Phylogenomics Uncovers Evolutionary Trajectory of Nitrogen Fixation in Cyanobacteria

Meng-Yun Chen, Wen-Kai Teng, Liang Zhao, Bo-Ping Han, Li-Rong Song, and Wen-Sheng Shu

1Institute of Ecological Science, Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, Guangdong Provincial Key Laboratory of Biotechnology for Plant Development, School of Life Sciences, South China Normal University, Guangzhou, PR China
2State Key Laboratory of Biocontrol, Guangdong Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou, PR China
3Department of Ecology and Institute of Hydrobiology, Jinan University, Guangzhou, PR China
4Key Laboratory of Algal Biology, Institute of Hydrobiology, Chinese Academy of Science, Hubei, PR China

*Corresponding authors: E-mails: shuwensheng@m.scnu.edu.cn; lrsong@ihb.ac.cn; tbphan@jnu.edu.cn.
Associate editor: Fabia Ursula Battistuzzi

Abstract

Biological nitrogen fixation (BNF) by cyanobacteria is of significant importance for the Earth’s biogeochemical nitrogen cycle but is restricted to a few genera that do not form monophyletic group. To explore the evolutionary trajectory of BNF and investigate the driving forces of its evolution, we analyze 650 cyanobacterial genomes and compile the database of diazotrophic cyanobacteria based on the presence of nitrogen fixation gene clusters (NFGCs). We report that 266 of 650 examined genomes are NFGC-carrying members, and these potentially diazotrophic cyanobacteria are unevenly distributed across the phylogeny of Cyanobacteria, that multiple independent losses shaped the scattered distribution. Among the diazotrophic cyanobacteria, two types of NFGC exist, with one being ancestral and abundant, which have descended from diazotrophic ancestors, and the other being anaerobe-like and sparse, possibly being acquired from anaerobic microbes through horizontal gene transfer. Interestingly, we illustrate that the origin of BNF in Cyanobacteria coincide with two major evolutionary events. One is the origin of multicellularity of cyanobacteria, and the other is concurrent genetic adaptations with massive gene gains and expansions, implicating their key roles in triggering the evolutionary transition from nondiazotrophic to diazotrophic cyanobacteria. Additionally, we reveal that genes involved in accelerating respiratory electron transport (coxABC), anoxenic photosynthetic electron transport (sqr), as well as anaerobic metabolisms (pf or, hemN, nrdG, adhE) are enriched in diazotrophic cyanobacteria, representing adaptive genetic signatures that underpin the diazotrophic lifestyle. Collectively, our study suggests that multicellularity, together with concurrent genetic adaptations contribute to the evolution of diazotrophic cyanobacteria.

Key words: Cyanobacteria, biological nitrogen fixation, evolution, comparative genomics.

Introduction

Biological nitrogen fixation (BNF) is a critical process in the nitrogen biogeochemical cycle that impacts primary productivity in both marine and terrestrial ecosystems (LeBauer and Treseder 2008; Canfield et al. 2010). Although the process is mediated by various archaeal and bacterial lineages, diazotrophic cyanobacteria are considered to be of paramount importance in the modern nitrogen cycle. Nitrogen fixation in the oceans is largely attributed to a small group of marine cyanobacteria (e.g., Trichodesmium, Richelia, UCYN-A ( unicellular cyanobacteria Candidatus Atelocyanobacterium thalassa), and Crocosphaera) (Sohm et al. 2011; Martínez-Pérez et al. 2016; Zehr and Capone 2020), and cyanobacteria in crypto- gamic covers are recognized as major players that contribute to nearly half of the total BNF on land (e.g., Microcoleus, Leptolyngbya, and Pseudanabaenaceae) (Elbert et al. 2012; Pietrasiak et al. 2013). Therefore, elucidating the evolutionary history of this important trait within Cyanobacteria sets the ground for understanding the evolution of Earth’s biogeochemical nitrogen cycle (Stüeken et al. 2016).

BNF is solely catalyzed by nitrogenase enzymes (Canfield et al. 2010). There are three forms of nitrogenase, characterized by different metal contents of the active-site cofactor: molybdenum-iron nitrogenase (Mo-nitrogenase), vanadium-iron nitrogenase (V-nitrogenase), and iron-iron nitrogenase (Fe-nitrogenase) (Edy 1996; Raymond et al. 2004). Among these, Mo-nitrogenase is the most effective and dominant enzyme complex. The Mo-nitrogenase enzyme consists of two components: the electron-transfer...
Fe protein (dinitrogenase reductase) encoded by *nifH*, and the MoFe protein (dinitrogenase) encoded by *nifD* and *nifK* genes. The Fe protein provides the driving force for electron transfer, which is a homodimer that contains two adenosine triphosphate (ATP)-binding sites and a 4Fe-4S cluster. The MoFe protein is a $\alpha_2\beta_2$ tetramer that contains two metal clusters: the FeMo-co in the active site, and the $\beta$ cluster for transferring electrons from the Fe protein to the FeMo-co. In addition, the biosynthesis of the FeMo-co in the MoFe protein depends on three additional genes, *nifENB* (Raymond et al. 2004; Esteves-Ferreira et al. 2017). The nitrogenase is an oxygen-sensitive enzyme. Only under microaerobic or anaerobic conditions, the synthesis of nitrogenase and its catalytic reaction occur (Fay 1992). During the BNF process, the Fe protein is reduced by electron donor, such as ferredoxin. Then, with hydrolysis of two ATPs, electrons are further transferred from the Fe protein to the MoFe, resulting in N$_2$ being reduced to NH$_3$ (Duval et al. 2013). Hence, cyanobacteria that are capable of BNF usually satisfy the following conditions: 1) possessing a genetic toolbox (*nifHDKENB*) that encodes for nitrogen-fixing enzyme nitrogenase; 2) having strategies that protect oxygen-sensitive nitrogenase from atmospheric oxygen and oxygen produced in oxygenic photosynthesis process (Berman-Frank et al. 2001, 2003; Zehr et al. 2008; Bandyopadhyay et al. 2013; Bergman et al. 2013; Cornejo-Castillo and Zehr 2019; Inomura, Deutsch, et al. 2019; Inomura, Wilson, et al. 2019). However, the genetic factors that are evolutionarily correlated with BNF remain poorly characterized.

Here, on the basis of the representative genomic data set and the backbone phylogeny of Cyanobacteria that we recently published (Chen et al. 2021), we address the knowledge gaps in understanding the origin and evolutionary history of BNF in Cyanobacteria and explore the underlying genetic players affecting the evolutionary dynamics of BNF. We carried out comparative genomic analysis of the Mo-nitrogenase gene family (*nifHDKENB*) across 650 cyanobacterial genomes and determined the phylogenetic distribution of diazotrophic cyanobacteria. Then plausible evolutionary scenarios of BNF were evaluated according to the phylogenetic patterns of Mo-nitrogenase genes. By inferring the ancestral gene repertoire, ancestral states of BNF and multicellularity, we explored the potential evolutionary forces that shaped the evolution of BNF. Using gene enrichment analysis, we further uncovered genes correlated with BNF trait, which are likely to be involved in evolutionary adaptation to diazotrophic lifestyle.

**Results**

**BNF Capacity Is Unevenly Distributed among Cyanobacteria With Different Taxonomy Classifications and Morphologies**

To assess the potential BNF capacity of cyanobacterial strains, searches for homologs of six core Mo-nitrogenase-related genes (*nifHDKENB*) (Rubio and Ludden 2008; Dos Santos et al. 2012) were carried out for the genomes of 604 Oxyphotobacteria strains and 46 Melainabacteria/Sericytobacterium strains, which represent a broad taxonomic of Cyanobacteria. Nitrogenase-related genes were detected in 266 genomes (fig. 1; supplementary table S1, Supplementary Material online), of which 257 contained intact copies of the six core genes for nitrogenase (*nifHDKENB*), 9 only possessed the genes that encode the nitrogenase catalytic components (*nifHDK*) (supplementary fig. S1, Supplementary Material online).

The six core nitrogenase-related genes typically formed a conserved gene cluster (nitrogen fixation gene cluster [NFGC]; supplementary fig. S2 and table S2, Supplementary Material online). The highly conserved NFGCs were missing in Melainabacteria/Sericytobacterium strains. In contrast, the presence of NFGC was found across all of the major lineages in Oxyphotobacteria except basal Gloeobacterales and Gloeomargaritales orders. Mapping NFGCs to the reference
Orders

- Gloeobacteriales
- Synechococcales
- Gloeomargaritales
- Oscillatoriales
- Chroococcales
- Pleurocapsales
- Spirulinales
- Chroococidiopsidales
- Nostocales
- Melainabacteria

Morphology

- Unicellular
- Baecystous
- Filamentous
- Heterocystous
- Ramified
- Unknown

Table 1: Distribution of nitrogenase complexes across the phylogeny of Cyanobacteria. The figure depicts the phylogenetic distribution of genes encoding molybdenum-dependent nitrogenase. The phylogenetic tree and taxonomic classification are based on our recent study (Chen et al. 2021). The inner layer denotes the morphological type of each strain. Brown circles indicate genomes with the potential to fix nitrogen using molydbenum-based nitrogenase based on the presence of nifHDKENB genes. Yellow circles indicate genomes with nifHDK genes. Right bottom panels show the proportion of genomes that have the potential capacity to fix nitrogen across diverse orders and morphologies. "HQGenomeCounts" denotes the number of high-quality genomes. "NHQgenomeCounts" denotes the number of high-quality genomes that have the potential capacity to fix nitrogen.
cyanobacterial phylogenetic tree showed that their phylogenetic distribution was uneven (fig. 1). The NFGC was nearly consistently present in the monophyletic clade, *Nostocales* (158 of 181 high-quality genomes, fig. 1), which is capable of forming heterocysts or ramified filaments to facilitate the coexistence of BNF and photosynthesis, and some strains even carry two sets of NFGC (supplementary fig. S3, Supplementary Material online). On the other hand, among the polyphyletic order, *Synechococcales*, which is typically unicellular, NFGCs were rare (39 of 181 high-quality genomes, fig. 1), and NFGCs were restricted to particular clades, including basal unicellular lineage (*Synechococcales I*, fig. 1) and filamentous lineage (*Synechococcales IV*, fig. 1).

Additionally, BNF capacity was unevenly distributed among cyanobacteria with different morphologies. NFGCs were rare in the genomes of unicellular cyanobacteria (26 of 147 high-quality genomes, fig. 1). In addition to the aforementioned unicellular *Synechococcales I* lineage, the unicellular diazotrophic genomes were concentrated in the *Chroococcales IV* lineage from marine ecosystems, which contains the well-known strains responsible for oceanic N₂ fixation (UCYN-A and *Crococphaera*) (Zehr and Capone 2020). On the other hand, roughly one-third of nonheterocyst-forming filamentous cyanobacteria had NFGCs in their genomes (51 of 158 high-quality genomes, fig. 1), and NFGCs were prevalent in cyanobacteria with heterocyst structures (142 of 163 high-quality genomes) or ramified filaments (14 of 15 high-quality genomes, fig. 1).

We further examined whether the presence of NFGCs was correlated with a specific environment. We noticed that NFGCs were present in the majority of cyanobacteria that live in thermal springs (19 of 26 high-quality genomes; supplementary table S3, Supplementary Material online), suggesting that BNF capacity might confer a selective advantage to cyanobacteria in these habitats (Estrella Alcamin et al. 2015). The pervasiveness of BNF capacity was also found in cyanobacteria with symbiotic lifestyles (21 of 35 high-quality genomes; supplementary table S3, Supplementary Material online). Consistent with this result, many symbiotic partnerships were found between diazotrophic cyanobacteria and various eukaryotic organisms, such as diatoms, green algae, corals, and plants (Rai et al. 2000; Fiore et al. 2010; Harding et al. 2018; Warshen et al. 2018; Harke et al. 2019), indicating that diazotrophic cyanobacteria could enter symbiosis more easily, since they expand their host’s metabolic capabilities.

In addition to the *Mo*-nitrogenase genes, genes encoding alternative *V*-nitrogenase (*vmhfKDC*) were found in a few cyanobacterial genomes (all present in the BNF gene-enriched clade, *Nostocales*, supplementary fig. S3 and table S4, Supplementary material online). In agreement with previous findings (Gagunashvili and Andrésson 2018), these genomes also harbor genes encoding the *Mo*-nitrogenase enzyme, suggesting that cyanobacteria might benefit from equipping with additional vanadium-dependent nitrogenase under molybdenum-deprived conditions. Collectively, our results suggest that BNF capacity is a widespread trait in *Cyanobacteria* and that metabolic function and related genes are unevenly distributed across clades and diverse morphological groups.

### Evolutionary Events Drive the Emergence of BNF in *Cyanobacteria*

To infer the evolutionary trajectory of BNF trait in *Cyanobacteria*, we performed ancestral state reconstruction using both the maximum likelihood (ML) method and Bayesian inference (BI). In contrast to an earlier study based on limited taxon sampling (Latysheva et al. 2012), we found that the marginal probabilities of BNF at the LCAC were low, indicating a nondiazotrophic ancestor (0.33 and 0.73 for ML and BI, respectively; fig. 2). This hypothesis is further supported by the pattern that the closest outgroup lineages (*Melainabacteria* and *Sericyctothromatia*) (Matheus Carnevali et al. 2019) as well as the basal cyanobacterial lineage are nondiazotrophic organisms. Our result showed that BNF evolved in the most recent common ancestor of *Synechococcales IV* and *Synechococcales V* lineages (Node 8, ML: 0.96, BI: 0.99, respectively; fig. 2). According to our analyses, the LCAC is a nondiazotrophic cyanobacterium that is capable of oxygenic photosynthesis (Latysheva et al. 2012; Harel et al. 2015; Esteves-Ferreira et al. 2017; Soo et al. 2017). Therefore, we can expect that the oxygenated conditions created by oxygenic photosynthesis might act as strong selective forces that affecting the coexistence of oxygen-sensitive BNF and oxygenic photosynthesis in ancestral cyanobacteria (Berman-Frank et al. 2003; Zehr et al. 2008; Schirrmeister et al. 2011; Bandyopadhyay et al. 2013). Thus, we inferred the evolution of cyanobacterial multicellularity, since multicellularity may serve as the launching pad for the separation of two incompatible processes that could not occur simultaneously in one cell. As expected from our analyses and previous study on the trait evolution of cyanobacterial multicellularity, the LCAC is a unicellular cyanobacterium, and the origin of multicellularity (Node 8, ML: 1.0, BI: 0.97, respectively; fig. 2) is synchronized with the origin of BNF (Node 8, ML: 0.96, BI: 0.99, respectively; fig. 2) (Hammerschmidt et al. 2021). To further examine whether BNF and multicellularity coevolved in *Cyanobacteria*, we performed Pagel’s binary traits correlation test (Pagel 1994). The model that the evolution of BNF depended on that of multicellularity best explained the data (weighted AIC = 0.8378, fig. 2). Specifically, much higher instances of the presence of BNF were found when multicellularity was present. These results collectively indicate that multicellularity might act as an important driver of BNF in *Cyanobacteria*, that facilitate the separation of oxygenic photosynthesis and oxygen-sensitive BNF.

In addition, gene family analysis of cyanobacterial genomes revealed that a large number of gene gain and expansion events occurred at the phylogenetic nodes where the BNF and multicellularity emerged (1010 events, Node 8; fig. 2) and its preceding node (1056 events, Node 7; fig. 2). Of the increased gene families at Node 8 and its preceding node (Node 7), gene ontology analysis showed that more than half of them (∼57% for Node 8, ∼56% for Node 7) were identified as hypothetical proteins. Of note, we found that several gene families associated with iron uptake and regulation (*fur, efeU, feoB, and afuA*) and *Mo* transport-related functions (*mapA*) were acquired and
FIG. 2. Reconstruction of BNF across cyanobacterial evolution. The figure depicts the inferred capacity of BNF during cyanobacterial evolution. The gray branch denotes that the strain is unable to fix nitrogen, the orange color denotes that the strain is a diazotrophic cyanobacterium, and the black color indicates that the ancestral status of BNF capacity remains uncertain in the corresponding node. Colored circles on the nodes indicate trait transitions. The first confidently inferred ancestor with BNF trait was found on Node 8 (probability cutoff of 0.95), and the transition from unicellular to multicellular cyanobacteria was inferred on the same node (probability cutoff of 0.95), indicating the concurrent origin of BNF and multicellularity. The bottom right panel displays the results of page's test for evolutionary correlation between multicellularity and BNF. The thickness of arrows represents the rate of change from one combination of trait states (i.e., multicellular and non-nitrogen-fixing) to another combination (i.e., multicellular and nitrogen-fixing). The higher the rate the thicker the arrow. Bayes factor analysis shows that the transition rate of BNF trait gain [q01] is significantly lower than the transition rate of BNF trait loss [q10]. The numbers next to nodes connecting major clades of the tree represent reconstructed gene gains/expansions and losses events of selected lineages, that “+” represents gain of gene family and “−” represent gene loss event.
expanded at these two steps (supplementary table S5, Supplementary Material online). These gene gains and expansions meet the high demand for iron and Mo in photosynthesis and BNF, respectively. Other expanded gene families including genes involved in lipid metabolisms to maintain membrane fluidity and rigidity (mlaE, hpnP; supplementary table S5, Supplementary Material online), were also reported to be highly related to cyanobacteria with BNF capacity (Malinverni and Silhavy 2009; Ricci et al. 2014). These correlated changes suggest that the burst of gene gains and expansions might facilitate the transition from nondiazotrophic cyanobacteria to diazotrophic cyanobacteria.

Parallel Losses Contributed to the Scattered Phylogenetic Distribution of BNF in Cyanobacteria

Ancestral state reconstruction also revealed multiple independent losses of BNF under the course of cyanobacterial evolution (fig. 2), which included both trait losses in deep clades (e.g., on the branch leading to Synechococcales IV group) and trait losses that occurred recently (e.g., within the order Nostocales). The rates of BNF gain and BNF loss showed significant differences, in which the rate of transition from nitrogen-fixing to non-nitrogen-fixing was approximately five times higher than the transition rate from non-nitrogen-fixing to nitrogen-fixing (trait loss [q10]: 3.47; trait gain [q01]: 0.66). This finding suggests that loss of BNF is strongly favored over acquisition of BNF in cyanobacterial lineages (Kunin and Ouzounis 2003). Specifically, we found that the BNF trait is subject to differential losses among cyanobacteria. Although sporadic trait losses exist in multicellular cyanobacteria, the unicellular cyanobacteria were observed to be more prone to undergo loss of BNF in deep clades than multicellular cyanobacteria (e.g., Synechococcales and Chroococcales; fig. 2). The discrepancies observed here might reflect the difficulty in coordinating oxygen-sensitive BNF and oxygenic photosynthesis in one cell, thus reinforced the scenario that the evolution of BNF largely depended on that of multicellularity. It should be noted that the use of incomplete genomes could potentially result in false negatives for diazotrophic cyanobacteria, and subsequently, lead to overestimation of trait loss. The most efficient way to evaluate its impact is that sequence whole genomes of corresponding strains or do PCR amplification to examine whether the strains have NFGCs, unfortunately we do not have DNA samples of these strains at hand. Thus we tried to find clues based on the physical locations of nitrogenase genes in contigs and scaffolds. For example, in the case of Nostocales strains, there are 30 strains have been classified as nondiazotrophic cyanobacteria. We examined the presence of nitrogenase genes (nifHDKENB) and their physical locations in corresponding genomes. We found that 3 genomes of nondiazotrophic strains (Raphidiopsis curvata NIES-932, Cylindrospermopsis raciborskii CENA303, Raphidiopsis brookii D9) do not contain any nitrogenase genes though they are high-quality assemblies, even one of them has been reported as a complete genome, indicating that the absence of NFGC unlikely to be a result of use of incomplete genomes (number of contigs: 1, 47, 77, respectively; genome completeness: 100%, 99.37%, 99.37%, respectively). For the rest 27 genomes of nondiazotrophic strains, they possessed at least one nitrogenase gene for each genome. Tracking genomic locations of these genes in corresponding genomes, we found that the incomplete set of NFGC is usually located at the middle position of large contigs (contigs length >40 kb, 16 of 27 genomes). Therefore, it is less likely that the incomplete set of NFGC results from the incompleteness of genome, instead the incomplete set of NFGC might reflect instances of gene losses in corresponding strains. We did find that 9 genomes of nondiazotrophic strains, their incomplete sets of NFGC are located at the edges of contigs, leaving open the possibility that the incompleteness of genomes results in false negative data. Overall, we argue that the use of incomplete genomes would not change the general trend of our results, but the potential false negative rate brought by the incompleteness of genomes is worthy of attention and further investigation.

Transfer of Nitrogenases across Cyanobacterial Lineages

Given that horizontal gene transfer (HGT) has played an important role in the spread of BNF between domains (bacteria and archaea) (Raymond et al. 2004) and HGT would occur more frequently between closely related species (Popa and Dagan 2011), we tested whether HGT has a large impact on the distribution of BNF in cyanobacteria. We evaluated the impact of horizontal transfer to BNF evolution by analyzing the phylogenetic conflicts between gene trees of six nitrogenase-related genes (nifHDKENB) and organismal topologies. Phylogenies of nitrogenase genes showed that sequences from the same subgroup tended to be closely related (supplementary fig. S4, Supplementary Material online), suggesting that vertical inheritance is still the major contributor to the evolution of BNF (Washburne et al. 2018).

Nevertheless, phylogenies provide clues about candidate intraphylum horizontal gene transfer. In the aforementioned case that some strains affiliated with the order Nostocales harboring two sets of NFGC in their genomes (nif1 and nif2 nitrogenases), we found that these two sets of NFGC have different evolutionary scenarios. All phylogenies of nitrogenase genes supported the topology in which sequences of Nostocales were separated into two clusters instead of as a monophyletic group supported by the backbone phylogeny of cyanobacteria (supplementary figs. S4 and S5, Supplementary Material online). Sequences of nif1 nitrogenase were embedded within the major Nostocales cluster, whose phylogenetic placements were in line with their taxonomic identifications, whereas sequences of nif2 nitrogenase clustered with non-Nostocales sequences, though the closest related organism remains uncertain (in nifHDK gene trees, nif2 sequences were placed as a sister group to the sequences of Synechococcales V lineage with
moderate or high bootstrap support (supplementary figs. S4 and S5, Supplementary Material online). The phylogenetic incongruence between gene tree and species tree suggested that HGT events might occur between cyanobacteria strains, from non-Nostocales strains to Nostocales strains, result in the additional nitrogenase in some strains. The findings here support the inference that cyanobacteria to cyanobacteria HGT events were more likely to occur in strains that already had a set of NFGCs (Raymond et al. 2004). These results also provide plausible explanation for additional nitrogenase synthesized only under anaerobic conditions in vegetative cells for some Nostocales strains, that the corresponding nitrogenase might be derived from nonheterocyst-forming diazotrophic cyanobacteria (Thiel et al. 1995).

Transfer of Nitrogenases across Phyla

Moreover, a closer examination of phylogenies of nitrogenase genes showed that nitrogenase genes exist in two different forms. All unrooted nitrogenase gene trees formed a highly supported cluster that was distinct from the rest of the sequences, with an extremely long internal branch, indicating a different evolutionary scenario (supplementary figs. S4 and S6, Supplementary Material online). Therefore, we designated this cluster as “distinct group,” and the rest of sequences as “major group.” Aside from phylogenetic position, the “distinct group” was also suggested by sequence analysis (supplementary fig. S7, Supplementary Material online). Within the “distinct group,” thirteen strains were commonly found (supplementary table S6, Supplementary Material online), including Microcoleus chthonoplastes PCC 7420, whose NFGCs are thought to be acquired from Deltaproteobacteria via HGT (Bolhuis et al. 2010). Mapping the thirteen strains to the reference Cyanobacteria phylogeny showed that lineages of the “distinct group” were scattered throughout the Cyanobacteria phylogeny (supplementary fig. S8, Supplementary Material online). In addition, we found that cyanobacterial strains possessing distinct NFGCs were isolated from various sites around the world (supplementary table S6, Supplementary Material online). These results collectively imply a scenario that the “distinct group” cyanobacteria acquired their NFGC via one or more HGT events in their early evolution.

To further trace the source of HGTs, distinct sequences of nitrogenase genes were compared to those in the NCBI protein database using BLASTp. The BLASTp result for each nitrogenase gene returned top 500 hits, and over 90% of the hits were identified as noncyanobacterial groups. The top hits derived from sequences of Deltaproteobacteria suggest that HGT of nitrogenase genes possibly occurred from Deltaproteobacteria to Cyanobacteria (supplementary table S7, Supplementary Material online). To further confirm the interphylum HGT, we compiled and curated three data sets (nifHDK) from the NCBI protein database, spanning vast phylogenetic diversity (see Material and Methods). All gene phylogenies revealed that sequences belonging to the distinct group were confidently embedded within Proteobacteria. In contrast, the rest of the sequences formed a monophyletic group, indicating an independent origin of nitrogenase (fig. 3; supplementary figs. S9 and S10, Supplementary Material online). Thus, these results provided strong evidence that distinct nitrogenase genes have been horizontally transferred from a group of Deltaproteobacteria to some cyanobacteria.

In addition, we compared the organization of NFGCs of thirteen strains (“distinct group”) with canonical NFGCs of cyanobacteria (“major group,” fig. 3). The structures of “distinct group” NFGCs were highly conserved among strains (supplementary fig. S8, Supplementary Material online). Interestingly, two nifI genes (nifI1 and nifI2), involved in the switch-off of nitrogenase activity reported in methanogenic archaea (Dodsworth and Leigh 2006), were detected in the downstream of nifH in all “distinct group” NFGCs except one strain, yet nifI genes were not present in the canonical NFGC of “major group” (fig. 3, supplementary table S2, Supplementary Material online). Moreover, investigations into the gene content and order of NFGC revealed that cyanobacterial strains from the “distinct group” possessed smaller operons than their counterparts (fig. 3; supplementary fig. S8, Supplementary Material online). The structural features of “distinct group” NFGCs, including the more compact operon and the presence of nifI genes located between nifH and nifD, were similar to the structure of NFGCs for group II nitrogenase, which is predominantly found in methanogenic archaea and anaerobic bacteria (Raymond et al. 2004). Consistent with this observation, a unique insertion of nearly 50 residues shared by group II nifD sequences was also found in nifD sequences from the “distinct group” (supplementary fig. S11, Supplementary Material online). Therefore, these results indicate that group II nitrogenase genes for BNF were present in the genomes of some cyanobacteria. Taken together, multiple independent lines of evidence strongly suggest that one or more ancient interphylum HGT events of anaerobe-like nitrogenase (group II nitrogenase) (Harel et al. 2015) genes occurred between obligate anaerobes and cyanobacteria.

Adaptive Genetic Signatures Associated With Diazotrophic Lifestyle

To investigate the underpinning genetic mechanisms related to adaptation to diazotrophic lifestyle, we compared the genome contents of cyanobacteria that possess NFGC with those that lack NFGC using gene enrichment analysis. Overall, genes that were enriched in diazotrophic genomes could be categorized into two groups (supplementary table S8, Supplementary Material online). The first group comprised genes located in the NFGC region involved in the BNF process, including nifTVWXZ, the ferredoxin-encoding fdxN and the gene encoding metalloprotein with unknown function (DUF683) (Mulligan and Haselkorn 1989) (supplementary fig. S12, Supplementary
These highly conserved genes located in the NGFC region suggested that there is a strong selective constraint operated at the NGFC region.

The remaining genes, which are frequently found in diazotrophic cyanobacteria but largely missing in genomes that lack NGFC, might reflect metabolic adaptations to BNF (fig. 4; supplementary table S8, Supplementary Material online). Specifically, *hemN*, encodes oxygen-independent coproporphyrinogen III oxidase, leading to biosynthesis of protoheme under anoxic conditions (Fujita et al.

**Fig. 3.** Phylogenetic evidence for interphylum horizontal transfer of nitrogenase based on global phylogeny of the *nifH* gene. The phylogenetic tree was inferred from 3,159 *nifH* protein sequences, representing 8,497 sequences deposited in the NCBI protein database and clustered at 97% sequence identity. Taxonomic classifications of sequences are shown in the outer ring with different colors; white color denotes sequences without taxonomic information. Arrows point out the phylogenetic positions of *nifH* sequences from the major cyanobacteria group and distinct cyanobacteria group. The red pentacle labeled on the tree denotes that the distinct cyanobacterial sequences were embedded in *Proteobacteria* lineages, indicating that interphylum HGT events occurred from *Proteobacteria* to *Cyanobacteria*. The inset displays the differences in structural features between representative NFGC of distinct cyanobacteria group and canonical NFGC of major cyanobacteria group. The signature feature of the NFGCs for group II nitrogenase commonly found in anaerobic bacteria, that two genes (*nifI1* and *nifI2*) located between *nifH* and *nifD*, are highlighted with dashed red lines.
Evolution of Nitrogen Fixation in Cyanobacteria · https://doi.org/10.1093/molbev/msac171

NrdG encodes class III anaerobic ribonucleotide reductase (RNR) small complex, which has been shown to be involved in DNA synthesis and repair under anaerobic conditions in facultative and obligate anaerobes (Eliasson et al. 1990). NifJ, encodes oxygen-sensitive pyruvate:ferredoxin oxidoreductase (PFOR), catalyzing the oxidation of pyruvate to acetyl-coenzyme A (acetyl-CoA) and carbon dioxide, possibly supporting the metabolism of strains under anoxic conditions (Lu and Imlay 2021). The enrichment of aforementioned genes in diazotrophic cyanobacteria (fig. 4; supplementary fig. S13, Supplementary Material online) might confer fitness benefits when cyanobacteria create microaerobic or anaerobic conditions to proceed with the BNF process. We also observed the enrichment of genes involved in iron and molybdate transport in diazotrophic cyanobacteria (fig. 4; supplementary fig. S13, Supplementary Material online), which possibly denotes the cofactors for enzymes involved in photosynthesis and BNF (Barron et al. 2009).

In addition, we further grouped diazotrophic cyanobacteria into heterocyst-forming diazotrophic cyanobacteria and non-heterocyst-forming diazotrophic cyanobacteria, then we compared their genomic contents with those of nondiazotrophic cyanobacteria, respectively (supplementary table S8).

Fig. 4. Schematic representation of enriched pathways in diazotrophic cyanobacteria. Solid colored circles indicate genes significantly enriched in genomes containing NFGCs compared to genomes without NFGCs. Specifically, beige and brown squares indicate genes enriched in genomes of heterocyst-forming diazotrophic cyanobacteria and genes enriched in genomes of non-heterocyst-forming diazotrophic cyanobacteria, respectively. Pathways occurred under anoxic or aerobic conditions are marked with different colors, that black color indicates shared pathway.

Statistical analyses were limited to high-quality genomes to avoid potential bias. hemN, oxygen-independent coproporphyrinogen III oxidase; pdh, pyruvate dehydrogenase; PFL, formate C-acetyltransferase; adhE, acetaldehyde dehydrogenase/alcohol dehydrogenase; TCA variant, tricarboxylic acid cycle variant; NDH, NDH-1 complex; SDH, succinate dehydrogenase; PQ, plastoquinone; PSI, photosystem II; SQR, sulfide:quinone oxidoreductase; B6F, cytochrome b6f; PC, plastocyanin; OX, cytochrome c oxidase; CBB cycle, calvin-benson-bassham cycle; HupLS, uptake hydrogenase.
and fig. S13, Supplementary Material online). We discovered that genes involved in carbohydrate fermentations (pfl, adhE) (Stal and Moezelaar 1997) were significantly correlated with non-heterocyst-forming diazotrophic cyanobacteria, the correlation have not been observed in heterocyst-forming diazotrophic cyanobacteria and nondiazotrophic cyanobacteria, implicating their important roles in supporting energy-requiring processes (BNF) under anoxic conditions for non-heterocyst-forming diazotrophic cyanobacteria. We also observed that sqr gene which encodes a sulfide quinone oxidoreductase (SQR) was enriched in non-heterocyst-forming diazotrophic cyanobacteria. It has been shown that SQR oxidizes sulfide to sulfur and provides electrons to photosystem I (PSI) (Hamilton et al. 2018), the universal presence of sqr in non-heterocyst-forming diazotrophic cyanobacteria suggests that the potential capability to use alternative electron donors for photosynthesis (anoxygenic photosynthesis) other than water (oxygenic photosynthesis), conferring a selective advantage in non-heterocyst-forming diazotrophic cyanobacteria to perform oxygen-sensitive BNF. On the other hand, genes involved in accelerating oxygen consumption (coxABC) (Valladares et al. 2003; Inomura et al. 2017) and consuming hydrogen produced by the nitrogenase (hupLS) (Puente-Sánchez et al. 2018) during BNF were enriched in genomes of heterocyst-forming diazotrophic cyanobacteria (fig. 4; supplementary fig. S13, Supplementary Material online). For heterocyst-forming cyanobacteria, vegetative cells perform oxygenic photosynthesis to supply carbon and energy for growth and BNF, and heterocysts create microoxic environment to perform BNF (Herrero et al. 2016), thus specifically expressing additional cytochrome c oxidase (coxABC) to maintain anoxic environment of heterocyst and recycling the hydrogen produced by BNF (hupLS) to minimize energy losses obviously confer selective advantages in heterocyst-forming cyanobacteria. Together with our observations, we illustrated the universal and primordial genetic adaptations to BNF in cyanobacteria (genes largely conserved in genomes of diazotrophic cyanobacteria), and displayed the divergent genetic adaptations between heterocyst-forming and non-heterocyst-forming cyanobacteria to diazotrophic lifestyle.

Discussion

In this study, we asked what evolutionary pattern of BNF can be found in Cyanobacteria. Our dataset, which includes genomes from Cyanobacteria with extensive taxon sampling, enabling us to perform a detailed analysis of the evolutionary history of BNF in Cyanobacteria. As a result of our analyses, we propose the following evolutionary trajectory of BNF in Cyanobacteria. First, our results evidenced that BNF was not an ancestral feature of the LCAC (Shi and Falkowski 2008; Falcón et al. 2010), the capacity to fix nitrogen evolved in filamentous cyanobacteria (fig. 2), which recently estimated to appeared around 2.2 billion years ago, following the great oxidation event (GOE) (Boden et al. 2021). In the presence of oxygen arose in GOE, ammonium would be biologically converted to nitrite and nitrate, promoting the diversification of organisms which could utilize the newly available oxidants (Ren et al. 2019). On the other hand, organic nitrogen returned to atmosphere via enhanced denitrification and anaerobic ammonium oxidation processes, thus a modern-style aerobic nitrogen cycle dominated by nitrogen loss initially prevail globally (Zerkle et al. 2017). In this scenario, the cyanobacteria with the capacity to produce oxygen and fix both CO2 and N2 thus would be fundamental players in the global C and N cycles since Proterozoic eon. Subsequently, massive independent losses of BNF occurred at broad phylogenetic scales, as it is costly to maintain BNF, including the high energy cost of BNF and the sensitivity of nitrogenase to oxygen poisoning, resulting in the uneven distribution of BNF among cyanobacterial lineages (fig. 1). The maintenance of BNF in specific lineages is often accompanied by evolving more sophisticated mechanisms to coordinate BNF and oxygenic photosynthesis, such as forming specialized heterocysts to fix nitrogen (Nostocales; fig. 1) (Herrero et al. 2016), exhibiting a circadian rhythm of BNF in unicellular cyanobacteria (Bandyopadhyay et al. 2013). It therefore seems that the differential trait losses reflected the trade-off between photosynthesis and BNF in some cyanobacteria (Albalat and Cañestro 2016). Meanwhile, we found evidence for several HGT events generating diversified nitrogenases through the evolution of Cyanobacteria, including interphylum HGT events that contributed to the nitrogenase reported in obligate anaerobes (fig. 3) and interclade HGT events that provided additional nitrogenase (supplementary fig. S5, Supplementary Material online). Altogether, the findings here provided insight into the potential roles of vertical inheritance, differential loss and HGT in the evolution of BNF in Cyanobacteria.

On the basis of a wealth of genomic data of diazotrophic and nondiazotrophic cyanobacteria, we further explored the genetic and mechanistic basis underlying the evolutionary dynamics of BNF. Our results indicate that the two episodes of extensive gene gains and expansions (Node 8 and its preceding node, fig. 2), likely enlarged functional capabilities in ancestral cyanobacteria, contributing to the emergence of BNF and multicellularity. Indeed, gene families which facilitated BNF were equipped during the events. Since BNF evolved in cyanobacteria, adaptation to diazotrophic lifestyle drove changes in genomic contents. As a result, a small set of genes was found to be universally enriched in diazotrophic cyanobacteria, which remodeled the metabolic core to provide fitness advantages to diazotrophic cyanobacteria under anaerobic conditions (fig. 4). In subsequent evolutionary steps, further genomic adaptation appeared in response to different morphologies and phenotypes of diazotrophic cyanobacteria (e.g., nonheterocyst-forming diazotrophic cyanobacteria and heterocyst-forming diazotrophic cyanobacteria; fig. 4). The findings here further improved our understanding of the evolutionary adaptation of Cyanobacteria from nondiazotrophic to diazotrophic.

Materials and Methods

Identification of Nitrogenase Genes

A total of 650 cyanobacterial genomes were used in this study (supplementary table S1, Supplementary Material online).
Details of sample information including taxonomic classifications, morphological traits and data processing procedures were described in our recently published work (Chen et al. 2021). Genome-specific nitrogen fixation potential was determined by screening the core nitrogenase genes nifH, nifD, nifK, nifE, nifN and nifB using the following procedures: 1) Generating custom HMM profiles. Reference protein sequences for each gene were downloaded from the NCBI database. Sequences were aligned using the l-ins-i algorithm in MAFFT (Katoh and Standley 2013), followed by alignment trimming using TrimAl with the “gappyout” option (Capella-Gutierrez et al. 2009). Custom HMM profiles were built based on the trimmed alignments using hmmbuild implemented in HMMER 3.1b1 (Mistry et al. 2013); 2) hmmssearch was employed to search genomes with the custom HMM profiles (cutoff E-value = 1e-5), and all hits above the preset score cutoff were retained; 3) Corresponding sequences were manually curated with additional searches against the NCBI conserved domain database (CDD) (Marchler-Bauer et al. 2015) to confirm homology using a domain-based approach RPS-BLAST (cutoff E-value = 1e-5). Homologs that met the above criteria were prepared for subsequent analysis.

Phylogenetic Analysis of Nitrogenase Genes
For each nitrogenase gene, multiple sequence alignment was built using the l-ins-i algorithm as implemented by MAFFT, and then the ambiguously aligned positions were excluded using TrimAl with the “gappyout” option. Individual gene trees were inferred by IQTree v1.6.6 (Nguyen et al. 2015) with automatic choice of the best-fit model (-m MFP). Ultrafast bootstrap support values were calculated from 1000 replicates (-bb 1000 –bnni).

Horizontal Gene Transfer Analyses
Interphylum Detection
To detect horizontal transfer events that occurred across cyanobacterial strains, the topology of the gene tree estimated from individual nitrogenase genes and the Cyanobacteria phylogeny derived from our recent study were compared. For each gene tree, sequences that were identified as putative interphylum transfer sequences were used as outgroups since they were evolutionarily distantly related to the rest of the data. To discriminate HGT events that are presumably false positives caused by insufficient phylogenetic signals in individual gene data sets, we defined interphylum HGT as the existence of discordant topologies at the order level with high bootstrap support between the gene tree and reference cyanobacterial phylogeny.

Ancient State Reconstruction
To infer the evolution of the BNF trait in Cyanobacteria, we performed ancestral state reconstruction using two different methods: 1) marginal posterior probability approximation algorithm (MPPA) implemented in PastML (Ishikawa et al. 2019) with the F81 model as recommended by the authors (https://pastml.pasteur.fr/) and 2) multistate Markov Chain Monte Carlo model implemented in BayesTraits v3 (Pagel et al. 2004). The reference cyanobacterial phylogenetic tree and binary data set representing the capacity of BNF for each strain were used as inputs for both methods. The MPPA method takes the uncertainty of the ancestral state into account and predicts the ancestral state based on the Brier score. The major characteristics of this method, including less computationally demanding, comparable accuracy and easily interpretable results, make it applicable to inferring ancestral traits.
Identification of Genetic Signatures Associated with Diazotrophic Lifestyle

To discover statistically significant differences between genomes of diazotrophic and nondiazotrophic cyanobacteria, gene enrichment analyses were performed at the Kyoto Encyclopedia of Genes and Genomes, Clusters of Orthologous Genes and domain levels with two approaches: binary matrix (presence/absence data) and abundance matrix (gene/domain copy numbers), using hypergeometric tests and PhyloGLM tests implemented in R with an false discovery rate-adjusted P-value cutoff of 0.05. Analyses were limited to high-quality genomes that were nearly complete (completeness ≥90%) with low contamination (less than 5% contamination). Since the cyanobacteria used in this study originated from different habitats and might have different genetic backgrounds, we applied stringent criteria to distinguish genes/domains that clearly covaried with nitrogen fixation function. We defined genes/domains enriched in genomes with the NFGC as genes/domains that were either shared by nearly all nitrogen fixation genomes (≥80% nitrogen fixation genomes) but not presented in the majority of nonfixer genomes (≤50% nonfixer genomes), or the abundances of genes/domains in nitrogen fixation genomes were significantly higher than those of nonfixer genomes.

Supplementary Material
Supplementary data are available at Molecular Biology and Evolution online.

Acknowledgments
We thank C. He for discussions on the results and comments on the manuscript. This project was supported by the National Natural Science Foundation of China (Grant numbers 31900002 and 41830318) to M.Y.C. and W.S.S., the China Postdoctoral Science Foundation (Grant number 2021M701275) to M.Y.C., the Natural Science Foundation of Guangdong Province (2022A1515010464) to M.Y.C.

Author Contributions
M.C., W.S., L.S. and B.H. conceived the study; M.C. led data analysis; M.C. led the results interpretation and paper writing; M.C., W.S., B.H., L.Z. and W.T. contributed to the writing and discussion of the paper.

Data Availability
All data are provided with this paper are available upon request.

References
Albalat R, Cañestro C. 2016. Evolution by gene loss. Nat Rev Genet. 17:379–391.
Bandyopadhyay A, Elvitigala T, Liberton M, Pakrasi HB. 2013. Variations in the rhythms of respiration and nitrogen fixation in members of the unicellular diazotrophic cyanobacterial genus Cyanothece. Plant Physiol. 161:1334–1346.
Barron AR, Wurzburger N, Bellenger JP, Wright SJ, Kraepiel AML, Hedin LO. 2009. Molybdenum limitation of symbiotic nitrogen fixation in tropical forest soils. Nat Geosci. 2:42–45.
Fujita Y, Tsujimoto R, Aoki R. 2015. Evolutionary aspects and regulation of tetrapyrrole biosynthesis in cyanobacteria under aerobic and anaerobic environments. *Life* 5:1172–1203.

Gagunashvili AN, Andrésón ÖS. 2018. Distinctive characters of Nostoc genomes in cyanolichens. *BMC Genomics* 19:1–18.

Hamilton TL, Klatt JM, De Beer D, McFadyn J, Li J. 2018. Cyanobacterial photosynthesis under sulfidic conditions: insights from the isolate Leptolyngbya sp. strain hensoni. *ISME J.* 12:568–584.

Hammerschmidt K, Landan G, Domingues Kümmel Tria F, Alcorta J, Dagan T. 2021. The order of trait emergence in the evolution of cyanobacterial multicellularity. *Genome Biol Evol.* 13:1–24.

Harding K, Turk-Kubo KA, Sipler RE, Mills MM, Bronk DA, Zehr JP. 2018. Symbiotic unicellular cyanobacteria fix nitrogen in the Arctic Ocean. *Proc Natl Acad Sci U S A* 115:13371–13375.

Harel A, Karkar S, Cheng S, Falkowski PG, Bhattacharya D. 2015. Deciphering primordial cyanobacterial genome functions from protein network analysis. *Curr Biol.* 25:628–634.

Harke MJ, Frischkorn KR, Haley ST, Aylward FO, Zehr JP, Dyhrman ST. 2019. Periodic and coordinated gene expression of a diazotroph and its diatom host. *ISME J.* 13:118–131.

Herrero A, Stavans J, Flores E. 2016. The multicellular nature of filamentous heterocyst-forming cyanobacteria. *FEMS Microbiol Rev.* 40:831–854.

Inomura K, Bragg J, Follows MJ. 2017. A quantitative analysis of the direct and indirect costs of nitrogen fixation: a model based on *Azotobacter vinelandii*. *ISME J.* 11:166–175.

Inomura K, Deutsch C, Wilson ST, Masuda T, Lawrenz E, Büchinska L, Sobotka R, Gauglitz JM, Saito MA, Prášil O, et al. 2019. Quantifying oxygen management and temperature and light dependencies of nitrogen fixation by *Crocospheara watsonii*. *mSphere* 4:e00531-19.

Inomura K, Wilson ST, Deutsch C. 2019. Mechanistic model for the coexistence of nitrogen fixation and photosynthesis in marine trichomycetes. *mSystems* 4:1–13.

Ishikawa SA, Zhukova A, Iwasaki W, Gascul O. 2019. A fast likelihood method to reconstruct and visualize ancestral scenarios. *Mol Biol Evol.* 36:2069–2085.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 30:772–780.

Komárek J, Kaštovský J, Mareš J, Johansen J. 2014. Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia* 86:295–335.

Kunin V, Ouzounis CA. 2003. The balance of driving forces during genome evolution in prokaryotes. *Genome Res.* 13:1589–1594.

Latsycheva N, Junker VL, Palmer WJ, Codd GA, Barker D. 2012. The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics* 28:603–606.

LeBauer D, Treseder K. 2008. Nitrogen limitation of net primary productivity. *Ecology* 89:371–379.

Letunic I, Bork P. 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47:W256–W259.

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.

Lu Z, Imlay JA. 2021. When anaerobes encounter oxygen: mechanisms of oxygen toxicity, tolerance and defence. *Nat Rev Microbiol.* 19:774–785.

Malinverni JC, Silhavy TJ. 2009. An ABC transport system that maintains lipid asymmetry in the Gram-negative outer membrane. *Proc Natl Acad Sci U S A.* 106:8009–8014.

Marchler-Bauer A, Derbyshire MK, González NR, Lu S, Chitsaz F, Geer LY, He J, Gwadz M, Hurwitz DI, et al. 2015. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 43:D222–D226.

Martínez-Pérez C, Mohr W, Löschér CR, Dekaezemacker J, Littmann S, Yilmaz P, Lehnen N, Fuchs BM, Lavik G, Schmitz RA, et al. 2016. The small unicellular diazotrophic symbiont, UCN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol.* 1:16163.
