ISME

ARTICLE   OPEN

Distribution and survival strategies of endemic and cosmopolitan diazotrophs in the Arctic Ocean

Takuhei Shiozaki[1,7 ✉], Yosuke Nishimura [2,7], Susumu Yoshizawa [1], Hideto Takami[1,3], Koji Hamasaki [1,4,5], Amane Fujiwara[6], Shigeto Nishino[6] and Naomi Harada[1,6]

Dinitrogen (N$_2$) fixation is the major source of reactive nitrogen in the ocean and has been considered to occur specifically in low-latitude oligotrophic oceans. Recent studies have shown that N$_2$ fixation also occurs in the polar regions and thus is a global process, although the physiological and ecological characteristics of polar diazotrophs are not yet known. Here, we successfully reconstructed diazotroph genomes, including that of cyanobacterium UCYN-A (*Candidatus* 'Atelocyanobacterium thalassa'), from metagenome data corresponding to 111 samples isolated from the Arctic Ocean. These diazotrophs were highly abundant in the Arctic Ocean (max., 1.28% of the total microbial community), suggesting that they have important roles in the Arctic ecosystem and biogeochemical cycles. Further, we show that diazotrophs within genera *Arcobacter*, *Psychromonas*, and *Oceanobacter* are prevalent in the <0.2 μm fraction in the Arctic Ocean, indicating that current methods cannot capture their N$_2$ fixation. Diazotrophs in the Arctic Ocean were either Arctic-endemic or cosmopolitan species from their global distribution patterns. Arctic-endemic diazotrophs, including Arctic UCYN-A, were similar to low-latitude-endemic and cosmopolitan diazotrophs in genome-wide function, however, they had unique gene sets (e.g., diverse aromatics degradation genes), suggesting adaptations to Arctic-specific conditions. Cosmopolitan diazotrophs were generally non-cyanobacteria and commonly had the gene that encodes the cold-inducible RNA chaperone, which presumably makes their survival possible even in deep, cold waters of global ocean and polar surface waters. This study shows global distribution pattern of diazotrophs with their genomes and provides clues to answering the question of how diazotrophs can inhabit polar waters.

INTRODUCTION

Dinitrogen (N$_2$) fixation is the process by which specialized prokaryotes (diazotrophs) convert dinitrogen gas to ammonia. It has long been thought that N$_2$ fixation occurs mainly in the nitrogen (N)-depleted tropical and subtropical regions where cyanobacterial diazotrophs are prevalent [1]. However, recent studies have demonstrated that N$_2$ fixation occurs even in the N-rich polar regions [2–5].

In the Arctic Ocean, sea ice exists throughout the year, and the seawater temperature remains low (average: ~0–1 °C in August [6]). The Arctic surface layer is relatively stable despite the high latitude due to the presence of low-salinity water, and some regions become oligotrophic in summer [7, 8]. During the autumn and winter, mixing processes become active due to the frequent strong wind and atmospheric cooling, which break down water stratification, and nitrogenous nutrients are supplied to the surface layer throughout the Arctic Ocean [8]. Furthermore, underwater irradiance varies dramatically throughout the year, with solar radiation disappearing in winter. The Arctic environment is largely different from that in tropical and subtropical oligotrophic regions, in which the seawater temperature in the mixed layer is maintained above

25 °C throughout the year, the thick N-depleted surface layer is rarely disturbed due to the stable water structure, and underwater irradiance is also stable throughout the year [9]. Given this environmental difference, diazotrophs could have unique strategies for adapting to the Arctic environment.

Information on diazotrophs in the Arctic Ocean is currently limited to the *nifH* sequence. *nifH* encodes the iron protein subunit of nitrogenase, which has a highly conserved sequence that can be used for species identification [10]. Many *nifH* sequences retrieved from the Arctic Ocean differ from those at lower latitudes [2, 3, 11], suggesting the presence of Arctic-endemic diazotrophs. However, *nifH* alone does not explain how diazotrophs have adapted to the Arctic environment. Furthermore, *nifH* alone is not necessarily an indicator of microbial N$_2$ fixation [12, 13]. Mise et al. [13] recently demonstrated that ~20% of bacterial genomes with *nifH* in the public database do not have *nifD* and/or *nifK*, which encode essential subunits of nitrogenase, indicating that these genomes represent non-diazotrophs. Although their study primarily used data from the terrestrial environment, it raises important implications for marine N$_2$ fixation studies as well. Collectively, current understandings of Arctic diazotrophs based only on *nifH* information are incomplete.

[1]Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa 277-8564, Japan. [2]Research Centre for Bioscience and Nanoscience, Japan Agency for Marine-Earth Science and Technology (JAMSTEC), Yokosuka 237-0061, Japan. [3]Center for Mathematical Science and Advanced Technology, JAMSTEC, Yokohama 236-0001, Japan. [4]Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, 277-8564 Kashiwa, Japan. [5]Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, 113-8657 Bunkyo-ku, Japan. [6]Research Institute for Global Change, JAMSTEC, Yokosuka 237-0061, Japan. [7]These authors contributed equally: Takuhei Shiozaki, Yosuke Nishimura. ✉email: shiozaki@g.ecc.u-tokyo.ac.jp

The recent development of culture-independent genome reconstruction techniques such as metagenome-assembled genomes (MAGs) has changed our understanding of microbial ecology, including that of diazotrophs. Genome-based approaches have been applied in tropical and subtropical studies, revealing previously unknown diazotrophs (e.g. lineage within Planctomycetes) and their physiology and ecology [14, 15]. Recently, we have newly built a marine MAG catalog, which contains >50,000 genomes derived from 8466 prokaryotic species that were derived from various marine oceanic regions including the Arctic Ocean [16]. Here, we explored the genome catalog for Arctic diazotrophic species using *nifH* sequences as an initial marker gene. Notably, we successfully retrieved the genome of symbiotic cyanobacterial diazotroph UCYN-A (*Candidatus* 'Atelocyanobacterium thalassa') from metagenomic data of the Arctic Ocean. UCYN-A is one of the major diazotrophs in subtropical regions [17] and is currently divided into six subclades based on its *nifH* sequences [18], of which UCYN-A1 and -A2 occur exclusively in and fix N$_2$ in the Arctic Ocean [3, 4]. The *nifH* sequences of UCYN-A1 and -A2 in the Arctic Ocean are identical to those in the subtropical Ocean [4]. However, genomic information on UCYN-A in the Arctic Ocean has not been revealed, and thus there are no clues for adaptation mechanisms of UCYN-A to the Arctic environment. We examined the distribution patterns using a global metagenomic database and characterized diazotrophs in the Arctic Ocean by comparative genomic analyses.

MATERIALS AND METHODS
Arctic Ocean metagenomes and MAG construction
The 111 metagenome data used in this study were derived from the water column of the Arctic Ocean (Supplementary Table S1). These samples were originally published in the Tara polar project ($n = 68$) [19, 20], Polar marine reference gene catalog ($n = 31$) [21], and Canada Basin cruise ($n = 12$) [22] and were reanalyzed in a large-scale marine metagenome study (the OceanDNA MAG study) [16]. The samples were collected across the entire Arctic Ocean, except for those collected during the Canada Basin cruise. Size-fractionated samples were collected during each project (0.2–3 and <0.2 μm for the Tara polar project; ≥0.2 μm for the Polar marine reference gene catalog; 0.2–3 μm for the Canada Basin cruise). The samples were collected wide depth range from 0 to 3800 m.

The OceanDNA MAG study reconstructed 52,325 qualified prokaryotic MAGs using 2057 metagenomes derived from various marine environments [16]. We focused on the 111 Arctic Ocean samples, which yielded 6816 MAGs, which include 1095 species representatives. Completeness and contamination of genomes were estimated by taxon-specific sets of single-copy marker genes through the lineage-specific workflow of CheckM v1.0.13 [23]. We explored new diazotroph genomes among the species representatives derived from the Arctic Ocean samples. The *nifH* gene was identified with a significant (<1e−05) and best hit to TIGR01287 among the TIGRFAMs HMM library [24] using hmmsearch (HMMER v3.3.2). The MAGs containing *nifH* were deposited in the UTokyo Repository (https://doi.org/10.15083/0002005808).

Marine diazotroph genomes from the lower latitudes
Diazotroph genome from low-latitude samples were reconstructed from metagenomic data, which included not only existing cyanobacterial diazotrophs but also previously unknown non-cyanobacterial diazotrophs (NCDs) [14, 15], although genomes of *Crocosphaera subtropica* (UCYN-C) and *C. chwakensis* (formerly *Cyanothece* CCY0110) were not included. We then used these genomes as the reference genomes for the Arctic diazotroph MAGs. The genomes of *C. subtropica* (Genbank ID: GCA_000017845.1) and *C. chwakensis* (GCA_000169335.1) were downloaded from the NCBI website. To perform a comparative genomics analysis of UCYN-A, we downloaded all UCYN-A genomes in the NCBI database for which the collection location could be identified.

Taxonomic and gene annotation and metabolic and physiological potential and their clustering
All genomes used in this study were taxonomically classified using the classify workflow (classify_wf) of GTDB-Tk v1.7.0 (GTDB release 202) [25].

Protein coding genes were predicted using prodigal (10.1186/1471-2105-11-119) v2.6.3 with the option "-p single". Gene annotation was performed with hmmsearch (HMMER v3.3.2) using HMMs of Pfam [26], TIGRFAMs [24], and KOfam [27] databases (e-value <1e−5). We further estimated the minimum doubling time using Growthpred [28] with default settings by following the method of Royo-Llonch et al. [29]. The lists of gene annotations for each genome were deposited in the UTokyo Repository (https://doi.org/10.15083/0002005808).

To examine the metabolic and physiological potential of each diazotroph genome, the multi-FASTA file of amino acid sequences of the genes was subjected to Genomaple (formerly MAPLE) ver. 2.3.2 [30]. Genomaple is available through a web interface (https://maple.jamstec.go.jp/maple/maple-2.4.0/) [31] or as a stand-alone package from Docker Hub (https://hub.docker.com/r/genomaple/genomaple). Genes were mapped to 814 functional modules defined by the KEGG [32], resulting in 310 pathways, 298 complexes, 167 functional sets, and 49 signatures. The module completion ratio (MCR) was automatically calculated according to a Boolean algebra–like equation [33], and the Q-value was also calculated to evaluate the MCR in Genomaple. We note that a Q-value near zero indicates a high working probability of the module [34]. All calculations in Genomaple were performed with default settings. We then characterized the overall MCR pattern of all diazotroph genomes. The complete-linkage clustering method was used for the functional classification of diazotrophs with pairwise Euclidean distances between the overall MCR patterns for each genome using hclust function of "stats" and as.phylo function of "ape" in an R statistical package ver. 4.1.2 [35]. KEGG modules with MCR values of 0% for all genomes were excluded from this analysis.

Usage of codons and amino acids
We also conducted an in-depth analysis on the usage patterns of codons and amino acids to characterize the Arctic diazotroph MAGs. To examine the differences in codon usage between the Arctic diazotroph MAGs and other diazotroph genomes, we conducted a principal component analysis using the prcomp function of R v4.0.3 [35] with default parameters, and the first and the second primary components were evaluated. For amino acid usage, we examined the percentage of seven amino acids (I, V, Y, W, R, E, L) relative to the overall amino acid composition in the diazotroph MAGs, which is reported to be the most influential signature of optimal growth temperature [36].

Homology between *nifH* of Arctic diazotroph MAGs and universal *nifH* primers
The *nifH* sequence of Arctic diazotroph MAGs was tested in silico if it was detectable with the existing universal *nifH* primers using gen_primer_match_report.py [14] on anvi'o ver. 7.1 [37].

Phylogenetic analysis
Two types of phylogenetic tree were constructed, one for *nifH* only and one for the whole genome. The complete *nifH* sequences were aligned with MUSCLE in the MEGA11 package [38] with default settings. Then, an *nifH* phylogenetic tree was constructed using the maximum likelihood method, and bootstrap values were determined using 100 iterations implemented in MEGA11. A whole-genome-based phylogenetic tree was constructed by PhyloPhlAn v3 [39], which uses ~400 conserved marker genes. The phylogenetic tree was built with default settings and a rapid bootstrap test of 100 replicates.

Comparative genome analysis of UCYN-A
This study used all UCYN-A genomes for which sampling locations have been clearly described (as of September 2022) (Supplementary Table S2). The method for gene annotation is detailed above. Average nucleotide identity was calculated using the ANI calculator [40]. To visualize the UCYN-A pangenome, the anvi'o ver. 7.1 [37] pangenomic workflow was used. Details of the method are provided (https://merenlab.org/2016/11/08/pangenomics-v2/).

Genome abundance of MAGs in metagenomes
We assessed the fraction of metagenomic reads recruited onto diazotroph genomes. Sequence reads of the 2,057 metagenomes used in the OceanDNA MAG study [16] were mapped onto 59 diazotroph genomes (Supplementary Table S3). Read mapping was performed with

Table 1. Summary of *nifH* detected MAGs in the Arctic Ocean.

Genome	Length (Mb)	Genome completeness	*nifH* primer compatibility	*nifH* cluster	*nifDK*, *nifENB*	Distribution pattern	GTDB-based taxonomy
Arc-UCYN-A2[b]	1.48	74.3 (99.3)[a]	○	I	○	Arctic-endemic	Cyanobacteria; *Atelocyanobacterium thalassa*
Arc-Bactero[b]	4.55	97.3	xnifH4	III	○	Arctic-endemic	Bacteroidota; genus *Sunxiuqinia*
Arc-Campylo[b]	2.92	96.8	○	I	○	Arctic-endemic	Campylobacterota; genus *Arcobacter*
Arc-Alpha[b]	3.18	97.7	○	I	○	Cosmopolitan	Proteobacteria; genus *Novosphingobium*
Arc-Gamma-01[b]	3.21	97.6	○	I	No *nifK*	Arctic-endemic	Proteobacteria; genus *Immundisolibacter*
Arc-Gamma-02[b]	4.14	93.9	xnifH4	I	○	Arctic-endemic	Proteobacteria; genus *Psychromonas*
Arc-Gamma-03[b]	3.74	96.4	xnifH4	I	○	Arctic-endemic	Proteobacteria; genus *Oceanobacter*
Arc-Gamma-04	3.77	91.8	xnifH4	I	No *nifB*	–	Proteobacteria; genus *Motiliproteus*
Arc-Myxo	3.77	84.6	xnifH1,2,3	III	No *nifDK*, or *nifENB*	–	Myxococcota; family JABWCM01

Genome completeness was estimated by CheckM [23].

nifH primer compatibility was examined using the existing universal *nifH* primers, in which nifH1,2,3, and 4 are most widely used primers.

[a]Calculated by regarding the complete genome of the UCYN-A2 (CPSB-1, Supplementary Table S2) as 100.

[b]MAGs were identified as diazotrophs through *nif* gene-based screening. Arc-Gamma-01 may be missing *nifK* due to a fragmented genome assembly (see text).

bowtie2 v2.3.5.1 [41] with the default setting using the quality-controlled paired-end reads of each run. If multiple sequencing runs were performed for one sample, only the run with the largest scale was used. If the sequencing run was >5 Gbps, a subset of 5 Gbps was randomly sampled. Then, the mapping results were sorted using samtools v1.9, and mapped reads with ≥95% identity, of ≥80 bp, and with ≥80% aligned fraction of the read length were extracted using msamtools bundled in MOCAT2 v2.1.3 [23, 42]. Then, the mapped reads were counted using featureCounts [43] bundled in Subread v2.0.0. [44]. The genome abundance in each sample was calculated as a length-normalized count per million microbe genomes of a total community (CPMM) by the following equation,

$$CPMM = CPMT \cdot (\text{mapped read count on the 59 genomes})/(\text{total read count})$$

$$CPMT = 10^6 \cdot G_i/(\text{sum of } G_i \text{ of the 59 genomes})$$

$$G_i = 1000 \cdot C_i/L_i$$

where C_i is a count of mapped reads on genome_i, and L_i is the length of genome_i. CPMT is a length-normalized count per million targeted genomes (i.e., 59 genomes). The concept of CPMT is similar to the 'TPM' measure, which is frequently used in transcriptome analysis [45]. For example, when CPMM is 1000, the genome abundance is estimated as 0.1% of the total microbe community.

Shipboard observation and qPCR assay

We performed shipboard observations of western Arctic Ocean samples to examine the distribution of major diazotrophs in the Arctic metagenome and the environmental factors contributing to their distribution. Sampling was carried out on board the R/V Mirai MR15-03 (06 Sep to 03 Oct 2015), MR16-06 (30 Aug to 22 Sep 2016), and MR17-05C (26 Aug to 21 Sep 2017) cruises. Seawater was collected from depths corresponding to 100%, 10%, 1%, and 0.1% of surface light intensity, and from near bottom in the shelf region or from 100 m in the off-shelf region with Niskin-X bottles and a bucket. The light profiles were determined using a submersible PAR sensor (PRR-800 (Biospherical Instruments) in 2015, C-OPS (Biospherical Instruments) in 2016 and 2017) just before water sampling. The depth profiles of temperature, salinity, and dissolved oxygen were measured with an SBE 911 plus CTD system (Sea-Bird Electronics). Samples for nutrient and chlorophyll *a* analyses were collected in 10-mL acrylic tubes and 290-mL dark bottles and were analyzed immediately onboard. For DNA analysis, 2 L of seawater was collected and filtered onto 0.2-μm pore size Sterivex-GP pressure filters (Millipore). Total DNA was extracted using the ChargeSwitch Forensic DNA Purification kit (Invitrogen). Quantitative PCR (qPCR) analysis targeted the *nifH* sequences of six species (Arc-Bactero, Arc-Alpha, Arc-Gamma-01, Alpha-02, Alpha-04, and *C. chwakensis*) that were particularly highly abundant in the Arctic Ocean in the >0.2-μm fraction and targeted *nifH* of Arc-UCYN-A2. The TaqMan probe and primer sets used here were newly designed except for Arc-UCYN-A2 (Supplementary Table S4). The qPCR analysis was conducted in triplicate using a LightCycler 480 System (Roche Applied Science, Penzberg, Germany). The r² values for the standard curves ranged from 0.990 to 1.000. The efficiency of the qPCR analyses ranged from 93.8 to 100%.

RESULTS AND DISCUSSION
Overall characteristics of MAGs containing *nifH* that were retrieved from the Arctic Ocean

Metagenomic data derived from 111 samples collected in the Arctic Ocean yielded 6816 prokaryotic MAGs, which include 1095 species according to the OceanDNA MAGs catalog [16]. Of those MAGs from the Arctic Ocean, *nifH* was detected in nine MAGs (Table 1 and Supplementary Table S3). Four of the nine *nifH* sequence (Arc-UCYN-A2, Arc-Campylo, Arc-Alpha, and Arc-Gamma-01) was matched with the existing *nifH* universal primers sequences, but the remaining five was not, indicating that those could not be detected with previous PCR approaches (Table 1 and Supplementary Table S5).

Diazotrophs must have at least *nifDK* and *nifEBN* in addition to *nifH* for N₂ fixation [12]. Of the nine MAGs, Arc-Gamma-01, Arc-

Gamma-04, and Arc-Myxo lacked *nifK*, *nifB*, and *nifDK* and *nifENB*, respectively, suggesting that these may not be genomes of diazotrophs. However, upon careful inspection, Arc-Gamma-01 had a contig that included *nifD* at the end of the contig. Because *nifHDK* generally form an operon structure in the genome, this MAG may seem to lack *nifK* due to the limitations of a fragmented genome assembly. As Arc-Gamma-01 was categorized into the genus *Immundisolibacter* and the same genus genome was reconstructed with a different method in which all *nif* genes were isolated from Arctic metagenomic data (https://anvio.org/blog/targeted-binning/), we assumed that Arc-Gamma-01 is a diazotroph. However, Arc-Gamma-04 and Arc-Myxo, neither of which is likely to be a diazotroph, were excluded from the subsequent analysis. In summary, seven MAGs (Arc-UCYN-A2, Arc-Bactero, Arc-Campylo, Arc-Alpha, and Arc-Gamma-01, -02, and -03) were identified as diazotrophs through *nif* gene-based screening.

Among the Arctic diazotroph MAGs, Arc-Bactero belongs to Cluster III and the rest belong to Cluster I based on the *nifH* sequence [10] (Fig. 1a and Table 1). These MAGs were taxonomically classified to Cyanobacteria, Alphaproteobacteria, Gammaproteobacteria, Campylobacterota (formerly categorized as Epsilonproteobacteria), and Bacteroidota at the phylum or class level using the Genome Taxonomy Database [25]. Diazotrophs belonging to the phylum Bacteroidota have not been reported thus far from the marine water column [46]. The cyanobacterial MAG (Arc-UCYN-A2) was classified to UCYN-A belonging to the subclade UCYN-A2. The remaining six Arctic diazotrophs may represent new species within genera *Arcobacter*, *Sunxiuqinia*, *Psychromonas*, *Oceanobacter*, *Immunodisolibacter*, *Novosphingobium* (Table 1 and Supplementary Table S3). Thus far, *nifH* of *Arcobacter* has been found in oceans around the globe including the Arctic Ocean [11, 47, 48], and *nifH* of *Novosphingobium* was recently detected in deep waters in the subtropical ocean [49]. The diazotrophs within genera *Sunxiuqinia*, *Psychromonas*, *Oceanobacter*, and *Immunodisolibacter* were newly discovered in this study.

The phylogenetic tree based on concatenated sequences of a set of marker genes conserved in all MAGs formed various clusters categorized into every phylum or class in the case of Proteobacteria, as expected, and this result supported the correctness of the phylogenetic assignment for each MAG (Fig. 1a). In contrast, the clusters on the phylogenetic tree based on *nifH* genes did not necessarily reflect the phylogenetic placement because the genes assigned to different taxonomic groups were nested in one cluster. For example, Arc-Gamma-01, which is categorized as belonging to Gammaproteobacteria, was buried within the cluster composed of Alphaproteobacteria, which does suggest that the *nifH* gene in Arc-Gamma-01 was acquired by horizontal gene transfer. The horizontal gene transfer of *nif* genes is well-documented in diazotrophs in terrestrial environments [50, 51], however, knowledge in marine environments is scarce due to the paucity of species with available whole genome information. The mismatch between *nifH* and conserved marker genes phylogeny was also observed in low-latitude diazotrophs (e.g. Delta-03 and Alpha-03, Fig.1), implying that the horizontal gene transfer of *nif* genes occurs widely in marine environment. In addition to the phylogenetic analysis, we also carried out hierarchical clustering analysis of MAGs based on the pattern of completion ratios of functional modules as calculated by the Genomaple system [34] (Fig. 1a), which results in the functional classification of the genomes [52]. Although most clusters consisted of a single taxonomic group, some were made up of phylogenetically distant groups such as Planctomyces, Verrucomicrobia, and Bacteroidota. This indicated that these particular MAGs possess a similar physiological and metabolic potential regardless of their phylogenetic relationships. In addition, as Arctic diazotroph MAGs did not form their own cluster, they did not possess common functional traits that can distinguish them from those from lower

latitudes. We also examined codon usage among diazotroph MAGs from Arctic and those from other regions, and found no significant differences in codon usage between the two groups (Supplementary Fig. S1).

Global distribution of Arctic diazotroph MAGs

We examined the relative abundance of each MAG in a global metagenome database (Supplementary Fig. S2a) [16]. Samples were assessed based on size fractionation—total (≥0.2 μm; although the viral fraction was not included, this sample is referred to as "total" for convenience), bacterial (0.2–3 μm), and viral (<0.2 μm) fractions—and water column depth—shallow (≤200 m), intermediate (200–1000 m), and deep (≥1000 m) layers. In general, water temperature at the surface varied widely across the sampled oceanic regions, but in the deep sea it was uniformly low in all regions (Supplementary Fig. S2b). Similarly, in the metagenome database, water temperatures in the deep layers showed less variation and lower temperatures (median, 2.02 °C) than in other layers (Supplementary Fig. S2c). The genome abundance of Arctic diazotroph MAGs varied greatly across different size fractions and increased in high latitudes except for Arc-Alpha, which also increased in deep water (≥1000 m) in low latitudes (Fig. 2). In the total fraction (≥0.2 μm), the abundance of Arc-UCYN-A2, Arc-Bactero, Arc-Alpha, and Arc-Gamma-01 was high (maximum CPMM: 479, 355, 3213, and 12841, respectively) especially in the Arctic Ocean. The maximum CPMM of Arc-Gamma-01 means that the genome abundance reached 1.28% of the total microbial community, which is greater than the maximum of those from low latitudes (1.14%, *C. watsonii* of the total community, Supplementary Table S6). Although a high abundance does not necessarily indicate high activity [53], given the importance of N₂ fixation in the biogeochemical cycle at low latitudes, the importance of diazotrophs in the Arctic Ocean is also inferred. In contrast, in the bacterial fraction (0.2–3 μm), most arctic diazotroph MAGs were rarely found. This size-dependent difference in the abundance could be due to the cell size of these bacteria or their symbiotic/particle-attached lifestyle [4, 54–57]. For example, UCYN-A is a symbiotic diazotroph with haptophytes, and its size including its host is >3 μm [4, 54]. Therefore, UCYN-A was detected in the total fraction but was rarely detected in the bacterial fraction. Arc-Bactero, Arc-Alpha, and Arc-Gamma-01 also showed high abundance mainly in >3 μm fraction, suggesting that their lifestyle or cell size may be involved in this size selection as well.
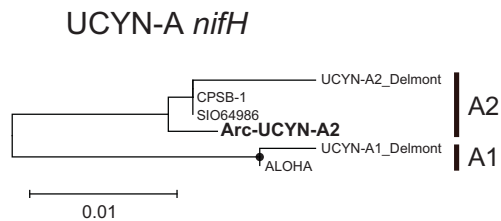
In the viral fraction (<0.2 μm), the relative abundance of Arc-Bactero, Arc-Campylo, Arc-Gamma-02, and Arc-Gamma-03 was particularly high in the Arctic Ocean samples (maximum CPMM: 14291, 62203, 2530, and 152821, respectively). The CPMM value indicates that Arc-Gamma-03 accounted for up to 15.2% of the viral fraction community. Arc-Bactero was also abundant in the total fraction, which may be attributable to its cell shape (i.e., a slender filamentous cell [58]). Meanwhile, as Arc-Campylo, Arc-Gamma-02, and Arc-Gamma-03 were rarely found in the other fractions, their cell size is likely to be very small, and thus they would have passed through a 0.2-μm filter [58]. Interestingly, recent studies reported that *nif* genes are detected in the viral fraction [11, 59], and those are more abundance in the Arctic Ocean [11]. Of the two *nifH* sequences observed in the viral fraction by Pierella Karlusich et al. [11], one was identical to the *nifH* sequence of Arc-Campylo, thus confirming our result.

We further examined the abundance of diazotroph genomes obtained at lower latitudes (Supplementary Fig. S3). Cyanobacterial diazotrophs were abundant mainly at low latitudes in the total and bacterial fractions as shown in previous microscopy and *nifH*- and genome-based studies [11, 15, 60]. The exception was *C. chwakensis*, which was widely found in the Arctic Ocean. Interestingly, some NCDs were as abundant or more abundant in the Arctic Ocean than in low latitudes. For example, the relative

a

nifH   Conserved marker genes   MCR pattern

Bootstrap value ● >80%

b

UCYN-A *nifH*
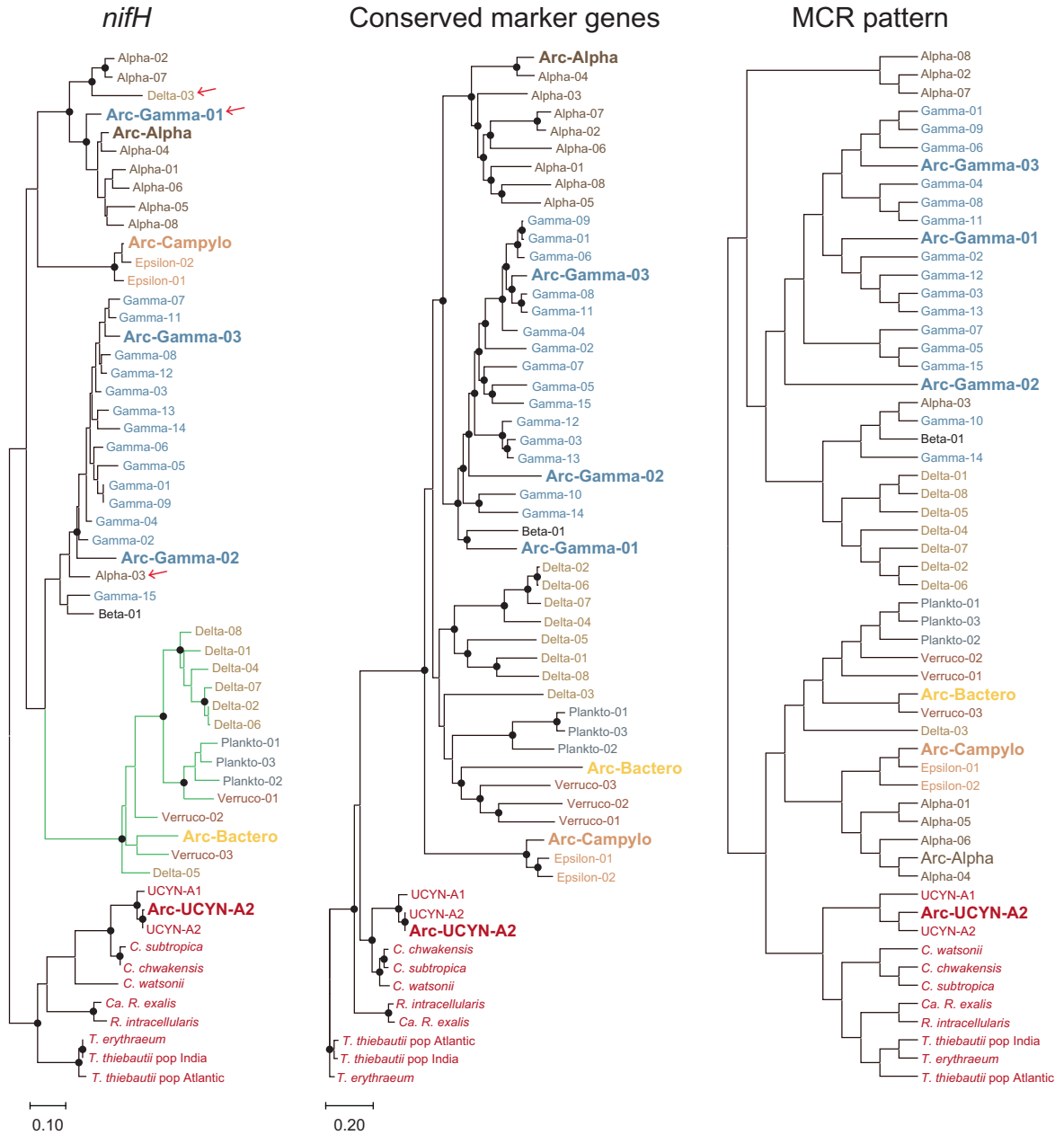
abundance of Alpha-02 and -04 became high in the Arctic Ocean (maximum CPMM: 324 and 843, respectively) in the total fractions. These NCDs also occurred in deep water at low latitudes, and this distribution pattern was similar to that of Arc-Alpha. Among the known diazotrophs, no species was found to have a high relative abundance in only the viral fraction, as was

noted for the Arctic diazotrophs Arc-Campylo, Arc-Gamma-02, and Arc-Gamma-03.

In summary, we noted two general distribution patterns of diazotrophs that exist in the Arctic Ocean: (1) diazotrophs that occur almost exclusively in the Arctic Ocean (i.e., Arctic-endemic: Arc-Bactero, Arc-Campylo, and Arc-Gamma-01, -02, and -03) and (2)

Fig. 1 Phylogeny and functional module clustering of diazotrophs. a Three types of dendrograms representing *nifH* phylogeny, conserved marker gene phylogeny, and hierarchical clustering based on the functional module completion ratio (MCR) [33]. The seven Arctic MAGs listed in Table 1 (shown in bold) and cyanobacterial diazotrophs *C. chwakensis* and *C. subtropica* are included, along with MAGs constructed from a metagenomic dataset from samples collected in the subtropical ocean [14, 15]. Label colors represent phylum- or class-level taxonomy. The *nifH* phylogenetic tree is rooted with cyanobacterial sequences to be consistent with the other trees. The red arrow indicates MAGs with mismatch between *nifH* and conserved marker genes phylogeny. **b** *nifH* phylogenetic tree of UCYN-A, for which nearly the whole genome has been assembled (Supplementary Table S2). The phylogenetic tree was estimated with the maximum-likelihood method based on the full length of the *nifH* sequence and the conserved marker genes using PhyloPhlAn. The branches in the *nifH* phylogenetic trees belonging to Clusters I and III are shown with black and green lines, respectively.
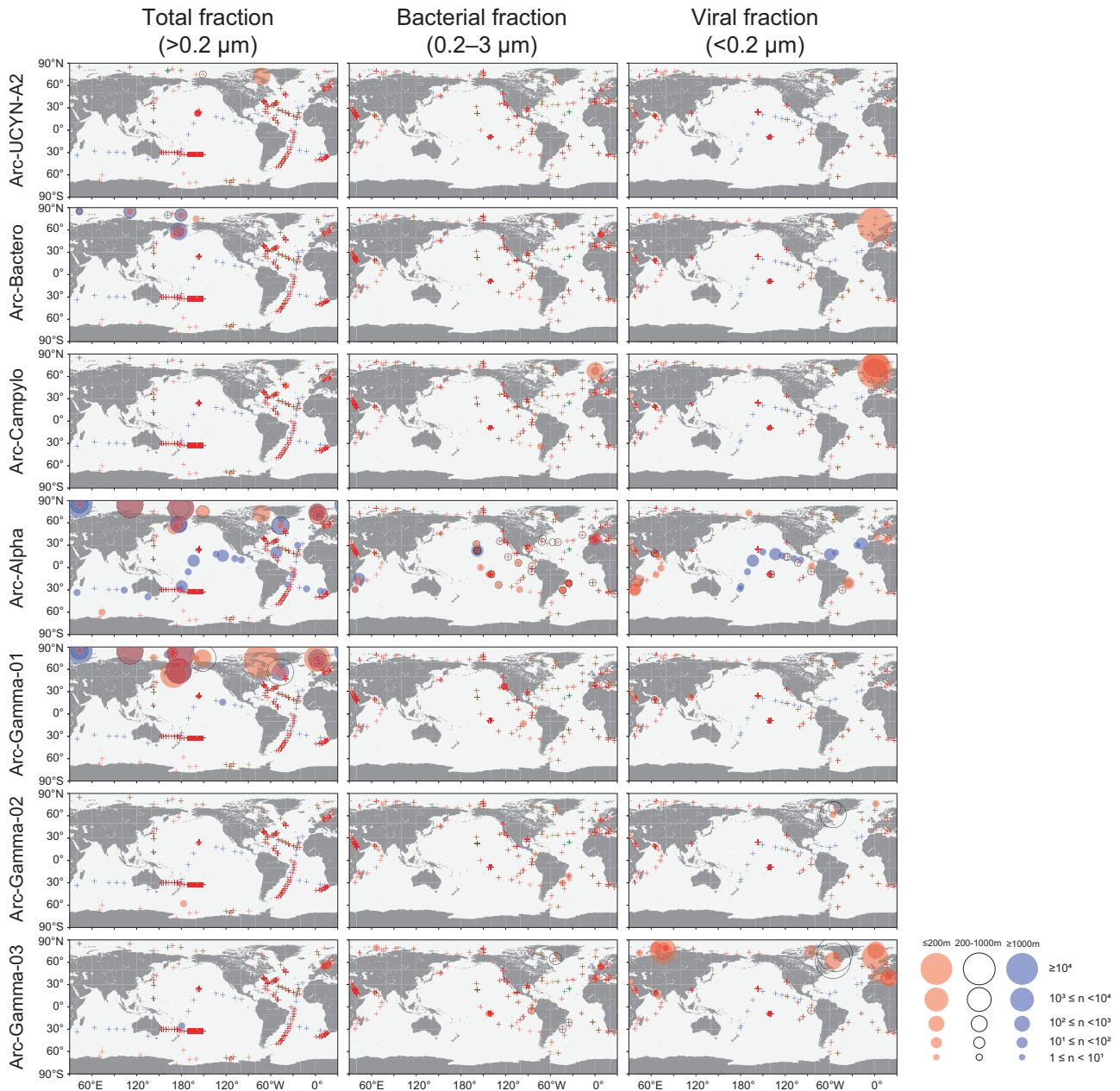

Fig. 2 Depth-resolved abundance of Arctic diazotroph MAGs across the global oceans. The abundance data are divided into three size fractions and three depths. Total, bacterial, and viral fractions indicate ≥0.2, 0.2–3, and <0.2 μm filter size fractions used for sample collection, respectively. Sampled depths are divided into ≤200, 200–1000, and ≥1000 m, and are indicated in red, black open circle, and blue, respectively. The genome abundance was calculated as a length-normalized count per million microbe genomes of a total community (CPMM) in each fraction. The area of each circle is proportional to its indicated abundance. The plus signs indicate the location with CPMM lower than 1.

diazotrophs that occur both in the Arctic Ocean and at lower latitudes (i.e., cosmopolitan: Arc-Alpha, *C. chwakensis*, Alpha-02 and -04, and so on) (Table 1 and Supplementary Fig. S3). The size-fractionation data provided insight into the cell size and/or lifestyle of Arctic diazotrophs; Arc-UCYN-A2, Arc-Bactero, Arc-Alpha, and Arc-Gamma-01 may have a large cell size of >3 μm or have symbiotic/particle-attached lifestyle, and Arc-Campylo and Arc-Gamma-02 and -03 are likely to have a very small cell size passable 0.2 μm filter.

What characteristics allow diazotrophs to occur in the Arctic Ocean?

Arctic-endemic microbes other than diazotrophs have previously been described based on 16S rRNA genes and MAGs [21, 29, 61]. Royo-Llonch et al. [29] recently investigated the genomic characteristics of Arctic-endemic microbes through a large-scale comparison of MAGs from the Arctic Ocean and lower latitudes and showed that Arctic-endemic microbes generally have larger genome sizes and shorter minimum doubling times, implying a copiotrophic lifestyle. The genome size (1.4–4.5 Mb, Table 1) and minimum doubling time (1.0–12 h, Supplementary Table S3), which was estimated using Growthpred [28], of Arctic diazotrophs were not significantly different from those of low-latitude microbes in Royo-Llonch et al. [29], suggesting no trend of a copiotrophic lifestyle. This may not be surprising for diazotrophs because they can thrive in an oligotrophic environment.

Other possible factors that may make microbes more likely to inhabit the Arctic Ocean are their psychrophilic (grow optimally at <15 °C) or psychrotolerant (survive below the freezing point but grow optimally at 20–25 °C) characteristics [21, 62]. Indeed, *Psychromonas* (Arc-Gamma-02) is a representative of the psychrophilic bacteria, many species of which have been isolated from low-temperature marine environments including polar regions and the deep sea (e.g., refs. [63, 64]). Microbes in cold environments generally have specialized proteins that function alone or in combination to adapt to their growing conditions [62, 65, 66]. We examined the genes encoding cold-inducible proteins [65] in each MAG (Supplementary Table S3), but we note that some of these cold-inducible proteins are not found only in cold-adapted microbes but can also be present in thermophilic or mesophilic microbes (e.g., DEAD box protein [67]). On the other hand, among these proteins, the cold-inducible RNA chaperone (CspA), which is a protein involved in maintaining RNA structure at low temperatures, is known as one of the representative proteins shared among microbes in cold environments [62, 65, 66]. *cspA* gene was found in all Arctic diazotroph MAGs except that of Arc-UCYN-A2. Furthermore, among the low-latitude diazotrophs examined here, most of the NCDs had *cspA* (Supplementary Table S3). Obviously, various genes in addition to *cspA* are involved in cold-environment adaptation, but these results suggested that most marine NCDs have the potential to adapt to cold environments. Most microbes except phototrophs can occur from the surface to the deep sea, and thus it is not surprising that NCDs detected at low latitudes also have *cspA* to adapt to the cold environment of the deep sea (median, 2.02 °C in our dataset). In contrast, cyanobacterial diazotrophs do not have *cspA* except for *C. chwakensis* and *C. subtropica*. Cyanobacteria can grow only in the euphotic layer (<200 m), as they need to perform photosynthesis. In addition, cyanobacterial diazotrophs prefer high-temperature (>20 °C) regions [1, 60] and thus do not need to have *cspA*. *C. chwakensis* was found in high abundance in the Arctic Ocean, suggesting that it acquired *cspA* to adapt to the cold environment. Although *C. subtropica* was rarely found in the Arctic Ocean, it is phylogenetically very close to *C. chwakensis* (Fig. 1a and [68]) and may also have *cspA*. The presence of *cspA* in the genome can explain distribution patterns of diazotrophs that are present both in the Arctic Ocean and at lower latitudes; these include not only *C. chwakensis* but also Arc-Alpha and Alpha-02 and -04. However, *cspA* alone does not explain the occurrence of Arctic-endemic diazotrophs. There were no genes encoding cold-inducible proteins that were specific to Arctic diazotroph MAGs with the exception of *ipxP* of Arc-Gamma-02 (Supplementary Table S3), which encodes a protein responsible for lipid A synthesis under low temperatures [65]. Further, the proportion of the seven amino acids (IVYWREL) in the total amino acids which is a signature of optimal growth temperature [36] was not significantly different between Arctic and other diazotroph MAGs (Supplementary Fig. S4). Therefore, in addition to cold-environment adaptations, there are likely to be other reasons for the occurrence of Arctic-endemic diazotrophs.

Environmental uniqueness may also be related to characteristics of Arctic diazotrophs. The Arctic Ocean has sea ice throughout the year and is prone to being stratified despite the cold environment due to the fresh water input from melting sea ice and rivers. Arctic rivers also supply a large amount of terrestrial materials to the ocean. Interestingly, Arc-Gamma-01 has a large number of glycosyltransferase genes (57 genes) as compared with low-latitude gammaproteobacterial diazotrophs (14–43 genes) (Supplementary Table S3). Glycosyltransferase is involved in polysaccharide production [69]. *Crocosphaera* and *Trichodesmium*, which also have a high number of members of this gene family (Supplementary Table S3), produce extracellular polysaccharides (EPSs) and form aggregates [70, 71] in which EPSs assist with adherence [69]. Similarly, Arc-Gamma-01 could produce EPSs and may form aggregates and/or attach itself to sea ice or particles. Another unique feature of Arctic diazotroph MAGs is their diverse aromatic degradation capabilities. The metabolic pathways of aromatic degradation were also found in diazotroph MAGs from low latitude, such as benzoate degradation and catechol cleavage functions in Alpha-07, Gamma-05, -07, and -08 (Supplementary Table S7). However, in addition to these pathways, Arctic-endemic diazotrophs had the pathways of carbazole (Arc-Gamma-01) and salicylate (Arc-Gamma-03) degradation, which were not found any other diazotrophs (Supplementary Table S7). Arctic diazotrophs could be exposed to more aromatic compounds than in other oceans, and thus they may develop the ability to degrade various aromatic compounds. Colatriano et al. [22] recently showed that Chloroflexi, a major bacterial phylum in the Arctic Ocean, may have acquired aromatics degradation genes horizontally from terrestrial bacteria and was subsequently able to grow using material of terrestrial origin. The same thing could have happened to the Arctic diazotrophs. Collectively, the Arctic-endemic diazotrophs may have expanded their metabolic potential to adapt to the unique environment of the Arctic Ocean.

It should be note that some Arctic-endemic diazotrophs (Arc-Campylo and Arc-Gamma-02 and -03) are expected to have a very small cell size (<0.2 μm), as was also inferred from an *nifH*-based study [11]. One characteristic of very small bacteria is a reduced genome size (<2 Mb) [58]. The genome size of Arc-Gamma-03 (3.7 Mb) is smaller than that of the isolated species in the genus *Oceanobacter* (4.5 and 5.1 Mb), but it is not particularly small. Those of Arc-Gamma-02 (4.1 Mb) and Arc-Campylo (2.9 Mb) are within the range of the genus *Psychromonas* (3.9–5.5 Mb) and *Arcobacter* (2.2–3.2 Mb). Therefore, these Arctic diazotrophs with potentially small cell sizes do not have small genomes. In contrast, the cell size of the Arctic-endemic diazotrophs is presumably smaller than that of isolated non-diazotroph species within the same lineage [63, 64, 72–75]. In general, bacteria in the field tend to have smaller cell sizes than do cultured strains due to nutrient limitations and predation [76]. Given that variations in cell size were inferred but appreciable variations in genome size were not, one possibility is that environmental stress might reduce the cell size of the Arctic-endemic diazotrophs.

Comparative genome analysis of UCYN-A

We obtained a nearly complete genome of UCYN-A2 from the Arctic Ocean. The *nifH* phylogenetic tree using full-length *nifH* showed that the *nifH* sequence of Arc-UCYN-A2 is distinct from that of low-latitude UCYN-A2 (Fig. 1b), which contradicts the conclusion of a previous study [4]. This is because the previous study examined only a short PCR-amplified region of the gene. Indeed, the *nifH* sequence targeted by PCR of Arc-UCYN-A2 is identical to that of the low-latitude one.

We further performed a comparative genomic analysis using existing UCYN-A genomes (Supplementary Table S2 and Supplementary Fig. S5). As noted previously [77], substantial differences were found between the UCYN-A1 and -A2 genomes (Supplementary Fig. S5). In contrast, the intraclade difference was very small; the Arc-UCYN-A2 genome was almost identical to that of

UCYN-A2 at low latitudes (ANI 99.42–99.79% as compared with low-latitude UCYN-A2). However, we found that Arc-UCYN-A2 had a *gph* gene, which is used in DNA repair [78], that was absent from the low-latitude UCYN-A2. No differences were found except for this *gph* gene when comparing Arc-UCYN-A2 with the low-latitude UCYN-A2 genome. Regarding the *gph* gene, cold-adapted microbes tend to have more genes to repair DNA damage caused by reactive oxygen species, which generally increase inside the cell in cold environments [65]. Although Arc-UCYN-A2 does not have the *cspA* gene, the *gph* gene in Arc-UCYN-A2 may be used for another strategy for adaptation to the cold Arctic environment. Further, Arc-UCYN-A2 has a symbiotic lifestyle [4], and so they may receive some benefit from their host for adaptation to low temperatures, which was not coded in their genome.

nifH gene abundance among diazotroph MAGs isolated from Arctic Ocean samples

We quantified the *nifH* copy number of MAGs that showed a particularly high genome abundance in the total fraction of Arctic Ocean and of Arc-UCYN-A2 using qPCR technique. We collected the samples in the Pacific side of the Arctic Ocean in summer for three years. The observed sample sites were located mainly in open-water areas. The sea surface temperature ranged from −1.2 to 7.6 °C (Supplementary Table S8). Although nitrogenous nutrients were sporadically high (>1 μM) in the surface water of the Bering Strait, they were generally depleted (<0.1 μM) in the north of 70°N (Supplementary Table S8). Of the targeted diazotrophs, Arc-UCYN-A2, Arc-Alpha, and Alpha-04 were detected in samples collected during each cruise, indicating that these diazotrophs were indeed present in the Arctic Ocean. In contrast, Arc-Bactero, Arc-Gamma-01, *C. chwakensis*, and Alpha-02 were not detected in samples from either cruise. This absence might be related to their habitat area and season, because samples were collected only in the Pacific side open-water area in summer. For example, Arc-Gamma-01 could produce EPSs as mentioned above and thus could attach to the sea ice or particles, which would preclude its sampling from open-water samples. The *nifH* of *C. chwakensis* (former genus *Cyanothece*) was detected in the Atlantic side of the Arctic Ocean [79].

UCYN-A2 was found in samples from most of the stations during each cruise and was distributed vertically (Fig. 3a). The maximum *nifH* abundance of UCYN-A2 was 2.9×10^3, 5.8×10^5,
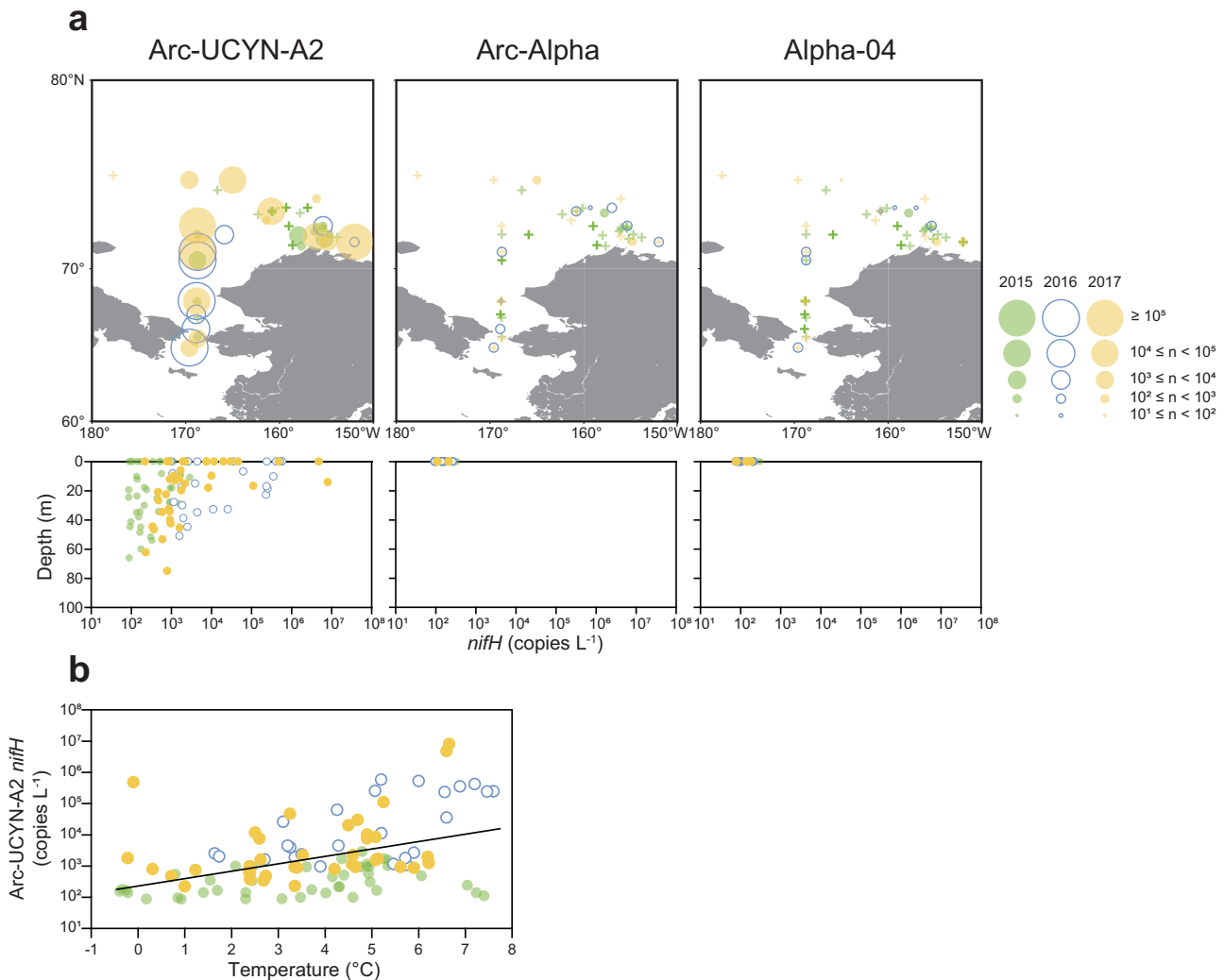

Fig. 3   *nifH* abundance of diazotroph MAGs in the Arctic Ocean. a. Spatial and vertical distribution of *nifH* abundance of Arc-UCYN-A2, Arc-Alpha, and Alpha-04. The upper panels show the spatial distribution of the maximum of each *nifH* abundance in 2015 (green), 2016 (blue circle), and 2017 (yellow). The lower panels show the vertical distribution of each *nifH* abundance by the year. The *nifH* of Arc-Alpha and Alpha-04 were not detected except the surface water. The *nifH* of Arc-Bactero, Arc-Gamma-01, Alpha-02, and *C. chwakensis* were not detected in our study region. **b.** Relationship between *nifH* abundance of Arc-UCYN-A2 and temperature. The regression line was generated from data pooled from all years (2015 (green), 2016 (blue circle), and 2017 (yellow)).

and 8.0×10^6 copies L^{-1} in 2015, 2016, and 2017, respectively, and was found near the surface. Arc-Alpha and Alpha-04 were widely distributed in the study region but were detected only at the surface. The maximum for Arc-Alpha (3.3×10^2 copies L^{-1}) and Alpha 04 (3.0×10^2 copies L^{-1}) was significantly lower than that of UCYN-A2 ($p < 0.05$). We examined the relationship between *nifH* abundance and environmental parameters and found that UCYN-A2 had a significant positive correlation with temperature ($p < 0.05$), although other diazotrophs had no significant relationship with any environmental parameters ($p > 0.05$). This relationship between UCYN-A abundance and temperature was also reported at low latitudes, although the temperature range (22–30 °C) differed from that in the Arctic Ocean (–0.4–7.6 °C) [80, 81]. Hence, the controlling factor seems to be the same even if the clade is different. The positive correlation with temperature indicated that UCYN-A2 can expand its habitat range and should be able to increase in the Arctic Ocean with future warming.

CONCLUSION
This study successfully reconstructed the diazotroph genomes from metagenomic data in the Arctic Ocean, some of which would not have been detected by a conventional PCR-based approach. Discovery of diazotroph within phylum Bacteroidota indicate that diazotrophs are more diverse in the ocean than previously thought. We found a MAG in which *nifH* gene was detected, but no other *nif* genes were present. This result implies that the detection of *nifH* alone cannot necessarily confirm the presence of diazotrophs in a marine environment (c.f. [13]). In addition, our results inferred that horizontal gene transfer of *nif* genes also occur in the ocean.

We show the distribution and survival strategies of diazotrophs in the Arctic Ocean based on the MAG information. Although the environments in the Arctic and tropics and subtropics are very different, we first found cosmopolitan diazotrophs that were distributed all over the world that have the potential for cold adaptation. In contrast, we also found Arctic-endemic diazotrophs, which have specific genes that help them adapt to the Arctic environment. Considering that cyanobacterial diazotrophs occur mainly in low latitudes, the biogeographic distribution of diazotrophs follows one of three patterns: low-latitude-endemic, Arctic-endemic, and cosmopolitan.

The Arctic-occurring diazotrophs are remarkably abundant among the total metagenomic reads from the Arctic Ocean samples, indicating that they are important species in the Arctic microbial community and thus are likely to substantially contribute to biogeochemical cycles. We further showed that diazotrophs detected in the 0.2 µm passable fraction are particularly abundant in the Arctic Ocean, indicating that the standard (and even recently improved) method of using GF/F filters (pore size, 0.3–0.7 µm) [82–84] can markedly underestimate the rate of N_2 fixation. The Arctic Ocean is one of the most rapidly changing oceans on Earth, and N_2 fixation in the pan-Arctic Ocean is still unknown. Our findings increase our understanding of current and possible future Arctic N_2 fixation.

DATA AVAILABILITY
All data used in this study are available in the supplementary materials or the UTokyo Repository (https://doi.org/10.15083/0002005808).

REFERENCES
1. Sohm JA, Webb EA, Capone DG. Emerging patterns of marine nitrogen fixation. Nat Rev Microbiol. 2011;9:499–508.
2. Blais M, Tremblay J-É, Jungblut AD, Gagnon J, Martin J, Thaler M, et al. Nitrogen fixation and identification of potential diazotrophs in the Canadian Arctic. Glob Biogeochem Cycles. 2012;26:GB3022.
3. Shiozaki T, Fujiwara A, Ijichi M, Harada N, Nishino S, Nishi S, et al. Diazotroph community structure and the role of nitrogen fixation in the nitrogen cycle in the Chukchi Sea (western Arctic Ocean). Limnol Oceanogr. 2018;63:2191–205.
4. Harding K, Turk-Kubo KA, Sipler RE, Mills MM, Bronk DA, Zehr JP. Symbiotic unicellular cyanobacteria fix nitrogen in the Arctic Ocean. Proc Natl Acad Sci USA. 2018;115:13371–5.
5. Shiozaki T, Fujiwara A, Inomura K, Hirose Y, Hashihama F, Harada N. Biological nitrogen fixation detected under Antarctic sea ice. Nat Geosci 2020;13:729.
6. Timmermans M-L, Labe Z. Sea Surface Temperature. Arctic Report Card 2022, Druckenmiller, ML, Thoman, RL, and Moon, TA, Eds. 2022, https://doi.org/10.25923/p493-2548.
7. Aagaard K, Carmack EC. The role of sea ice and other fresh-water in the Arctic circulation. J Geophys Res Oceans. 1989;94:14485–98.
8. Tremblay JE, Anderson LG, Matrai P, Coupel P, Belanger S, Michel C, et al. Global and regional drivers of nutrient supply, primary production and CO_2 drawdown in the changing Arctic Ocean. Progr Oceanogr. 2015;139:171–96.
9. Karl DM, Church MJ. Microbial oceanography and the Hawaii Ocean Time-series programme. Nat Rev Microbiol. 2014;12:699–713.
10. Zehr JP, Jenkins BD, Short SM, Steward GF. Nitrogenase gene diversity and microbial community structure: a cross-system comparison. Environ Microbiol. 2003;5:539–54.
11. Pierella Karlusich JJ, Pelletier E, Lombard F, Carsique M, Dvorak E, Colin S, et al. Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. Nat Commun. 2021;12:4160.
12. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genom. 2012;13:162.
13. Mise K, Masuda Y, Senoo K, Itoh H. Undervalued pseudo-*nifH* Sequences in public databases distort metagenomic insights into biological nitrogen fixers. mSphere. 2021;6:e00785–21.
14. Delmont TO, Quince C, Shaiber A, Esen OC, Lee STM, Rappe MS, et al. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. Nat Microbiol. 2018;3:804–13.
15. Delmont TO, Karlusich JJP, Veseli I, Fuessel J, Eren AM, Foster RA, et al. Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. ISME J. 2022;16:1203.
16. Nishimura Y, Yoshizawa S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. Sci Data. 2022;9:305.
17. Zehr JP, Shilova IN, Farnelid HM, Munoz-Marin MC, Turk-Kubo KA. Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. Nat Microbiol. 2017;2:16214.
18. Turk-Kubo KA, Farnelid HM, Shilova IN, Henke B, Zehr JP. Distinct ecological niches of marine symbiotic N₂-fixing cyanobacterium *Candidatus* Atelocyanobacterium thalassa sublineages. J Phycol. 2017;53:451–61.
19. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh HJ, Cuenca M, et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell. 2019;179:1068–83.
20. Gregory AC, Zayed AA, Conceicao-Neto N, Temperton B, Bolduc B, Alberti A, et al. Marine DNA viral macro- and microdiversity from pole to pole. Cell. 2019;177:1109–23.
21. Cao SN, Zhang WP, Ding W, Wang M, Fan S, Yang B, et al. Structure and function of the Arctic and Antarctic marine microbiota as revealed by metagenomics. Microbiome. 2020;8:47.
22. Colatriano D, Tran PQ, Gueguen C, Williams WJ, Lovejoy C, Walsh DA. Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. Commun Biol. 2018;1:90.
23. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
24. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003;31:371–3.
25. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the genome taxonomy database. Bioinformatics. 2020;36:1925–7.
26. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The Pfam protein families database. Nucleic Acids Res. 2002;30:276–80.
27. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36:2251–2.
28. Vieira-Silva S, Rocha EPC. The systemic imprint of growth and its uses in ecological (meta) genomics. PLoS Genet. 2010;6:e1000808.
29. Royo-Llonch M, Sanchez P, Ruiz-Gonzalez C, Salazar G, Pedros-Alio C, Sebastian M, et al. Compendium of 530 metagenome-assembled bacterial and archaeal genomes from the polar Arctic Ocean. Nat Microbiol. 2021;6:1561–74.

30. Arai W, Taniguchi T, Goto S, Moriya Y, Uehara H, Takemoto K, et al. MAPLE 2.3.0: an improved system for evaluating the functionomes of genomes and metagenomes. Biosci Biotech Biochem. 2018;82:1515–7.

31. Takami H. MAPLE Enables Functional Assessment of Microbiota in Various Environments. In: Gojobori T, Wada T, Kobayashi T, Mineta K (eds). Marine metagenomics-technological aspects and applications. (Springer, 2019) pp 85-119.

32. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.

33. Takami H, Taniguchi T, Moriya Y, Kuwahara T, Kanehisa M, Goto S. Evaluation method for the potential functionome harbored in the genome and metagenome. BMC Genom. 2012;13:699.

34. Takami H, Taniguchi T, Arai W, Takemoto K, Moriya Y, Goto S. An automated system for evaluation of the potential functionome: MAPLE version 2.1.0. DNA Res. 2016;23:467–75.

35. R_Core_Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistica Computing; 2021; https://www.R-project.org/.

36. Zeldovich KB, Berezovsky IN, Shakhnovich EI. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol. 2007;3:e5.

37. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. Nat Microbiol. 2021;6:3–6.

38. Tamura K, Stecher G, Kumar S. MEGA11 molecular evolutionary genetics analysis version 11. Mol Biol Evol. 2021;38:3022–7.

39. Asnicar F, Thomas AM, Beghini F, Mengoni C, Manara S, Manghi P, et al. Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0. Nat Commun. 2020;11:2500.

40. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. PeerJ Preprints. 2016;4:e1900v1.

41. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

42. Kultima JR, Coelho LP, Forslund K, Huerta-Cepas J, Li SS, Driessen M, et al. MOCAT2: a metagenomic assembly, annotation and profiling framework. Bioinformatics 2016;32:2520–3.

43. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 2014;30:923–30.

44. Liao Y, Smyth GK, Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Res. 2013;41:e108.

45. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131:281–5.

46. Turk-Kubo KA, Gradoville MR, Cheung S, Cornejo-Castillo F, Harding KJ, Morando M, et al. Non-cyanobacterial diazotrophs: Global diversity, distribution, ecophysiology, and activity in marine waters. FEMS Microbiol Rev. 2022.

47. Fernandez-Mendez M, Turk-Kubo KA, Buttigieg PL, Rapp JZ, Krumpen T, Zehr JP, et al. Diazotroph diversity in the sea ice, melt ponds, and surface waters of the eurasian basin of the central Arctic Ocean. Front Microbiol. 2016;7:1884.

48. Jabir T, Vipindas PV, Krishnan KP, Hatha AAM. Abundance and diversity of diazotrophs in the surface sediments of Kongsfjorden, an Arctic fjord. World J Microbiol Biotechnol. 2021;37:41.

49. Acinas SG, Sanchez P, Salazar G, Cornejo-Castillo FM, Sebastian M, Logares R, et al. Deep ocean metagenomes provide insight into the metabolic architecture of bathypelagic microbial communities. Commun Biol. 2021;4:604.

50. Kechris KJ, Lin JC, Bickel PJ, Glazer AN. Quantitative exploration of the occurrence of lateral gene transfer by using nitrogen fixation genes as a case study. Proc Natl Acad Sci USA. 2006;103:9584–9.

51. Bolhuis H, Severin I, Confurius-Guns V, Wollenzien UI, Stal LJ. Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium Microcoleus chthonoplastes. ISME J. 2010;4:121–30.

52. Takami H, Arai W, Takemoto K, Uchiyama I, Taniguchi T. Functional classification of uncultured "Candidatus Caldiarchaeum subterraneum" using the MAPLE System. PLoS One. 2015;10:e0132994.

53. Munson-McGee JH, Lindsay MR, Sintes E, Brown JM, D'Angelo T, Brown J, et al. Decoupling of respiration rates and abundance in marine prokaryoplankton. Nature. 2022;612:764–70.

54. Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, et al. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. Science. 2012;337:1546–50.

55. Geisler E, Bogler A, Rahav E, Bar-Zeev E. Direct detection of heterotrophic diazotrophs associated with planktonic aggregates. Sci Rep. 2019;9:9288.

56. Cornejo-Castillo FM, Zehr JP. Intriguing size distribution of the uncultured and globally widespread marine non-cyanobacterial diazotroph Gamma-A. ISME J. 2021;15:124–8.

57. Harding KJ, Turk-Kubo KA, Mak EWK, Weber PK, Mayali X, Zehr JP. Cell-specific measurements show nitrogen fixation by particle-attached putative non-cyanobacterial diazotrophs in the North Pacific Subtropical Gyre. Nat Commun. 2022;13:6979.

58. Nakai R. Size Matters: Ultra-small and filterable microorganisms in the environment. Microbes Environ. 2020;35:ME20025.

59. Lannes R, Olsson-Francis K, Lopez P, Bapteste E. Carbon fixation by marine ultrasmall prokaryotes. Genome Biol Evol. 2019;11:1166–77.

60. Zehr JP, Capone DG. Changing perspectives in marine nitrogen fixation. Science. 2020;368:eaay9514.

61. Ghiglione JF, Galand PE, Pommier T, Pedros-Alio C, Maas EW, Bakker K, et al. Pole-to-pole biogeography of surface and deep marine bacterial communities. Proc Natl Acad Sci USA. 2012;109:17633–8.

62. De Maayer P, Anderson D, Cary C, Cowan DA. Some like it cold: understanding the survival strategies of psychrophiles. EMBO Rep. 2014;15:508–17.

63. Auman AJ, Breezee JL, Gosink JJ, Kampfer P, Staley JT. Psychromonas ingrahamii sp nov., a novel gas vacuolate, psychrophilic bacterium isolated from Arctic polar sea ice. Int J Syst Evol Microbiol. 2006;56:1001–7.

64. Miyazaki M, Nogi Y, Fujiwara Y, Horikoshi K. Psychromonas japonica sp nov., Psychromonas aquimarina sp nov., Psychromonas macrocephali sp nov and Psychromonas ossibalaenae sp. nov., psychrotrophic bacteria isolated from sediment adjacent to sperm whale carcasses off Kagoshima, Japan. Int J Syst Evol Microbiol. 2008;58:1709–14.

65. Barria C, Malecki M, Arraiano CM. Bacterial adaptation to cold. Microbiology 2013;159:2437–43.

66. Siddiqui KS, Williams TJ, Wilkins D, Yau S, Allen MA, Brown MV, et al. Psychrophiles. Annu Rev Earth Planet Sci. 2013;41:87–115. Vol 41

67. Nagaoka E, Hidese R, Imanaka T, Fujiwara S. Importance and determinants of induction of cold-induced DEAD RNA helicase in the hyperthermophilic archaeon Thermococcus kodakarensis. J Bacteriol. 2013;195:3442–50.

68. Mares J, Johansen JR, Hauer T, Zima J, Ventura S, Cuzman O, et al. Taxonomic resolution of the genus Cyanothece (Chroococcales, Cyanobacteria), with a treatment on Gloeothece and three new genera, Crocosphaera, Rippkaea, and Zehria. J Phycol. 2019;55:578–610.

69. Yakovlieva L, Walvoort MTC. Processivity in bacterial glycosyltransferases. ACS Chem Biol. 2020;15:3–16.

70. Berman-Frank I, Rosenberg G, Levitan O, Haramaty L, Mari X. Coupling between autocatalytic cell death and transparent exopolymeric particle production in the marine cyanobacterium Trichodesmium. Environ Microbiol. 2007;9:1415–22.

71. Sohm JA, Edwards BR, Wilson BG, Webb EA. Constitutive extracellular polysaccharide (EPS) production by specific isolates of Crocosphaera watsonii. Front Microbiol. 2011;2:229.

72. Bowditch RD, Baumann L, Baumann P. Description of Oceanospirillum kriegii sp. nov. and O. jannaschii sp. nov. and assignment of two species of Alteromonas to this genus as O. ommune comb. nov. and O. vagum comb. nov. Curr Microbiol. 1984;10:221–9.

73. Huang ZB, Huang YY, Lai QL, Chen XL, Dong CM, Huang XZ. Oceanobacter mangrovi sp. nov., a novel poly-β-hydroxybutyrate accumulating bacterium isolated from mangrove sediment. Curr Microbiol. 2022;79:100.

74. Collado L, Cleenwerck I, Van Trappen S, De Vos P, Figueras MJ. Arcobacter mytili sp. nov., an indoxyl acetate-hydrolysis-negative bacterium isolated from mussels. Int J Syst Evol Microbiol. 2009;59:1391–6.

75. Collado L, Levican A, Perez J, Figueras MJ. Arcobacter defluvii sp. nov., isolated from sewage samples. Int J Syst Evol Microbiol. 2011;61:2155–61.

76. Young KD. The selective value of bacterial shape. Microbiol Mol Biol Rev. 2006;70:660–703.

77. Bombar D, Heller P, Sanchez-Baracaldo P, Carter BJ, Zehr JP. Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. ISME J. 2014;8:2530–42.

78. Pellicer MT, Nunez MF, Aguilar J, Badia J, Baldoma L. Role of 2-phosphoglycolate phosphatase of Escherichia coli in metabolism of the 2-phosphoglycolate formed in DNA repair. J Bacteriol. 2003;185:5815–21.

79. Diez B, Bergman B, Pedros-Alio C, Anto M, Snoeijs P. High cyanobacterial nifH gene diversity in Arctic seawater and sea ice brine. Environ Microbiol Rep. 2012;4:360–6.

80. Moisander PH, Beinart RA, Hewson I, White AE, Johnson KS, Carlson CA, et al. Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. Science. 2010;327:1512–4.

81. Shiozaki T, Chen YLL, Lin YH, Taniuchi Y, Sheu DS, Furuya K, et al. Seasonal variations of unicellular diazotroph groups A and B, and Trichodesmium in the northern South China Sea and neighboring upstream Kuroshio Current. Cont Shelf Res. 2014;80:20–31.

82. Montoya JP, Voss M, Kahler P, Capone DG. A simple, high-precision, high-sensitivity tracer assay for N_2 fixation. Appl Environ Microbiol. 1996;62:986–93.

83. Mohr W, Grosskopf T, Wallace DWR, LaRoche J. Methodological underestimation of oceanic nitrogen fixation rates. PLoS One. 2010;5:e12583.

84. Bombar D, Paerl RW, Anderson R, Riemann L. Filtration via conventional glass fiber filters in $^{15}N_2$ tracer assays fails to capture all nitrogen-fixing prokaryotes. Front Mar Sci. 2018;5:6.

AUTHOR CONTRIBUTIONS
TS designed the study. YN performed most of the bioinformatics analysis. TS, SY, HT, and KH contributed to the bioinformatics analysis. AF, SN, and NH contributed to the coordination of the research cruises. TS and AF collected samples on board. TS performed the experiments. TS wrote the manuscript, with contributions from all co-authors.

COMPETING INTERESTS
The authors declare no competing interests.

ADDITIONAL INFORMATION
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41396-023-01424-x.

Correspondence and requests for materials should be addressed to Takuhei Shiozaki.

Reprints and permission information is available at http://www.nature.com/reprints

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