AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems

| Open Peer Review | Biotechnology | Research Article

Marine biofilms: cyanobacteria factories for the global oceans

Cheng Zhong,[1,2] Shun Yamanouchi,[3] Yingdong Li,[1,2] Jiawei Chen,[1,2] Tong Wei,[1,2] Ruojun Wang,[1,2] Kun Zhou,[1,2] Aifang Cheng,[1,2] Weiduo Hao,[4] Hongbin Liu,[1,2] Kurt O. Konhauser,[4] Wataru Iwasaki,[3,5] Pei-Yuan Qian[1,2]

AUTHOR AFFILIATIONS See affiliation list on p. 13.

ABSTRACT Marine biofilms were newly revealed as a giant microbial diversity pool for global oceans. However, the cyanobacterial diversity in marine biofilms within the upper seawater column and its ecological and evolutionary implications remains undetermined. Here, we reconstructed a full picture of modern marine cyanobacteria habitats by re-analyzing 9.3 terabyte metagenomic data sets and 2,648 metagenome-assembled genomes (MAGs). The abundances of cyanobacteria lineages exclusively detected in marine biofilms were up to ninefold higher than those in seawater at similar sample size. Analyses revealed that cyanobacteria in marine biofilms are specialists with strong geographical and environmental constraints on their genome and functional adaption, which is in stark contrast to the generalistic features of seawater-derived cyanobacteria. Molecular dating suggests that the important diversifications in biofilm-forming cyanobacteria appear to coincide with the Great Oxidation Event (GOE), "boring billion" middle Proterozoic, and the Neoproterozoic Oxidation Event (NOE). These new insights suggest that marine biofilms are large and important cyanobacterial factories for the global oceans.

IMPORTANCE Cyanobacteria, highly diverse microbial organisms, play a crucial role in Earth's oxygenation and biogeochemical cycling. However, their connection to these processes remains unclear, partly due to incomplete surveys of oceanic niches. Our study uncovered significant cyanobacterial diversity in marine biofilms, showing distinct niche differentiation compared to seawater counterparts. These patterns reflect three key stages of marine cyanobacterial diversification, coinciding with major geological events in the Earth's history.

KEYWORDS marine biofilms, cyanobacteria, metagenomics, biodiversity, evolution

Cyanobacterial genomes and their ecology and evolutions are emerging proxies to promote the understanding of geological events such as Earth's oxygenation (1). Since the Archean, cyanobacteria are likely among the most diverse and abundant microbes in the marine environment, colonizing vast swaths of the marine photic zone to sediments and other solid substrata in estuarine and coastal waters (1, 2). A previous experimental study of oxygenation from multiple benthic environments has suggested that benthic biomass played a crucial role in the initial rise of planetary oxygenation during the Archean, contributing significantly to the Great Oxidation Event (3). This finding was supported by phylogenetic and molecular clock analyses of cyanobacterial genomes available in the public databases, and the genome analyses further suggested that the emergence of planktonic cyanobacteria was associated with the Neoproterozoic Oxidation Event (4–6). Changes in oceanic chemistry and biosphere are plausible reasons for the dramatic divergence of the benthic and planktonic cyanobacteria (7, 8). However, the understanding of the full extent of how both benthic and planktonic cyanobacteria have interacted with Earth's evolutionary trajectory over a long geological period, such as the Proterozoic "boring billion" (1.8–0.8 Ga) remains limited.

Editor Michaeline B. N. Albright, Los Alamos National Laboratory, Los Alamos, New Mexico, USA

Address correspondence to Pei-Yuan Qian, boqianpy@ust.hk.

The authors declare no competing interests.

See the funding table on p. 14.

Received 25 March 2024
Accepted 6 September 2024
Published 15 October 2024

With the rapid increase in available genomic and metagenomic data, there is a significant opportunity to uncover new insights into cyanobacterial eco-evolutionary patterns and their interactions with Earth's atmospheric and oceanic changes (9–11). For example, *Tara Oceans* project analyzed 7.2 terabyte of metagenomic data to illuminate microbial diversity for global oceans (12).Following this, surface-associated microbes have increasingly been recognized for their profound historical and ongoing impact on the Earth's environment (13, 14). Within the surface-associated microbes' categories, more than 7,300 "species" identified in the marine biofilms are undetected in seawater, which implies marine biofilms are previously underappreciated niches of novel microbial species and functional potential (15). Marine biofilm microbes offer numerous ecological and evolutionary advantages, such as environmental protection, enhanced nutrient access, and increased interaction opportunities with other organisms, making them important forces for Earth's evolution (16). Despite these advancements, comprehensive analyses of cyanobacterial genomes within these metagenomic data sets are lacking.

Here, we reprocessed a previously established terabyte-scale metagenomic data set, comprising 101 marine biofilm and 91 seawater samples across Earth's oceans. Specifically, we quantified marine cyanobacteria diversity and their ecological characteristics to reveal if the marine biofilm serves as an important source of the taxonomic and functional diversity of marine cyanobacteria. Insights into phylogenetic relationships and the possible origin of the marine biofilm cyanobacteria were also obtained by reconstructing molecular-dated trees. Ultimately, we aim to use metagenomic data to elucidate detailed cyanobacterial divergences, their connections to key geological events, and highlight previously underexplored periods linking cyanobacterial evolution with Earth's geological history.

MATERIALS AND METHODS

Data source

The metagenomic data of the 101 biofilm and 91 seawater samples were described in previous studies (15, 17). Briefly, samples were originally collected from 10 marine regions (Fig. 1). For marine biofilm samples, three samples from Sapelo Island, North Atlantic Ocean; 12 samples from Red Sea; 70 samples from Hong Kong water, South China Sea; one sample from Yung Shu O Bay, South China Sea; and five samples from Zhuhai Xiangzhou Bay, South China Sea; five samples from South China Sea; and six samples from East China Sea. A total of 101 samples were developed from seven substrata, including Petri dish (70 samples), zinc panel (12 samples), aluminium panel (two samples), poly(ether-ether-ketone) panel (six samples), titanium panel (two samples), stainless steel panel (two samples), and rock panel (seven samples). The duration of the biofilm development has a range from 3 days to >30 days (rock panel). Besides, 91 of the 267 total seawater samples available in *Tara* Oceans, which were sampled nearby the marine biofilms, were also collected for parallel analysis. As reported previously, the sequencing depth of these biofilm and seawater samples had sufficient coverage for a thorough survey of entire microbial communities and allowed the comparison of these samples (15).

16S rRNA Illumina tag analysis

Taxonomic profiling

16S rRNA Illumina tags (miTags) were extracted from the unassembled metagenomic data, and the taxonomic classification adopted herein followed a previous study (15). The 16S rRNA miTags were mapped to a database that integrates GreenGenes with RDP and SILVA. OTUs with 97% similarity or above were selected. The 16S rRNA miTags predicted to the cyanobacteria phylum were selected, and the OTUs classified to chloroplasts were further trimmed off for all subsequent analyses owing to their eukaryotic origin. A Venn diagram implemented in R v.4.0.5 was used to assess the distribution of full OTUs across

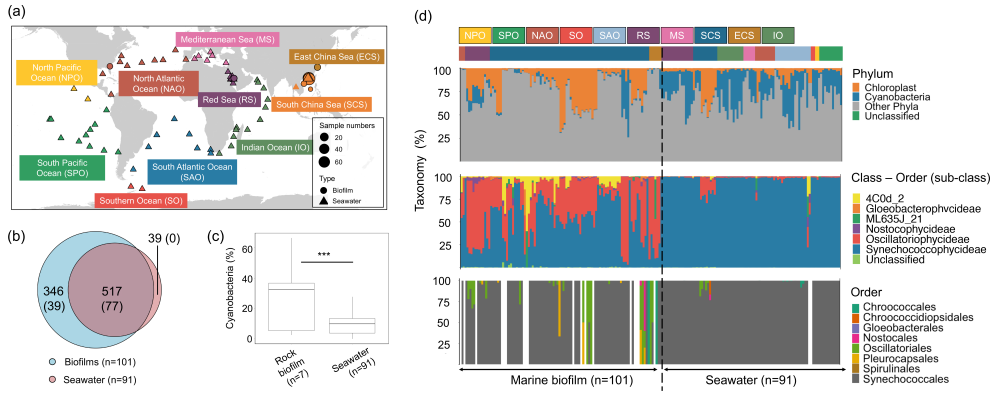

FIG 1 Cyanobacteria species across marine biofilm and seawater samples. (a) Sampling locations of the metagenomes. (b) Venn diagram showing the distribution of cyanobacterial 16S rRNA miTag OTUs and full-length 16S rRNA gene units (values presented inside parenthesis) across 101 biofilm and 91 epipelagic seawater of the total seawater samples at approximately equal size conditions. (c) Relative abundance of cyanobacteria on rock surfaces and seawater samples. (d) Phylum- and subclass-level (16S rRNA miTags) and class-level (protein marker genes) taxonomic profiling showing substantial differences in cyanobacterium community compositions between the biofilm and seawater samples. The color and text codes in the upper part of the taxonomic profiles refer to the oceans.

the 101 marine biofilm and 91 epipelagic seawater samples. The sequences of 101 marine biofilm and 91 epipelagic seawater samples were normalized to relative abundance for taxonomic profiling at the phylum, subclass, and genus levels. Metagenomic reads were also mapped to 40 universal single-copy marker genes to cross-validate the 16S rRNA miTag-based taxonomic profiling using mOTU v. 3.0.0.

Diversity, network, Mantel, and occupancy analyses

The metagenomic data of 101 marine biofilm and 91 epipelagic seawater samples were normalized to 10,000 sequences per sample for subsequent biogeographical analyses. Observed OTU, Chao1 richness, and ACE richness indices were used to compare community richness, whereas Shannon and inverse Simpson indices were utilized to compare community diversity. Jaccard distance was employed to compare OTU similarity. These diversity indicators were analyzed using phyloseq v.1.22.3 (18). The Bray–Curtis-based PCoA ordination with 95% confidence intervals was generated for dissimilarity analyses and pairwise comparison of marine biofilm and seawater cyanobacteria communities. The taxonomy-based PCoA ordination (Bray–Curtis distance) was generated to search for microbial species separating marine biofilm and seawater samples.

Network analyses of the OTUs associated with cyanobacteria from the normalized 16S rRNA miTags for the 101 biofilm and 91 epipelagic seawater samples were conducted using igraph (19) implemented in phyloseq. The Bray–Curtis distance and the maximum distance of 0.9 were used to build a network map. The Fruchterman–Reingold layout was utilized for network visualization. The objective distance and overall network topology, including average degree, closeness centrality, between centrality and eigenvector centrality, and between the biofilm and seawater samples, were compared.

Changes in the OTUs' co-presence numbers were analyzed across different sizes of the marine biofilm and seawater samples. With regard to changes in the biofilm and seawater samples, the linear regressions were used to assess their tendency from being generalists to specialists. The results of these analyses were confirmed by analyzing the non-normalized 16S rRNA miTags data set. Changes in the mean abundance of the OTUs were constructed as a function of their co-presence in the increased samples. The cutoff values for generalists and specialists were obtained from previous studies (20, 21). Cyanobacteria species were defined as generalists when they were present in >75% of the total sample size, whereas they were defined as specialists when they were present

in <25% of the total sample size with a relative abundance of >100 in log 10 scales. This applied classification aimed to distinguish the distribution of the generalists and specialists in the biofilm and seawater samples.

Partial Mantel correlations between the compositional data and the geographic distance (9,999 permutations) of the metagenomic data were computed using functions implemented in R. The Bray–Curtis distance and the Haversine distance were used for the compositional data and the geographic data, respectively. Then, seawater subsamples with coordinates similar to the biofilm samples, including samples falling in the intervals of the coordinates 0–2,500,000 km, 7,000,000–8,000,000 km, and 11,000,000–15,000,000 km, were collected to make the marine biofilm and seawater data sets more comparable to each other. Linear correlations were built for each data set. A Spearman-based matrix of correlation was also constructed across each phylum-level taxa for both samples by using the Hmisc package implemented in R (22).

Analysis of full-length 16S rRNA gene sequences

Distribution of full-length 16S rRNA gene units

The full-length sequences of 101 biofilms and 91 seawater samples were obtained using PhyloFlash v.3.0, and replicated sequences (100% similarity) were removed using CD-HIT (23). The method was adopted to build a Venn diagram as that for OTUs.

16S rRNA gene phylogenetic analysis

As for full-length 16S rRNA gene trees (posterior probability >80%), sequences were assembled from shotgun metagenomic reads of the 101 biofilm and 91 seawater samples and then clustered with 97% sequence identity to obtain 62 and 13 cyanobacterial representative sequences, respectively. The software used in this process was PhyloFlash v.3.0 with default settings, which was integrated with various types of software (e.g., SPAdes) and assigned the taxonomy on the basis of the modified versions of the SILVA SSU database (24). The sequences of the 16S rRNA genes associated with cyanobacteria, except for chloroplast, were selected for further analysis. This method minimized bias produced during the binning process of each metagenome sample, which supplemented useful results and additional validation for subsequent MAG-based analyses.

Alignment, Bayesian tree inference, divergence time estimation, and calibration

Briefly, Bayesian analysis of 16S rRNA genes was performed in three steps: first, inference of the tree topology, dating using the correction points indicated by a previous study, and finally estimation of the diversification rate along branches. We reconstructed a non-clock molecular phylogenetic tree to infer cyanobacterial 16S rRNA phylogenetic relationships (i.e., tree topology). The sequence data set consisted of 42 metagenome-derived cyanobacterial sequences clustered by 97% identity, 56 cyanobacterial sequences from NCBI GenBank, and three outgroup sequences (4). Sequences were aligned with MAFFT v.7.480 (25, 26) and subjected to the tree inference without gap trimming. Bayesian tree inference was performed with MrBayes v.3.2.7a (27) under the GTR + I + G4 model to obtain a 50%-majority rule consensus tree. Detailed methods to build the phylogenetic tree are presented in the supplemental material.

The divergence time of cyanobacteria was then estimated under a relaxed molecular clock. After removing three outgroup sequences from the data set, we again generated a multiple sequence alignment. Bayesian molecular clock analysis was conducted using BEAST v.2.6.3 (28). The GTR + I + G4 model was specified for the substitution model, and the clock rates were assumed to follow the uncorrelated log-normal distribution. Several constraints and calibration points were also imposed according to "Analysis 7" described in the previous study (4) after confirming that the phylogenetic tree obtained from the non-clock analysis was consistent with their trees for major subclades. MCMC runs were performed to obtain the maximum clade credibility tree. Note that we performed

another MCMC run in advance under a more moderate prior distribution to prepare a starting tree that satisfies their strict assumptions. For more details, see the supplemental material.

To improve the taxonomic resolution of seawater-derived cyanobacteria, we attempted the same analysis on a data set with metagenomic sequences clustered by 99% identity. However, we could not reconstruct the non-clock tree on this data set because MrBayes did not produce any properly converged MCMC run. The 97% identity sequence clusters and the 99% identity sequence clusters were compared to confirm that there was no discrepancy that would affect the calibration points or monophyletic constraints. Thus, the non-clock analysis was skipped, and BEAST clock analysis was directly performed to obtain a dated tree with a high resolution successfully.

Estimation of linage-specific diversification rates

The presence (or absence) of any phylogenetic disparity among cyanobacterial clades was tested by estimating lineage diversification rates via an established method (29), which modeled speciation, extinction, and change in rate categories as possible events along tree branches (the birth–death–shift process). Note that methods for modeling state-dependent diversification rates, such as the binary-state speciation and extinction (BiSSE) model, were not applicable to our data set because it contained too few state transitions that could be used to estimate parameters accurately.

The metagenomic samples were assumed to have uniformly sampled all extant cyanobacteria in the hydrosphere. On the basis of this assumption, the phylogenetic tree was pruned to include only metagenome-derived 16S rRNA clusters with 99% identity. With this phylogenetic tree as the input, Bayesian model inference was performed using RevBayes v.1.0.13. by following the tutorial (https://revbayes.github.io/tutorials/divrate/branch_specific.html, viewed on September 2021) (30). Detailed bioinformatic analyses and important assumptions are presented in the Supporting Information.

Analysis of metagenome-assembled genomes

Assembly, binning, and refinement

Quality-filtered sequences of 101 biofilms and 91 seawater metagenomes were assembled using metaWRAP v1.2, which integrates megahit and metaSPAdes, and further binned to MAGs using metaBAT2, CONCOCT, and MaxBin2 methods (31). Scaffolds ≥ 1,000 bp were retained for binning. MAGs with ≥50% completeness and ≤10% contamination scores were selected for refinement in metaWRAP. The refined medium- to high-quality MAGs were retained for downstream analyses (32). The process generated medium- to high-quality 77 marine biofilm-forming MAGs and 49 seawater MAGs.

Taxonomic and morphological classification and functional annotation

Functional capacities of cyanobacteria lineages were analyzed across 77 marine biofilm-forming MAGs and 49 seawater MAGs, as well as 60 cyanobacterial genomes of a previously reported data set (33). The taxonomic classification of the MAGs was analyzed using GTDB-Tk v.1.5.0 (34). The functions of MAGs and reference genomes were annotated using the DRAM software with default parameters (35). PCoA was performed to identify overall functional dissimilarities between 77 marine biofilm-forming MAGs and 49 seawater MAGs with completeness >50%–90% was investigated, and completeness >75% was selected to display. Subsequently, a detailed investigation of the presence and absence of functional genes across these MAGs was conducted. For important functional genes yet to be integrated into a known database, we built a database (https://github.com/jchenek/ref-seqs-chromatic-acclimation) for pigment type identification with the marker gene cpcBA, as well as genes implicated in chromatic acclimation genetic islands, including mpeZ, mpeY, mpeW, and mpeQ (36). Then, the coding sequence of the MAGs was predicted using Prodigal v.2.6.3 by using the "-p meta"

parameter (37) and annotated with Diamond v.2.0.4 (38) by using the "--sensitive, -k 1, -e 1e-20, and 1e-60" parameters against the customized database, and genes with coverage of less than 75% were discarded.

Phylogenetic analysis

To place the reconstructed biofilm-derived MAGs and seawater-derived MAGs into the phylogeny of cyanobacteria, we performed a phylogenetic analysis based on 27 widely used conservative genes for evolutionary analysis (39, 40). This method has higher accuracy of the phylogeny and timing of diversification compared to the phylogenetic analyses using only 16S rRNA genes. The reference sequences for these marker genes were downloaded from NCBI. The protein-coding genes of MAGs and reference genomes were predicted and extracted using Prodigal v2.6.3 (37). The extracted protein sequence of each coding gene was searched against the reference database of marker genes using PSI-BLAST with an e-value = $1e^{-05}$ and max_target_seqs = $1e^{7}$ to obtain the homologous genes in each genome (41). Subsequently, each set of homologous proteins was aligned using MAFFT v7.222 (42) and trimmed using trimAl v1.4 ((1, 43) using the auto options. A maximum likelihood (ML) phylogenetic tree was constructed using the IQ-TREE v2.1.2 (44). In addition, the best substitution model was searched and applied using the MFP + MERGE option, and a total of 1,000 ultrafast bootstrap replicates were sampled to evaluate the robustness of the phylogeny. The final results were visualized and rooted with iTOL v. 6.6 (45).

Molecular dating

The divergence time of cyanobacteria was estimated with MCMCTree v. 4.9e (39) on 27 conservative genes previously proposed to be valuable to date bacterial divergence (46) and cyanobacteria phylogenetic trees (33). Since molecular dating analysis is known to be intrinsically associated with calibration points (47), and the calibrations of cyanobacteria, such as the fitness of different molecular clock models, have been tested in the previous study (33), we followed it to use the independent rate model for molecular clock analysis of cyanobacteria in this study. The Bayesian molecular clock analysis has been run twice for verification with a burn-in of 50,000 and a total of 500,000 generations. Based on the well-established methodology for molecular dating of cyanobacteria, we were able to select the most precise estimates of cyanobacteria evolutionary timeline for illustration and further discussion. The tree was visualized in FigTree v.1.4.4. All files and procedures used for the molecular dating analysis were uploaded to the github (https://github.com/ylifc/Biofim_cyanobacteria_evolution).

Significant tests

For analyses associated with alpha-diversity and taxonomic profiling, ANOVA combined with TukeyHSD analysis was used to test whether the difference in diversity values observed between groups was significant. Similarly, ANOVA was used to test the significance of Principal Coordinate 1 (PC1) and Principal Coordinate 2 (PC2) used in PCoA. PERMANOVA was utilized to test the significance of PCoA results for the observed clusters. When PCoA was established, envift correlation implemented in Vegan v. 2.5–7 in R was employed to search for relevance between the tested samples/genomes and separate factors. In these statistical tests, $P < 0.05$ was regarded as a significant difference.

RESULTS

Taxonomic compositions

The niche differentiation was revealed by a Venn diagram analysis, in which 346 operational taxonomic units (OTUs) were unique to marine biofilms, which was nearly ninefold higher than the number of OTUs specific to seawater samples (Fig. 1b). The

distribution of the full-length 16S rRNA gene units was similar to that of the OTUs. While the comparison showed that the relative abundance of cyanobacteria in 101 marine biofilms was considerably lower than that in 91 seawater, the relative abundance of cyanobacteria in rock substrata of the biofilms was threefold higher than those in seawater samples (Fig. 1c). Cyanobacteria in marine biofilms on rock surfaces had a tenfold higher fraction (31.4% of the total community average) than those on other substrata (Fig. S1).

Further taxonomic analyses showed that the cyanobacteria communities in seawater and biofilms displayed a profound difference in taxonomic compositions (Fig. 1d). A taxonomic difference based on 16S rRNA miTag was revealed at the subclass level profiling: Oscillatoriophycideae (39.9% of total community on average) and Nostocophycideae (1.29% of total community on average) were major lineages in biofilms, whereas Synechococcophycideae (95.1% of the total cyanobacteria on average) were dominant in seawater samples (for the difference at the genus level, see Fig. S2). On average, the relative abundance of the cyanobacteria (excluding Chloroplasts) and the taxonomic differentiation in biofilms (e.g., Oscillatoriales), as represented by 16S rRNA miTag, were in line with that using protein marker genes (Fig. S3). However, the differentiation depicted by the protein-coding marker was less than that resolved by 16S rRNA-based classification, and the stricter classification resulted in lower detection of cyanobacteria of sub-levels in some biofilm and seawater samples.

Diversity analysis

The local diversity (i.e., alpha-diversity) of biofilm-forming cyanobacteria, including richness and diversity, was significantly ($P < 0.05$) lower than that of cyanobacteria in seawater samples. The Jaccard distance-based OTU composition dissimilarity, which represents the regional diversity of cyanobacteria in biofilms, was significantly higher ($P < 0.05$). Selected diversity indices are presented in Fig. 2a, and all measured values are presented in Table S1.

For beta-diversity analysis, the cyanobacterial community displayed a significant ($P < 0.001$) difference in principal coordinate analysis (PCoA) of the normalized OTU compositions for 10,000 sequences per sample (Fig. 2b). Results of further analyses show Synechococcophycideae and Oscillatoriophycideae species are key cyanobacterial lineages in separating marine biofilms and seawater samples on PCoA (summarized in Fig. S4).

Generalists and specialists

In the OTU-based network analysis, the biofilm samples were clustered geographically, while seawater samples were tightly connected to each other regardless of their origins (Fig. S5, Table S2). The average betweenness centrality of biofilm samples (160) was twofold higher than that of seawater samples, but the average degree (32) and the eigenvector centrality (0.14) were lower than the indices (0.63 and 0.67) of seawater. These centrality properties for an individual are shown in Figs. S6 through S9.

The notion of less biogeographic connectivity of marine biofilm cyanobacteria relative to seawater counterparts was supported by Mantel tests with a distance of up to 15,000 km (Fig. S10). With comparable data sets, we observed a stronger correlation between geographic distance and cyanobacterial composition dissimilarities in the biofilm samples ($R^2 = 0.29$) than in the seawater samples ($R^2 = 010$). This demonstrated that the biofilm-forming cyanobacteria had a stronger distance decay than the seawater cyanobacteria.

With an adopted definition (see Materials and Methods), seawater contained both specialists and generalists, while biofilms only contained specialists, and the specialists in biofilms generally had a higher abundance than those in seawater (Fig. 3a and b). As also evident in Fig. S11, the steeper decreasing trend of the same cyanobacteria OTU in increasing marine biofilms ($R^2 = 0.70$) than seawater ($R^2 = 0.95$) suggests that the cyanobacteria OTUs in seawater were simultaneously present in more samples than the

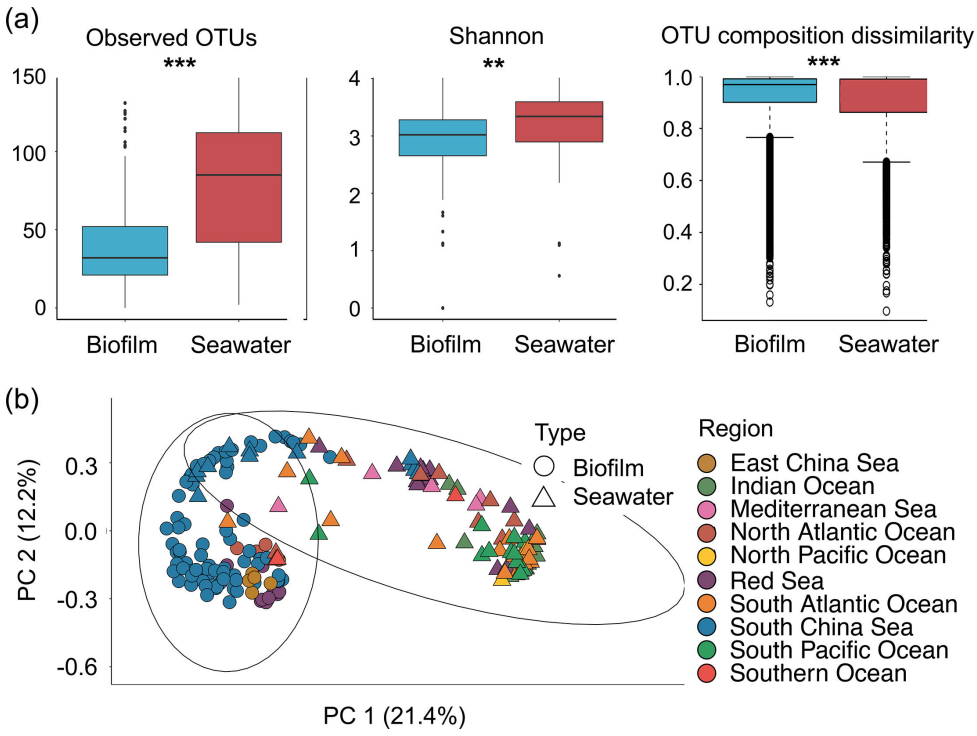

FIG 2 Diversity analysis between marine biofilm and seawater samples. (a) Bray–Curtis distance of the cyanobacterial communities shown by principal coordinate analysis (PCoA) of the normalized OTU matrix; 95% confidence intervals for the biofilm and seawater samples are present in the PCoA ordination. (b) Comparison of observed OTUs, Shannon diversity, and Jaccard distance-based OTU dissimilarity.

biofilm OTU (i.e., the same cyanobacteria OTU is present in more seawater samples than it could be present in the biofilm sample). More detailed analyses showed that the relative abundance of cyanobacteria is positively correlated to biofilm development duration, and a high abundance of cyanobacteria was found in the samples collected in spring and summer (Fig. S12). Biofilm-forming cyanobacteria in the Red Sea displayed a more diverse and niche partitioning pattern than seawater counterparts (Fig. S13, also see the detailed results and discussion in the supplemental material).

MAG characteristics and functional analysis

In total, we generated 1,047 marine biofilm MAGs and 1,601 seawater MAGs of medium–high quality (32); among these MAGs, 77 biofilms and 49 seawater MAGs were classified as cyanobacteria. This result substantially increased the total number of cyanobacteria MAGs derived from marine biofilms by eight times compared to a previous study, which used the same sequences to illustrate the overall biodiversity in marine biofilms (15). The detailed properties of newly generated MAGs are presented in the supplemental material. A broader comparison of the MAG sequences also showed substantial differences in taxonomy (Fig. 4a) and functional potential (Fig. 4b) of cyanobacteria across marine biofilms and seawater. Sequences of biofilm-forming cyanobacteria were assigned to diverse taxonomy, while seawater-derived cyanobacteria were mostly assigned to *Synechococcus* and *Prochlorococcus*. Based on the phylogenomic tree of cyanobacteria been referred from two sets of reference genomes (6, 33), the biofilm-forming MAGs covered the phylogenetic diversity of different evolutionary stages, especially the deep branches of this phylum, whereas seawater-derived MAGs mainly formed clusters in shallow branches. Furthermore, with the support of the bootstrap values, these phylogenetic trees consistently demonstrate the novelty of phylogeny of the cyanobacterial lineages in biofilm-derived MAGs (Fig. 4c; Fig. S14).

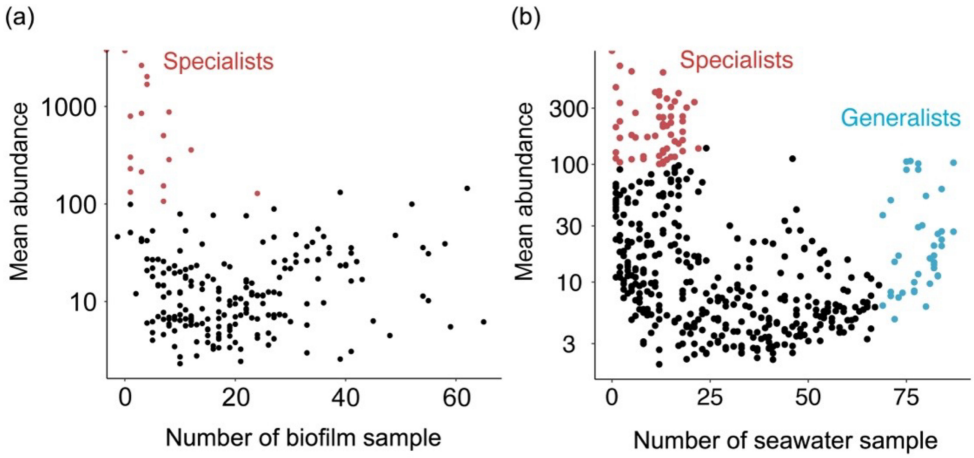

FIG 3 Occupancy patterns of the cyanobacterial communities across (a) biofilm and (b) seawater samples. Abundance (y-axis) and occupancy (x-axis) plots for the 16S rRNA miTags OTU. Habitat generalist OTUs (in blue) are defined in >75% of the total sample size, and specialist OTUs (in red) are defined in <25% of the total sample size with a relative abundance of >100 in log 10 scales.

The difference in functional gene contents between 43/77 biofilm-derived and 13/49 seawater-derived cyanobacteria was presented on the PCoA for MAGs with completeness >75% (Fig. 4d). The composition of functional genes of the biofilm MAGs with decreasing completeness tended to be similar to the composition of the seawater MAGs, which is consistent with the observation of substantial reductions in genome size and functional richness from biofilm MAGs to seawater counterparts (Fig. 4d). Completeness, contamination, and GC content of all MAGs are presented in Fig. S15. The major difference in the functional capacity between biofilm MAGs and seawater MAGs was that the biofilm MAGs generally had more complete functional genes related to Fe and Zn compared to seawater MAGs (Fig. 4d). The difference of other transport genes for oxygen-sensitive elements such as Ni and Mn (for Ni and Mn transport gene annotation, see Fig. S15), or nutrients such as P, between biofilm-forming and seawater MAGs was not observed. With regards to chromatic acclimation (light adaptation), seawater MAGs housed both pigment type 3 c and pigment type 3d genes, whereas pigment genes were not detected in any biofilm MAGs (Fig. 4d).

Molecular dating, diversification rate, and links with geological events

Bayesian molecular dated phylogenetic trees by using MAGs (Fig. 5a, see Fig. S16 for a full version tree) and full-length 16S rRNA genes (Fig. 5b, see Figs. S17 and S18 for full version trees) showed that marine biofilm-forming cyanobacteria evolution may be coincident with major changes in Earth's past climate. The OTU prevalence (Fig. 5b) is consistent with the distribution of MAG-based taxonomy (Fig. 4a), suggesting that a few genera of close phylogeny dominate seawater cyanobacteria. The age estimates at critical branching points were consistent between the two trees, but the MAG-based tree provided a higher phylogenetic resolution to identify previously unrecognized cyanobacteria diversification events.

Three critical time points, 2.5 Ga, 1.3 Ga, and 0.8 Ga, were identified for marine cyanobacteria evolution. The former coincides with the Great Oxidation Event (GOE) when Earth's atmosphere first became oxygenated (48). Cyanobacteria derived from marine biofilm MAG show them to be the first lineage among all studied genomes after the occurrence of GOE (Fig. 5a). This was followed at 1.3 Ga by a significant diversification of biofilm cyanobacterial lineages (Fig. 5a). Despite that, there is an interesting missing diversification activity of biofilm-forming cyanobacteria between GOE and 1.3 Ga. This is surprising given that this period of time includes two major oxygen-related events, the "overshot of oxygen" associated with the Lomagundi Event between 2.2 and 2.06 Ga

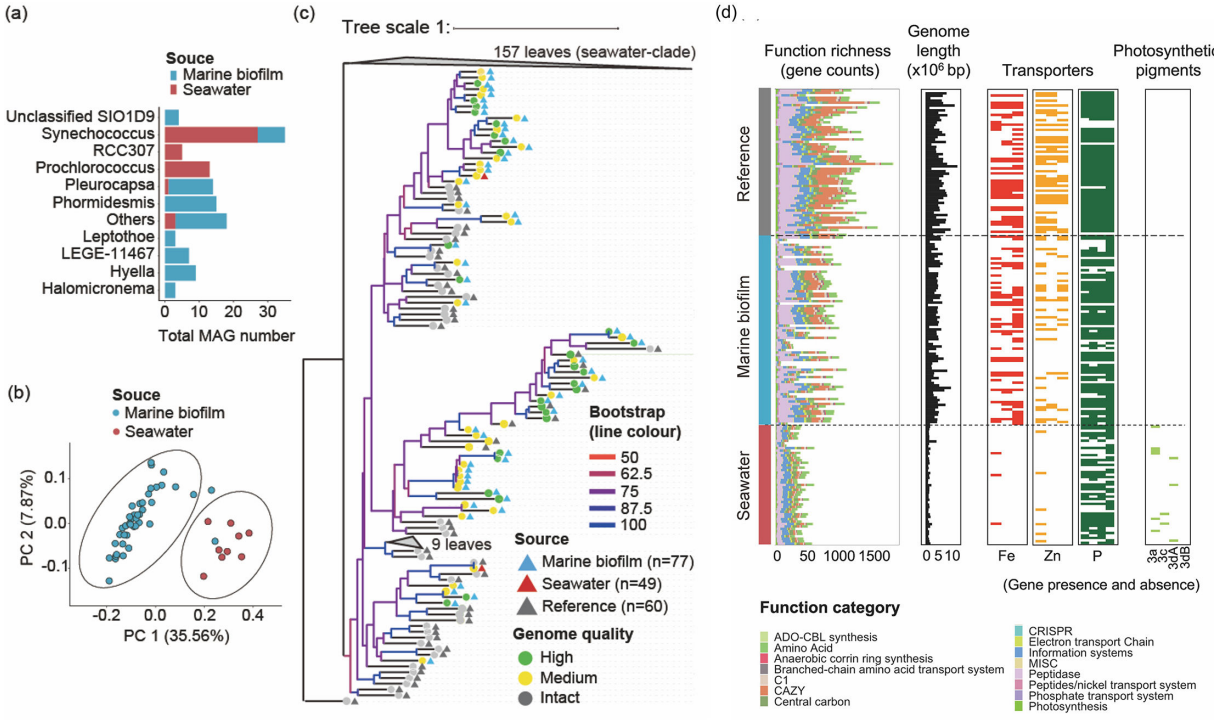

FIG 4 The novelty, characteristics, and functional analyses of 126 generated metagenome-assembled genomes (MAGs). The overall distribution of (a) GTDB-Tk (34) taxonomic assignments (total MAG numbers < 3 is defined as "others") and (b) functional genes for 77 cyanobacterial MAGs from marine biofilms and 49 cyanobacterial MAGs from seawater. The PCoA ordination is made by Bray–Curtis distance for biofilm and seawater MAGs > 75% completeness based on the sum of functional genes classified into carbon utilization, transporters, energy, organic nitrogen, and miscellaneous functional categories by using the DRAM software (35). (c) Phylogenetic tree of MAGs across marine biofilms and seawater with bootstrap values and morphology characterization. Branches with bootstrap values < 50 were collapsed. (d) MAG size and functional gene comparison with a focus on biological essential element transporters and photosynthesis adaptation across all MAGs and reference genomes. Functions with the top 15 abundant genes are retained for functional profiling overview.

where atmospheric O_2 levels may have increased above 10% of present atmospheric levels, and then there is a drop in atmospheric O_2 at 2.0 Ga (49). The third critical point marks the latest large diversification of marine biofilm cyanobacteria that occurred soon after the Neoproterozoic Oxidation Event (NOE) at about 0.8 Ga, which was in coincidence with the emergence and fast diversification of the seawater-derived MAGs (Fig. 5a). By estimating the lineage-specific diversification rates, we observed a significant increase in the speciation rate at the largest subclade of the seawater-derived cyanobacteria, which is coincident with the NOE (Fig. 5b).

DISCUSSION

Cyanobacterial biodiversity

We demonstrate that marine biofilms act as critical cyanobacteria factories in global oceans, underpinning complex biosphere and marine environments throughout history. This study revealed an overall picture of metacoupling of the evolutional, functional, and ecological patterns of marine cyanobacteria and critical geological events (Fig. 6). Compared with the analysis of isolated genomes, metagenomic analysis has made it possible to examine unculturable genomes and further link the genomes of individuals/groups to specific environments in an ecological context (50, 51). While exploration of marine biofilms has dramatically increased the known microbial species (15), with improved bioinformatic processing methods, this study further promotes these insights as represented by high-quality MAGs of cyanobacteria. The Venn diagram also supports that marine biofilms may hold a large and unique, underappreciated cyanobacterial

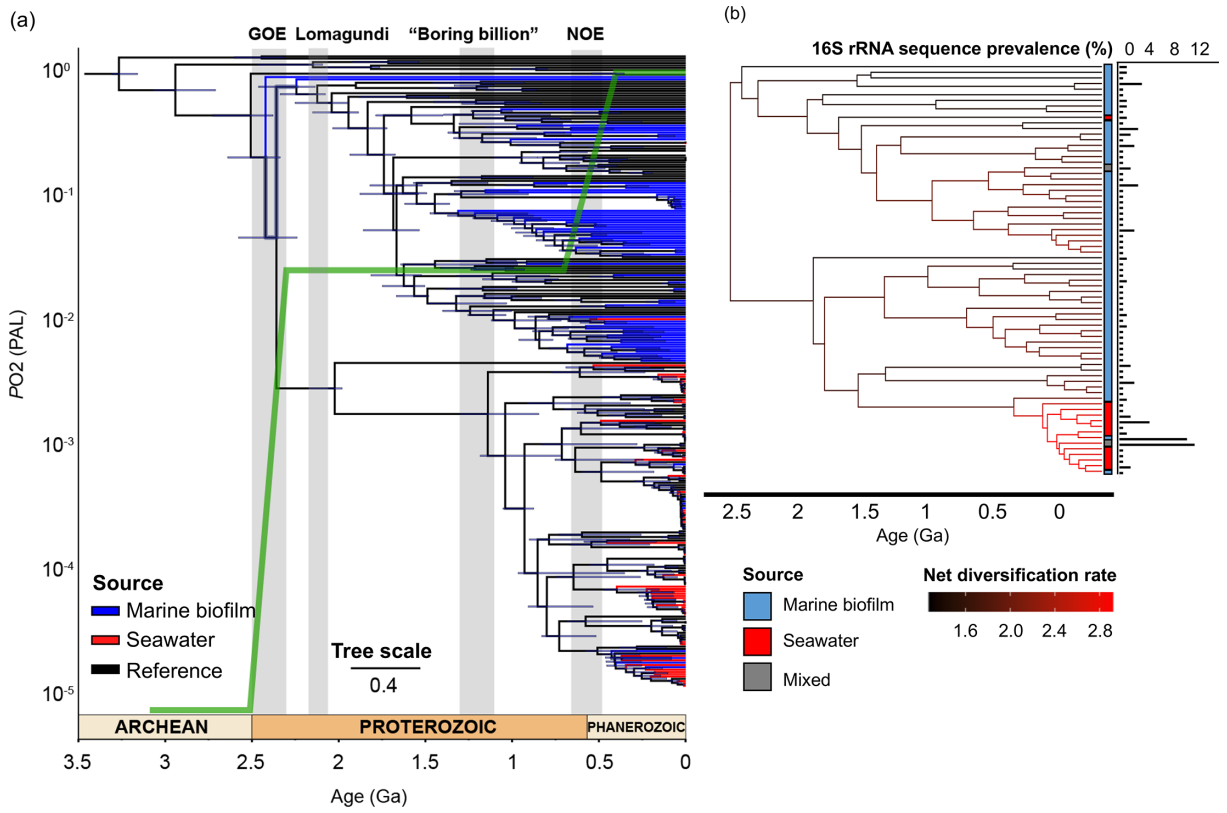

FIG 5 Bayesian molecular dated phylogenetic tree of cyanobacteria across marine biofilm and seawater samples. (a) Molecular dated phylogenetic tree of 77 MAGs derived from 101 marine biofilms and 49 MAGs derived from 91 seawater samples from global oceans with 60 references of full genome sequences (33). The MAG source is represented by the colored branches, and the diversification intensity for cyanobacteria from different sources is indicated by the color density. (b) Full-length 16S rRNA gene-based molecular dated phylogenetic tree of 62 sequences extracted from metagenomes of biofilm samples and 13 sequences from metagenomes of seawater samples with 56 references of 16S rRNA gene sequences (4); an estimation of changes in net diversification rates in major tree branches was done. Geological events associated with stepwise oxygenation (the green line), as represented by atmospheric oxygen (PAL), and their estimated occurrence time refer to previous studies (48).

abundance and diversity for coastal regions and estuaries compared to seawater. The implementation of full-length 16S rRNA genes (100% similarity) in the Venn diagram has mitigated the bias associated with planktonic cyanobacteria with highly similar genomes, such as *Prochlorococcus* data. Compared with other forms of benthic life, marine biofilm development is highly dynamic (52). The cyanobacteria functioning and their carbon export activity in the future ocean may be altered in the contemporary ocean ecosystem that undergoes rapid changes in trace metal contents due to the growing human population and industrialization (53). Thus, the results indicate that the unknown cyanobacterial population remains large, and it is critically important for re-assessments of primary production and other biogeochemical budgets of the oceans (54, 55).

Cyanobacterial adaptation represented by oceanic niche differentiation

Cyanobacterial adaption strategies are diverse, as previously recorded by *Synechococcus* and *Prochlorococcus* (56–58). The newly explored niche differentiation of cyanobacteria contributes to further understanding of cyanobacterial adaptation for global oceans. Analyses indicate more geographical constraints imposed on biofilm-forming cyanobacteria than their seawater counterparts and biofilm-forming lineages were less dispersed in the oceans. Generalists may promote the evolution of a microbial community (21), and the diversification rate estimation (Fig. 5b) fits well with this recent generalist–specialist evolution model, suggesting that marine biofilm-derived cyanobacteria may optimally

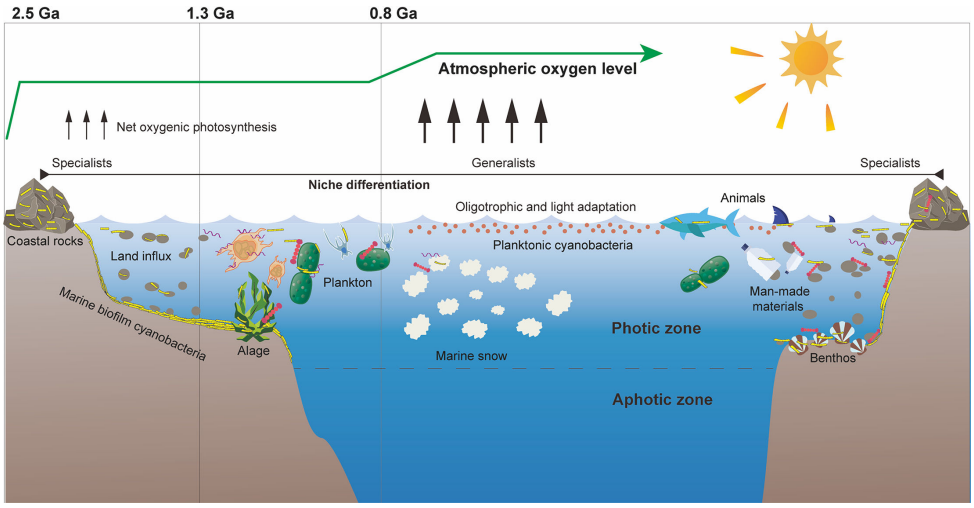

FIG 6 A concept model of the marine biofilms as cyanobacterial factories in the marine environment. This figure elucidates marine cyanobacteria ecology—evolution nexus across oceanic niches, spatial and geological time scales on the basis of the metagenomic data analysis, highlighting the seeds and diversification of cyanobacteria to the oceans through marine biofilms at critical time intervals at 2.5 Ga, 1.3 Ga, and 0.8 Ga, and various vectors (e.g., coastal rocks, suspended particles, animal surface, and seafloor) that may provide a surface for marine biofilms for shielding large cyanobacterial biomass in the ancient and modern oceans.

function in geographically constrained regions; in contrast, high diversification rates of seawater-derived cyanobacteria have a significant role in facilitating cell dispersion across environmental barriers (59). The observed reduced functional richness and genome size align with the reductive genome evolution phenomena (genomic streamlining) prevalent in marine bacterioplankton (60), and for marine planktonic cyanobacteria specifically, with new characteristics such as decreased cell diameter and loss of filamentous forms, and then evolved and spread across the oceans (61, 62). Our results show that the most significant differences in functional potential are related to various transporters (e.g., Fe), which may be caused by the uneven distribution of nutrients for global oceans and/or continuous changes in oceanic geochemistry throughout Earth's history. These ecological and functional potential patterns infer that "geo-bio" co-evolution may significantly determine the development of the biofilm-forming cyanobacteria.

Cyanobacteria evolution pattern in marine biofilms

Our results suggest that the newly explored cyanobacterial diversity is closely related to the evolution of cyanobacteria throughout Earth's history, and the reconstructed picture clearly displays the complexity of the marine cyanobacteria evolution. The rock record shows a substantial rise of planetary oxygenation toward the end of Archean, the so-called Great Oxidation Event, or GOE. It is widely accepted that a substantial population of planktonic cyanobacteria in the Archean water column was already required to generate that amount of oxygen (7, 63). Based on our result, the seeding of early planktonic cyanobacteria may have come from marine biofilms around the GOE, either growing on the seafloor in near-shore coastal settings and/or attached to increased flux of oceanic particles in seawater (3, 64, 65). Ancient cyanobacterial genomes older than 2.5 Ga were not found (for more discussion on geochemical aspects, see the supplemental material) (66).

The evolutionary trajectory of cyanobacteria challenges the view of slow biological evolution in the Earth's middle Proterozoic between 1.8 Ga and 0.8 Ga, that is, the so-called "boring billion", since there was a major diversification at 1.3 Ga (Fig. 5A). Our data suggest that the second remarkable diversification of marine biofilm-derived

cyanobacteria was followed by the emergence of eukaryotes that inhabited the world's oceans as early as 1.8–1.6 Ga, which was recorded by fossil records (7) and likely to be driven by the endosymbiotic event established ~1.9–1.25 Ga (63, 67). From a biological perspective, forming biofilms may have significant advantages in establishing a symbiotic relationship and other interactions with eukaryotes(52).

The long lag for planktonic cyanobacteria to evolve from marine biofilms/associated symbionts may essentially link to declining metal availability in Mesozoic seawater (8, 68, 69), as evidenced by the functional divergence herein, such as the minimized requirement for Fe and Zn transporters in seawater cyanobacteria (Fig. 4D) and the minimized proteomic requirement of these trace metals recorded elsewhere (55). This is also evidenced by the wide detection of cyanobacterial lineages located in the early branching points on the evolutionary tree in the Red Sea samples, as these samples are derived from upper water columns of the brine pools with uniquely high concentrations of metals (70–72). Cyanobacteria evolution is multidirectional in the modern oceans, as our data also suggest a previously overlooked and fast evolution event of marine biofilm-derived cyanobacteria after NOE.

Limitations and future steps

Our results highlight the importance of continuously exploring the unknown cyanobacterial population in marine biofilms and other underexplored niches in the marine environment. Geographical and substrata unevenness of sampling may cause bias in the results. For example, most of the biofilms are from SCS, and the seawater samples from SCS are much closer to the SCS biofilm samples than the others. More marine biofilm sampling worldwide is required. While we focused on the overall picture of the cyanobacteria in the marine biofilms, many details on the functional potential and bio-geo interactions are yet to be explicated. Also note that the molecular-dated trees adopted numerous methods previously established, and in fact, many areas, such as the key timing of cyanobacterial diversification, remain controversial. As for interesting patterns revealed in our results, implementing alternative clock models is the next step to constrain the important evolution timing. Nevertheless, given the prevalence of marine biofilms, this study shall serve as a foundation for further investigation of the critical roles of cyanobacteria and other key microbial participants in Earth's changing oceans and climates.

ACKNOWLEDGMENTS

We thank Dr. R. Zhang and Dr. M. Matsui, who provided technical assistance and invaluable comments on the manuscript.

This research was supported by the Major Project of Basic and Applied Basic Research of Guangdong Province (2019B030302004), Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) 880 (2021HJ01, SMSEGL20SC01) awarded to P.-Y.Q. W.I. was supported by Japan Science and Technology Agency (JPMJCR19S2) and Japan Society for the Promotion of Science (19H05688 and 18H04136). The work described in this paper was supported by a fellowship award from the Research Grants Council of the Hong Kong Special Administrative Region, China (HKUST PDFS2223-6S03), awarded to C.Z.

C.Z. wrote the paper with input from all authors. S.Y. performed 16S rRNA-based phylogenetic analyses, and Y.L. performed MAG-based phylogenetic analyses. J.C. performed cell counting, A.C. performed SEM imaging, and T.W., R.W., and K.Z. performed data processing. S.Y., W.H., H.L., W.I., K.O.K., and P.-Y.Q. edited the paper. C.Z. and P.-Y.Q. designed the study.

AUTHOR AFFILIATIONS

[1]Department of Ocean Science, The Hong Kong University of Science and Technology, Hong Kong, China

[2]Southern Marine Science and Engineering Guangdong Laboratory, Guangzhou, China
[3]Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo, Japan
[4]Department of Earth and Atmospheric Sciences, Faculty of Science, University of Alberta, Edmonton, Alberta, Canada
[5]Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Chiba, Japan

AUTHOR ORCIDs

Cheng Zhong http://orcid.org/0000-0002-8709-7444
Hongbin Liu https://orcid.org/0000-0002-3184-2898
Pei-Yuan Qian http://orcid.org/0000-0003-4074-9078

FUNDING

Funder	Grant(s)	Author(s)
Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (南方海洋科学与工程广东省实验室)	2021HJ01, SMSEGL20SC01	Pei-Yuan Qian
MEXT | Japan Science and Technology Agency (JST)	JPMJCR19S2	Wataru Iwasaki
MEXT | Japan Society for the Promotion of Science (JSPS)	19H05688, 18H04136	Wataru Iwasaki
Research Grants Council, University Grants Committee (研究資助局)	HKUST PDFS2223-6S03	Cheng Zhong

AUTHOR CONTRIBUTIONS

Cheng Zhong, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review and editing.

DATA AVAILABILITY

101 biofilm metagenomes and 24 adjacent seawater metagenomes refer to the publicly available data sets in the NCBI database under BioProject accession no. PRJNA438384 (BioSample no. SAMN08714533 for the 101 biofilm metagenomes; BioSample no. SAMN08714535 and SAMN08714533 for the 24 adjacent seawater metagenomes); 67 Tara Oceans seawater samples refer to the publicly available data sets in the EBI under the project identifiers PRJEB402 and PRJEB7988. Additional data are available in Figshare (https://doi.org/10.6084/m9.figshare.21485238.v2) as follows. The metadata of the metagenomes re-analyzed in this study are presented in Supplementary Data A. The data of 16S rRNA miTags taxonomic, diversity, and ecological analyses are provided in Supplementary Data B-D. The data and accession number of 77 marine biofilm-derived cyanobacterial MAGs and 49 seawater-derived cyanobacterial MAGs are provided, and their detailed properties are given in Supplementary Data E. The accession numbers of each of the reference sequences and genomes used for the 16S rRNA-based evolutionary analyses and the MAG-based evolutionary analyses are listed in Supplementary Data F. A full result of functional gene annotation of MAGs is listed in Supplementary Data G.

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Supporting information (mSystems00317-24-S0001.docx). Fig. S1-S18, Tables S1 and S2, and additional experimental details.

Open Peer Review

PEER REVIEW HISTORY (review-history.pdf). An accounting of the reviewer comments and feedback.

REFERENCES

1. Sánchez-Baracaldo P, Bianchini G, Wilson JD, Knoll AH. 2022. Cyanobacteria and biogeochemical cycles through Earth history. Trends Microbiol 30:143–157. https://doi.org/10.1016/j.tim.2021.05.008
2. Sessions AL, Doughty DM, Welander PV, Summons RE, Newman DK. 2009. The continuing puzzle of the Great Oxidation Event. Curr Biol 19:R567–74. https://doi.org/10.1016/j.cub.2009.05.054
3. Lalonde SV, Konhauser KO. 2015. Benthic perspective on Earth's oldest evidence for oxygenic photosynthesis. Proc Natl Acad Sci USA 112:995–1000. https://doi.org/10.1073/pnas.1415718112
4. Schirrmeister BE, de Vos JM, Antonelli A, Bagheri HC. 2013. Evolution of multicellularity coincided with increased diversification of cyanobacteria and the Great Oxidation Event. Proc Natl Acad Sci USA 110:1791–1796. https://doi.org/10.1073/pnas.1209927110
5. Blank CE, Sánchez-Baracaldo P. 2010. Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. Geobiology 8:1–23. https://doi.org/10.1111/j.1472-4669.2009.00220.x
6. Schirrmeister BE, Gugger M, Donoghue PCJ. 2015. Cyanobacteria and the Great Oxidation Event: evidence from genes and fossils. Palaeontology 58:769–785. https://doi.org/10.1111/pala.12178
7. Knoll AH. 2014. Paleobiological perspectives on early eukaryotic evolution. Cold Spring Harb Perspect Biol 6:a016121. https://doi.org/10.1101/cshperspect.a016121
8. Robbins LJ, Lalonde SV, Planavsky NJ, Partin CA, Reinhard CT, Kendall B, Scott C, Hardisty DS, Gill BC, Alessi DS, Dupont CL, Saito MA, Crowe SA, Poulton SW, Bekker A, Lyons TW, Konhauser KO. 2016. Trace elements at the intersection of marine biological and geochemical evolution. Earth Sci Rev 163:323–348. https://doi.org/10.1016/j.earscirev.2016.10.013
9. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol 2:1533–1542. https://doi.org/10.1038/s41564-017-0012-7
10. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi A, Gibbons SM, Ackermann G, et al. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551:457–463. https://doi.org/10.1038/nature24621
11. Nayfach S, Roux S, Seshadri R, Udwary D, Varghese N, Schulz F, Wu D, Paez-Espino D, Chen I-M, Huntemann M, et al. 2021. A genomic catalog of Earth's microbiomes. Nat Biotechnol 39:499–509. https://doi.org/10.1038/s41587-020-0718-6
12. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al. 2015. Structure and function of the global ocean microbiome. Science 348. https://doi.org/10.1126/science.1261359
13. Flemming HC, Wuertz S. 2019. Bacteria and archaea on Earth and their abundance in biofilms. Nat Rev Microbiol 17:247–260. https://doi.org/10.1038/s41579-019-0158-9
14. Flemming HC, Wingender J. 2010. The biofilm matrix. Nat Rev Microbiol 8:623–633. https://doi.org/10.1038/nrmicro2415
15. Zhang W, Ding W, Li YX, Tam C, Bougouffa S, Wang R, Pei B, Chiang H, Leung P, Lu Y, Sun J, Fu H, Bajic VB, Liu H, Webster NS, Qian PY. 2019. Marine biofilms constitute a bank of hidden microbial diversity and functional potential. Nat Commun 10:1–10. https://doi.org/10.1038/s41467-019-08463-z
16. Dang H, Lovell CR. 2016. Microbial surface colonization and biofilm development in marine environments. Microbiol Mol Biol Rev 80:91–138. https://doi.org/10.1128/MMBR.00037-15
17. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, et al. 2015. Structure and function of the global ocean microbiome. Science 348:1–10. https://doi.org/10.1126/science.1261359
18. McMurdie PJ, Holmes S. 2013. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8:e61217. https://doi.org/10.1371/journal.pone.0061217
19. Csárdi G, Nepusz T. 2006. The Igrph software package for complex network research. Comp Syst. https://doi.org/10.5281/zenodo.7682609
20. Barberán A, Bates ST, Casamayor EO, Fierer N. 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J 6:343–351. https://doi.org/10.1038/ismej.2011.119
21. Sriswasdi S, Yang CC, Iwasaki W. 2017. Generalist species drive microbial dispersion and evolution. Nat Commun 8:1162. https://doi.org/10.1038/s41467-017-01265-1
22. Harrell F. 2024. Hmisc: Harrell Miscellaneous. R package version 5.1-4. https://github.com/harrelfe/hmisc.
23. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659. https://doi.org/10.1093/bioinformatics/btl158
24. Gruber-Vodicka HR, Seah BKB, Pruesse E. 2020. phyloFlash: rapid small-subunit rRNA profiling and targeted assembly from metagenomes. mSystems 5:e00920-20. https://doi.org/10.1128/mSystems.00920-20
25. Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066. https://doi.org/10.1093/nar/gkf436
26. Katoh K, Standley DM. 2013. MAFFT: iterative refinement and additional methods. Mult Seq Align Methods 1079:131–146. https://doi.org/10.1007/978-1-62703-646-7_8
27. Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574. https://doi.org/10.1093/bioinformatics/btg180
28. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. 2018. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol 4:vey016. https://doi.org/10.1093/ve/vey016
29. Höhna S, Freyman WA, Nolen Z, Huelsenbeck JP, May MR, Moore BR. 2019. A Bayesian approach for estimating branch-specific speciation and extinction rates. bioRxiv. https://doi.org/10.1101/555805
30. Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Syst Biol 65:726–736. https://doi.org/10.1093/sysbio/syw021
31. Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6:158. https://doi.org/10.1186/s40168-018-0541-1
32. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, et al. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat Biotechnol 35:725–731. https://doi.org/10.1038/nbt.3893
33. Zhang H, Sun Y, Zeng Q, Crowe SA, Luo H. 2021. Snowball Earth, population bottleneck and *Prochlorococcus* evolution. Proc Biol Sci 288:20211956. https://doi.org/10.1098/rspb.2021.1956
34. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848

35. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, Liu P, Narrowe AB, Rodríguez-Ramos J, Bolduc B, Gazitua MC, Daly RA, Smith GJ, Vik DR, Pope PB, Sullivan MB, Roux S, Wrighton KC. 2020. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Microbiology. https://doi.org/10.1101/2020.06.29.177501
36. Grébert T, Nguyen AA, Pokhrel S, Joseph KL, Ratin M, Dufour L, Chen B, Haney AM, Karty JA, Trinidad JC, Garczarek L, Schluchter WM, Kehoe DM, Partensky F. 2021. Molecular bases of an alternative dual-enzyme system for light color acclimation of marine *Synechococcus* cyanobacteria. Proc Natl Acad Sci USA 118:e2019715118. https://doi.org/10.1073/pnas.2019715118
37. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119
38. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176
39. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088
40. Petitjean C, Deschamps P, López-García P, Moreira D. 2014. Rooting the domain archaea by phylogenomic analysis supports the foundation of the new kingdom Proteoarchaeota. Genome Biol Evol 7:191–204. https://doi.org/10.1093/gbe/evu274
41. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. https://doi.org/10.1093/nar/25.17.3389
42. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
43. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348
44. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
45. Letunic I, Bork P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49:W293–W296. https://doi.org/10.1093/nar/gkab301
46. Battistuzzi FU, Hedges SB. 2009. A major clade of prokaryotes with ancient adaptations to life on land. Mol Biol Evol 26:335–343. https://doi.org/10.1093/molbev/msn247
47. Schirrmeister BE, Sanchez-Baracaldo P, Wacey D. 2016. Cyanobacterial evolution during the Precambrian. Int J Astrobiol 15:187–204. https://doi.org/10.1017/S1473550415000579
48. Lyons TW, Reinhard CT, Planavsky NJ. 2014. The rise of oxygen in Earth's early ocean and atmosphere. Nature 506:307–315. https://doi.org/10.1038/nature13068
49. Bekker A, Holland HD. 2012. Oxygen overshoot and recovery during the early Paleoproterozoic. Earth Planet Sci Lett 317–318:295–304. https://doi.org/10.1016/j.epsl.2011.12.012
50. Allen EE, Banfield JF. 2005. Community genomics in microbial ecology and evolution. Nat Rev Microbiol 3:489–498. https://doi.org/10.1038/nrmicro1157
51. Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428:37–43. https://doi.org/10.1038/nature02340
52. Qian PY, Cheng A, Wang R, Zhang R. 2022. Marine biofilms: diversity, interactions and biofouling. Nat Rev Microbiol 20:671–684. https://doi.org/10.1038/s41579-022-00744-7
53. Teng YC, Primeau FW, Moore JK, Lomas MW, Martiny AC. 2014. Global-scale variations of the ratios of carbon to phosphorus in exported marine organic matter. Nature Geosci 7:895–898. https://doi.org/10.1038/ngeo2303
54. Hao W, Mänd K, Li Y, Alessi DS, Somelar P, Moussavou M, Romashkin AE, Lepland A, Kirsimäe K, Planavsky NJ, Konhauser KO. 2021. The kaolinite shuttle links the Great Oxidation and Lomagundi Events. Nat Commun 12:2944. https://doi.org/10.1038/s41467-021-23304-8
55. Zhang Q, Bendif EM, Zhou Y, Nevado B, Shafiee R, Rickaby REM. 2022. Declining metal availability in the Mesozoic seawater reflected in phytoplankton succession. Nat Geosci 15:932–941. https://doi.org/10.1038/s41561-022-01053-7
56. Chen J, Li Y, Jing H, Zhang X, Xu Z, Xu J, Liu H. 2022. Genomic and transcriptomic evidence for the diverse adaptations of *Synechococcus* subclusters 5.2 and 5.3 to mesoscale eddies. New Phytol 233:1828–1842. https://doi.org/10.1111/nph.17903
57. Huang S, Wilhelm SW, Harvey HR, Taylor K, Jiao N, Chen F. 2012. Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. ISME J 6:285–297. https://doi.org/10.1038/ismej.2011.106
58. Flombaum P, Gallegos JL, Gordillo RA, Rincón J, Zabala LL, Jiao N, Karl DM, Li WKW, Lomas MW, Veneziano D, Vera CS, Vrugt JA, Martiny AC. 2013. Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. Proc Natl Acad Sci USA 110:9824–9829. https://doi.org/10.1073/pnas.1307701110
59. Deng L, Cheung S, Kang CK, Liu K, Xia X, Liu H. 2022. Elevated temperature relieves phosphorus limitation of marine unicellular diazotrophic cyanobacteria. Limnol Oceanogr 67:122–134. https://doi.org/10.1002/lno.11980
60. Luo H, Huang Y, Stepanauskas R, Tang J. 2017. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nat Microbiol 2:17091. https://doi.org/10.1038/nmicrobiol.2017.91
61. Sánchez-Baracaldo P. 2015. Origin of marine planktonic cyanobacteria. Sci Rep 5:17418. https://doi.org/10.1038/srep17418
62. Sánchez - baracaldo P, Hayes PK, Blank CE. 2005. Morphological and habitat evolution in the cyanobacteria using a compartmentalization approach. Geobiology 3:145–165. https://doi.org/10.1111/j.1472-4669.2005.00050.x
63. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci USA 108:13624–13629. https://doi.org/10.1073/pnas.1110633108
64. Konhauser KO, Hamade T, Raiswell R, Morris RC, Grant Ferris F, Southam G, Canfield DE. 2002. Could bacteria have formed the Precambrian banded iron formations? Geology 30:1079. https://doi.org/10.1130/0091-7613(2002)030<1079:CBHFTP>2.0.CO;2
65. Konhauser KO, Robbins LJ, Alessi DS, Flynn SL, Gingras MK, Martinez RE, Kappler A, Swanner ED, Li Y-L, Crowe SA, Planavsky NJ, Reinhard CT, Lalonde SV. 2018. Phytoplankton contributions to the trace-element composition of Precambrian banded iron formations. GSA Bull 130:941–951. https://doi.org/10.1130/B31648.1
66. Planavsky NJ, Asael D, Hofmann A, Reinhard CT, Lalonde SV, Knudsen A, Wang X, Ossa Ossa F, Pecoits E, Smith AJB, Beukes NJ, Bekker A, Johnson TM, Konhauser KO, Lyons TW, Rouxel OJ. 2014. Evidence for oxygenic photosynthesis half a billion years before the Great Oxidation Event. Nature Geosci 7:283–286. https://doi.org/10.1038/ngeo2122
67. Sánchez-Baracaldo P, Raven JA, Pisani D, Knoll AH. 2017. Early photosynthetic eukaryotes inhabited low-salinity habitats. Proc Natl Acad Sci USA 114:E7737–E7745. https://doi.org/10.1073/pnas.1620089114
68. Boden JS, Konhauser KO, Robbins LJ, Sánchez-Baracaldo P. 2021. Timing the evolution of antioxidant enzymes in cyanobacteria. Nat Commun 12:4742. https://doi.org/10.1038/s41467-021-24396-y
69. Swanner ED, Mloszewska AM, Cirpka OA, Schoenberg R, Konhauser KO, Kappler A. 2015. Modulation of oxygen production in Archaean oceans by episodes of Fe(II) toxicity. Nature Geosci 8:126–130. https://doi.org/10.1038/ngeo2327
70. Lee OO, Wang Y, Yang J, Lafi FF, Al-Suwailem A, Qian PY. 2011. Pyrosequencing reveals highly diverse and species-specific microbial communities in sponges from the Red Sea. ISME J 5:650–664. https://doi.org/10.1038/ismej.2010.165
71. Qian PY, Wang Y, Lee OO, Lau SCK, Yang J, Lafi FF, Al-Suwailem A, Wong TYH. 2011. Vertical stratification of microbial communities in the Red Sea revealed by 16S rDNA pyrosequencing. ISME J 5:507–518. https://doi.org/10.1038/ismej.2010.112

72. Wang Y, Yang J, Lee OO, Dash S, Lau SCK, Al-Suwailem A, Wong TYH, Danchin A, Qian PY. 2011. Hydrothermally generated aromatic compounds are consumed by bacteria colonizing in Atlantis II Deep of the Red Sea. ISME J 5:1652–1659. https://doi.org/10.1038/ismej.2011.42