Host specificity of microbiome assembly and its fitness effects in phytoplankton. ISME J. 15, 774–788 (2020).
44. Koblížek, M. Ecology of aerobic anoxygenic phototrophs in aquatic environments. FEMS Microbiol. Rev. 39, 854–870 (2015).
45. Vigneron, A. et al. Multiple Strategies for Light-Harvesting, Photoprotection, and Carbon Flow in High Latitude Microbial Mats. Front. Microbiol. 9, 419618 (2018).
46. Motone, K. et al. A Zeaxanthin-Producing Bacterium Isolated from the Algal Phycosphere Protects Coral Endosymbionts from Environmental Stress. MBio (2020) doi:10.1128/mbio.01019-19.
47. Sultana, S., Bruns, S., Wilkes, H., Simon, M. & Wienhausen, G. Vitamin B12 is not shared by all marine prototrophic bacteria with their environment. ISME J. 17, 836–845 (2023).
48. Ken Sasaki, Facundo J. Marquez, Naomichi Nishio, Shiro Nagai. Promotive effect of 5-aminolevulinic acid on the growth and photosynthesis of Spirulina platensis. J. Ferment. Bioeng. 79, 453–457 (1995).

49. Rabalais, N. N. Nitrogen in aquatic ecosystems. Ambio 31, 102–112 (2002).
50. Blomqvist, S., Gunnars, A. & Elmgren, R. Why the limiting nutrient differs between temperate coastal seas and freshwater lakes: A matter of salt. Limnol. Oceanogr. 49, 2236–2241 (2004).
51. Elser, J. J. et al. Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. Ecol. Lett. 10, 1135–1142 (2007).
52. Adina Paytan, K. M. The Oceanic Phosphorus Cycle. (2007) doi:10.1021/cr0503613.
53. McGrath, J. W., Chin, J. P. & Quinn, J. P. Organophosphonates revealed: new insights into the microbial metabolism of ancient molecules. Nat. Rev. Microbiol. 11, 412–419 (2013).
54. Neff, J. C., Holland, E. A., Dentener, F. J., McDowell, W. H. & Russell, K. M. The origin, composition and rates of organic nitrogen deposition: A missing piece of the nitrogen cycle? Biogeochemistry 57, 99–136 (2002).
55. Zhao, L. et al. The facilitating role of phycospheric heterotrophic bacteria in cyanobacterial phosphonate availability and Microcystis bloom maintenance. Microbiome 11, 1–16 (2023).
56. Zecher, K., Hayes, K. R. & Philipp, B. Evidence of Interdomain Ammonium Cross-Feeding From Methylamine- and Glycine Betaine-Degrading Rhodobacteraceae to Diatoms as a Widespread Interaction in the Marine Phycosphere. Front. Microbiol. 11, 533894 (2020).
57. Skorupska, A., Janczarek, M., Marczak, M., Mazur, A. & Król, J. Rhizobial exopolysaccharides: genetic control and symbiotic functions. Microb. Cell Fact. 5, 1–19 (2006).
58. Ettema, T. J. G. & Andersson, S. G. E. The alpha-proteobacteria: the Darwin finches of the bacterial world. Biol. Lett. 5, 429–432 (2009).
59. Sullivan, J. T. & Ronson, C. W. Evolution of rhizobia by acquisition of a 500-kb symbiosis island that integrates into a phe-tRNA gene. Proc. Natl. Acad. Sci. U. S. A. 95, 5145–5149 (1998).
60. Bouma-Gregson, K. et al. Impacts of microbial assemblage and environmental conditions on the distribution of anatoxin-a producing cyanobacteria within a river network. ISME J. 13, 1618–1634 (2019).
61. Guillard, R. R. L. & Lorenzen, C. J. Yellow-Green algae with chlorophyllide c. J. Phycol. 8, 10–14 (1972).
62. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460–2461 (2010).
63. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv 081257 (2016) doi:10.1101/081257.
64. Yilmaz, P. et al. The SILVA and 'All-species Living Tree Project (LTP)' taxonomic frameworks. Nucleic Acids Res. 42, D643–D648 (2013).
65. Bolyen, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37, 852–857 (2019).
66. ggplot2. https://ggplot2.tidyverse.org/.
67. Website. http://microbiome.github.io.
68. GUniFrac: Generalized UniFrac Distances, Distance-Based Multivariate Methods and Feature-Based Univariate Methods for Microbiome Data Analysis. Comprehensive R Archive Network (CRAN) https://CRAN.R-project.org/package=GUniFrac.
69. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models Usinglme4. J. Stat. Softw. 67, (2015).
70. Community Ecology Package [R package vegan version 2.6-4]. (2022).
71. MuMIn: Multi-Model Inference. Comprehensive R Archive Network (CRAN) https://CRAN.R-project.org/package=MuMIn.
72. Lenth, R. V. Estimated Marginal Means, aka Least-Squares Means [R package emmeans version 1.8.9]. (2023).
73. Palarea-Albaladejo, J. & Martín-Fernández, J. A. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics Intellig. Lab. Syst. 143, 85–96 (2015).

74. Mallick, H. et al. Multivariable association discovery in population-scale meta-omics studies. PLoS Comput. Biol. 17, e1009442 (2021).
75. Olm, M. R. et al. Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. mSystems 5, (2020).
76. GUniFrac: Generalized UniFrac Distances, Distance-Based Multivariate Methods and Feature-Based Univariate Methods for Microbiome Data Analysis. Comprehensive R Archive Network (CRAN) https://CRAN.R-project.org/package=GUniFrac.
77. Microbiome@GitHub. http://microbiome.github.io.
78. A, J. N. & N, F. J. Sickle: Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. BBMap short read aligner vol. 198 1–11 Preprint at https://doi.org/10.1016/J.TOXICON.2021.04.023 (7 2011).
79. Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420–1428 (6 2012).
80. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834 (2017).
81. Camargo, A. P. et al. Identification of mobile genetic elements with geNomad. Nat. Biotechnol. 1–10 (2023).
82. Chen, L. & Banfield, J. F. COBRA improves the quality of viral genomes assembled from metagenomes. bioRxiv 2023.05.30.542503 (2023) doi:10.1101/2023.05.30.542503.
83. Hyatt, D. et al. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 1–11 (3 2010).
84. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28, 27 (1 2000).
85. The UniProt Consortium et al. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res. 51, D523–D531 (2022).
86. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359 (2012).
87. Wu, Y. W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607 (2 2016).
88. Alneberg, J. et al. Binning metagenomic contigs by coverage and composition. Nat. Methods 11, 1144–1146 (9 2014).
89. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165 (2015).
90. Nissen, J. N. et al. Improved metagenome binning and assembly using deep variational autoencoders. Nature Biotechnology 2021 39:5 39, 555–560 (1 2021).
91. Sieber, C. M. K. et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 2018 3:7 3, 836–843 (5 2018).
92. 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 (7 2015).
93. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 11, 2864–2868 (7 2017).
94. 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 (3 2020).
95. Lee, M. D. GToTree: a user-friendly workflow for phylogenomics. Bioinformatics 35, 4162–4164 (2019).
96. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. Nucleic Acids Res. 50, D785–D794 (2021).

97. 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).
98. 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).
99. GitHub - wwood/CoverM: Read coverage calculator for metagenomics. GitHub https://github.com/wwood/CoverM.
100. UMAP. <https://CRAN.R-project.org/package=umap.
101. Olm, M. R. et al. InStrain enables population genomic analysis from metagenomic data and sensitive detection of shared microbial strains. Nat. Biotechnol. 39, 727 (2021).
102. Grün, B., Kosmidis, I. & Zeileis, A. Extended beta regression inR: Shaken, stirred, mixed, and partitioned. J. Stat. Softw. 48, (2012).
103. Hothorn, T., Bretz, F. & Westfall, P. Simultaneous inference in general parametric models. Biom. J. 50, 346–363 (2008).
104. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–62 (2016).
105. Zhou, Z. et al. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. Microbiome 10, 33 (2022).
106. Diamond, S. et al. Soils and sediments host Thermoplasmata archaea encoding novel copper membrane monooxygenases (CuMMOs). ISME J. 16, 1348–1362 (2022).
107. Schmartz, G. P. et al. PLSDB: advancing a comprehensive database of bacterial plasmids. Nucleic Acids Res. 50, D273–D278 (2022).