Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods

Running title: Diazotrophs in *Tara* Oceans

Juan José Pierella Karlusich[1,2], Eric Pelletier[2,3,], Madeline Carsique[1], Etienne Dvorak[1], Sébastien Colin[4], Marc Picheral[2,5], Rainer Pepperkok[2,6], Eric Karsenti[1,2,6], Colomban de Vargas[2,4], Fabien Lombard[2,5,7], Patrick Wincker[2,3], Chris Bowler[1,2*], Rachel A Foster[8*]

[1] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

[2] CNRS Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/ *Tara* Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France

[3] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

[4] Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR 7144, ECOMAP, 29680 Roscoff, France

[5] Sorbonne Universités, CNRS, Laboratoire d'Océanographie de Villefranche (LOV), 06230 Villefranche-sur-Mer, France

[6] European Molecular Biology Laboratory, Heidelberg, Germany

[7] Institut Universitaire de France (IUF), Paris, France

[8] Department of Ecology, Environment and Plant Sciences, Stockholm University Stockholm Sweden

*corresponding authors: Rachel Foster (rachel.foster@su.se) and Chris Bowler (cbowler@bio.ens.psl.eu)

keywords: *Tara* Oceans, diazotroph, symbioses, diatom diazotroph associations, *Richelia*, *Trichodesmium*, non-cyanobacterial diazotrophs, spheroid bodies, ultrasmall bacteria

Abstract

Biological nitrogen fixation sustains ~50% of ocean primary production. However, our understanding of marine N_2-fixers (diazotrophs) is hindered by limited observations. Here, we developed a quantitative image analysis pipeline in concert with mapping of molecular markers for mining >2,000,000 images and >1,300 metagenomes from *Tara* Oceans, covering surface, deep chlorophyll maximum and mesopelagic layers across 6 organismal size fractions (0-2000 µm). Imaging and molecular data were remarkably congruent. Diazotrophs were detected from ultrasmall bacterioplankton (<0.2 µm) to mesoplankton (180 to 2000 µm). We identified several new high density regions of diazotrophs. Distributional and abundance patterns support the previous canonical view that larger sized diazotrophs (>10 µm) dominate the tropical belts, while unicellular diazotrophs were found in surface and mesopelagic samples. Multiple co-occurring diazotrophic lineages were frequently encountered, suggesting that complex overlapping niches are common. Overall, this work provides an updated global snapshot of marine diazotroph biogeographical diversity and highlights new sources and sinks of diazotroph-fueled new production.

Introduction

Approximately half of global primary production occurs in the oceans [1], fueling marine food webs, plankton decomposition and sequestration of fixed carbon to the ocean interior. Marine primary production is often limited by nitrogen (N) in the vast expanses of the open ocean (approximately 75% of surface ocean) [2,3]. In these regions, the biological reduction of di-nitrogen gas (N_2) to bioavailable N, a process called biological N_2 fixation (BNF), is a critical source of new N to the ecosystem and ultimately controls the uptake and sequestration of carbon dioxide (CO_2) [4–6].

In the upper sunlit ocean, the majority of BNF is mediated by a few groups of N_2-fixing (diazotrophic) cyanobacteria. Traditionally it was thought that marine BNF was largely restricted to the subtropical and tropical oceans and was predominantly mediated by relatively larger sized cyanobacterial organisms and holobionts (> 10 µm) such as colony-forming non-heterocystous *Trichodesmium* spp., and heterocystous cyanobacteria forming symbioses with diatoms, also called diatom diazotroph associations (DDAs) (*Richelia intracellularis, Calothrix rhizosoleniae*, hereafter *Richelia* and *Calothrix*) [7]. More recently, unicellular cyanobacteria (UCYN) have been detected in environmental samples outside the tropical belts by qPCR targeting the BNF marker gene *nifH* [8]. One of these UCYN groups is *Candidatus* Atelocyanobacterium thalassa (hereafter UCYN-A). Three UCYN-A strains (A-1, A-2, A-3) live in symbiosis with a small single celled eukaryote (haptophyte) [9–13]. UCYN-B is another unicellular group that is most closely related to *Crocosphaera watsonii* (hereafter *Crocosphaera*). UCYN-B lives singly, colonially or in symbioses with a large chain-forming diatom [14–16]. UCYN-C is the third marine unicellular group identified thus far by *nifH* sequence, and is most closely related to the free-living unicellular diazotroph *Cyanothece sp.* ATCC 51142 [17]. Finally, non cyanobacterial diazotrophs (NCDs), including Archaea and Bacterial lineages, co-occur with the cyanobacterial diazotrophs in the surface ocean and additionally below the photic layer. The distribution and *in situ* activity of NCDs are poorly constrained and difficult to estimate [18–21].

Luo et al [22] compiled the first database of diazotroph abundance in the global ocean for the MARine Ecosystem DATa (MAREDAT) project, consisting of 44 datasets (1966-2011), including microscopy-based counts and *nifH* qPCR studies. It should be noted that not all diazotrophs can be identified by microscopy, and qPCR has limitations, thus microscopy counts are only for *Trichodesmium*, *Richelia* and *Calothrix*, while the qPCR datasets additionally contain information about UCYN-A, UCYN-B and UCYN-C (Supplementary Fig. S1a). Recently, Tang and Cassar [23] updated the MAREDAT dataset with 17 additional qPCR datasets (2012-2018), resulting in more than doubling of the *nifH* observations. Both MAREDAT and the updated version have low coverage in vast regions of the global ocean, including the Mediterranean Sea (MS), the Red Sea (RS), the Arctic Ocean (AO), the Indian Ocean (IO), the South Atlantic Ocean (SAO) and the western Equatorial Pacific Ocean near South America (Supplementary Fig. S1a). Several of these poorly sampled areas were sampled during the *Tara* Oceans circumnavigation (2009-2013) [24] (Supplementary Fig. S1b).

Tara Oceans collected plankton samples separated into discrete size fractions using a serial filtration system [25]; some samples were used to generate parallel molecular and imaging datasets. The *Tara* Oceans gene catalog from samples enriched in free-living prokaryotes is based on the assembly of metagenomes and is highly comprehensive [26,27]. However, the larger plankton size fractions enriched in eukaryotes are genomically much more complex, and thus current *Tara* Oceans gene catalogs from these fractions are based only on poly-A-tailed eukaryotic RNA [28,29]. Hence, the prokaryotes from these larger size fractions have been unstudied and limited to specific taxa based on these poly-A assembled sequences [30–32]. The *Tara* Oceans imaging dataset [33] is also underutilized, especially due to the lack of well-established workflows. Overall, the cyanobacterial diazotrophs, especially those with diverse life histories (colonial, symbiotic, chain formers), have been poorly characterized (with the exception of UCYN-A; see [13,18,26,30,32,34]).

Here, we identify the diversity, abundance and distribution of symbiotic, colony-forming, and particle-associated diazotrophs in the World's ocean based on the *nifH* gene normalized to the bacterial single-copy housekeeping gene *recA* from >1,300 *Tara* Oceans metagenomes [26,28,29]. In parallel, we trained an image classification model and utilized it with the *in situ* images from an Underwater Vision Profiler (UVP) [35] and confocal microscopy [33] to generate a versatile analytical pipeline from images to genomics and genomics to images. Combined, our results provide an improved global overview of diazotroph abundances, diversity, and distribution (vertical and horizontal), as well as the environmental factors that shape these patterns.

Results and Discussion

Diazotroph abundance and biovolume based on imaging methods

We first used machine learning tools (see Methods; [33]) to search for diazotrophs in the *Tara* Oceans high-throughput confocal imaging dataset derived from 61 samples of the 20-180 µm plankton size fraction collected at 48 different sampling locations (Supplementary Fig. S2). We obtained >400 images of DDAs and almost 600 images of *Trichodesmium* free filaments (Fig. 1); all images were from the tropical and subtropical regions and were consistent with the molecular analyses and detected in several new locations not previously reported in diazotroph databases (see below; Fig. 2, Supplementary Figs. S3 and S4). In addition, we detected *Crocosphaera*-like colonies as well as the lesser-studied symbiosis between this diazotrophic cyanobacterium and the centric diatom *Climacodium* [14–16] on the Pacific side of the Panama Canal (TARA_140) (Fig. 1). It should be noted that there were only a few *Crocosphaera* cells (1-2 cells) seemingly embedded in the dense chloroplast fluorescence of each *Climacodium* host, demonstrating the high resolution of our image recognition model.

Abundance ranges based on number of images for the 3 main DDAs, *Hemiaulus-Richelia*, *Rhizosolenia-Richelia*, and *Chaetoceros-Calothrix,* were low and are representative of background densities (e.g., 1.5-20 symbiotic cells L^{-1}) (Fig. 2 and Supplementary Fig. S3a). The low densities and detection, especially C*haetoceros-Calothrix* which can form long chains (> 50 cells chain^{-1}) and the larger *Rhizosolenia-Richelia* symbioses, were not surprising given the pre-filtration step (180-µm mesh) in the sampling protocol that would exclude larger cells and chains. Although

Hemiaulus-Richelia was the most frequently detected, its chains were often short, and sometimes cell integrity was compromised. Variation in the number and length of the symbiont filaments (trichomes) was also observed (Fig. 3). This included observations of free *Richelia* and *Calothrix* filaments (Supplementary Fig. S5), which are rarely reported in the literature [36], but are not unexpected for facultative symbionts, which is the case for *Calothrix* and *Richelia* symbionts of *Chaetoceros* and *Rhizosolenia,* respectively [37,38].

DDAs were broadly distributed and detected in several new locations not previously reported in diazotroph databases [22,23]. These areas included several different stations of the IO, southwest SAO, the South Pacific gyre, and the Pacific side of the Panama Canal (Supplementary Fig. S3). Free *Richelia/Calothrix* filaments could also be quantified in the same samples and regions, as well as at station TARA_39 (IO), where symbiotic hosts were not observed (Supplementary Fig. S5). DDAs were also concentrated in surface samples, with the exception of two deeper samples showing *Hemiaulus-Richelia* densities as high as in the surface: one sample from 108 m at station ALOHA (TARA_131) and a second from 38 m at TARA_143 (Gulf Stream, North Atlantic) (Fig. 2b). Seasonal blooms of DDAs are well known at station ALOHA, with observations of DDAs in moored sediment traps below the photic zone [39–41]. However, observations of symbiotic diatoms in the Gulf Stream are more rare [42].

Observations of free filaments of *Trichodesmium* (1-40 filaments L^{-1}) co-occurred with DDAs in most stations from the IO and NPO (Fig. 2a and Supplementary Fig. S4), but they were also observed at sites where DDAs were not detected, such as in the Pacific North Equatorial Current (TARA_136), which was unexpected.

Trichodesmium is favored in warm (>26 °C) oligotrophic waters with low wind stress and a stable mixed layer (100 m or more) [7]. Tens to hundreds of *Trichodesmium* filaments often aggregate into fusiform-shaped colonies usually referred to as 'tufts' or 'rafts' or round-shaped colonies called 'puffs' (Fig. 1). The tremendous range in *Trichodesmium* colony diameter (from 200 μm to 5 mm) challenges our ability to collect and therefore consistently quantify/estimate their abundances. However, these dimensions were detectable and quantifiable by *in situ* imaging using the UVP5 (Fig. 2 and Supplementary Fig. S4a). Colonies were more prevalent in NAO and NPO, while IO stations were more enriched in free filaments (Fig. 2a and Supplementary Fig. S4a), probably related to the enhanced colony formation of *Trichodesmium* under nutrient limitation, as has been observed in culture experiments [43].

Single-cell free-living NCDs were estimated by combining flow cytometry estimates of free-living bacterial densities with diazotroph relative abundances derived from metagenomic sequencing of the 0.22-1.6/3 μm plankton size fractions (see Methods). We detected concentrations up to ~2.8x10^6 cells L^{-1}, with the highest values in the Pacific Ocean (Fig. 2a). Our estimates agree with recent reports based on the reconstruction of metagenome-assembled genomes [18].

The extensive imaging dataset from *Tara* Oceans also allowed us to convert abundance estimates into biovolumes. The comparison of individual abundance and biovolumes between the different diazotrophs from surface waters is shown in Fig. 2c. NCDs are by far the most abundant diazotrophs in the surface ocean, however DDAs and, in particular *Trichodesmium,* dominate in terms of biovolume (Fig. 2c). Cell density and biovolume of NCDs has not been reported previously at a global scale, so our work presented here expands our understanding about the relative contributions for these recently recognized important diazotrophs.

Diazotroph diversity and abundance using metagenomes from size fractionated plankton samples

To gain further insights into the abundance and distribution patterns of diazotrophs across the whole plankton size spectrum, we compared the imaging data with metagenomic reads from the 5 main size fractions mapped against a comprehensive catalog of 41,251 *nifH* sequences (see Methods). The *nifH* catalog represents most of the genetic diversity reported for diazotroph isolates and environmental clone libraries (although it has some redundancy, see Methods), with 30% of the sequences derived from marine environments and the rest from terrestrial and freshwater habitats. Less than 0.01% of these *nifH* sequences (406 out of 41,251) mapped with at least 80% similarity to the 1,192 metagenomes, retrieving a total of 87,810 mapped reads. Of the 406 sequences, 102 retrieved only one read. Mapped *nifH* reads were detected in slightly more than half of the samples (63% or 424 of 673), which highlights the broad distribution of diazotrophs in the *Tara* Oceans dataset (blue circles in Fig. 4a for surface waters; Supplementary Table S1).

We used the single-copy core gene *recA* to quantify the bacterial community in each sample; thus the read abundance ratio of *nifH/recA* provides an estimate for the relative contribution of diazotrophs (see Methods). Our analysis shows both a dramatic increase (up to 4 orders of magnitude) in diazotroph abundance and a dynamic compositional shift towards the larger size classes of plankton (Fig. 5). For example, diazotrophs comprise only a small proportion of the bacterial community in the 0.22-1.6/3 µm size fraction (minimum-maximum values of 0.004-0.8%), however, they increase to 0.003-40% in the 180-2000 µm size range (Fig. 5a). The increase is coincident with a change in taxonomy (Fig. 5b-c, Supplementary Table S1): proteobacteria and planctomycetes are the main components in the 0.22-1.6/3 µm size fraction (0.004-0.08% and 0.005-0.4%, respectively), while cyanobacterial diazotrophs dominate in the larger size fractions, including both filamentous (*Trichodesmium* and others) and non-filamentous types (free-living and symbiotic) (0.2-45% and 0.2-2%, respectively). When comparing the abundance patterns of these larger cyanobacterial diazotrophs based on our imaging methods with those based on metagenomic counts, the overlap was remarkable (Fig. 6, also compare panels a vs b in Supplementary Figs. S3 and S4). Hence, we developed a fully reversible pipeline from images to genomics and genomics to images to allow each to inform the other. The image analysis enables one to quickly identify which metagenomic sample(s) should contain a particular diazotroph. For populations like the cyanobacterial diazotrophs which are comparatively less abundant, this approach will reduce search time in genetic analyses.

The majority (95%) of the total recruited reads mapping to the *nifH* database corresponded to 20 taxonomic groups: 5 cyanobacteria, 2 planctomycetes, and 13 proteobacteria. For the NCDs, the 2 planctomycetes and 7 of the 13 proteobacterial types corresponded to recent metagenome-assembled genomes (named HBD01 to HBD09; [18]) which additionally were among the top contributors to the *nifH* transcript pool in the 0.22-1.6/3 µm size fraction of *Tara* Oceans metatranscriptomes [26]. We also found these taxa in the larger size fractions (Figs. 5c and 7). The 0.8 µm pore-size filter enriches for the larger bacterial cells, while letting pass the smaller ones (including more abundant taxa such as SAR11 and *Prochlorococcus*). However, it is interesting that we detected the NCDs in the three largest size fractions (5-20, 20-180 or 180-2000 µm), suggesting their attachment to particles (e.g. marine snow, facel pellets)[44] and/or larger eukaryotic cells/organisms, aggregation into colonies.

The main cyanobacterial taxa corresponded to *Trichodesmium*, *Richelia/Calothrix*, and UCYNs (UCYN-A1, UCYN-A2 and *Crocosphaera*). *Trichodesmium* represented the highest number of reads for *nifH* among all diazotrophs and constituted up to 40% of the bacterial community in the three largest size fractions (Figs. 5c and 7). Although *Trichodesmium* is widespread in the oceans, forming high density surface slicks and blooms, recent evidence for polyploidy has been shown in field and cultured *Trichodesmium* populations [45]. Hence polyploidy could influence the higher number of sequence reads for mapping, and therefore the higher numbers of *Trichodesmium* in our analysis.

Relative abundance of UCYN-A1 was highest in the smaller size fractions 0.2-1.6/3 µm and 0.8-5 µm, in accordance with the expected host size (1-3 µm; [13,46,47]), but was also detected in the larger size fractions (5-20, 20-180 and 180-2000 µm) (Figs. 5c and 7), probably related to particle association, which may subsequently sink to the deep ocean (see next section), or consumption by higher trophic levels [48].

Richelia displayed the highest relative abundance in the 20-180 µm size range, but was also detected in both the 5-20 and 180-2000 µm size fractions (Figs. 5c and 7). *Richelia* is associated with both small and large diatoms (*Hemiaulus and Rhizosolenia*, respectively; Figs. 1 and 3), and occasionally has been reported as free filaments [36–38]. Free filaments were also observed in our confocal analyses (Supplementary Fig. S5). Similar to *Richelia*, *Crocosphaera* was also found in multiple size fractions (0.8-5 µm, 5-20 µm, 20-180 µm, and 180-2000 µm), which is expected given its diverse life histories: free-living, colonial or symbionts of large *Climacodium* diatoms (Fig. 1) [14–16]. Other cyanobacterial symbionts of diatoms were also observed albeit in lower abundance, such as *Calothrix*, found in the 20-180 µm size range (Figs. 5c and 7) due to its association with chains of *Chaetoceros* (Fig. 1 and 3).

Unexpectedly, we recruited reads with sequence similarity to *nifH* from 'spheroid bodies' (Supplementary Fig. S7a), which are cyanobacteria that have lost photosynthesis [49] and heretofore have only been reported as N_2-fixing endosymbionts in a few freshwater rhopalodiacean diatoms [50–52]. To our knowledge, this is the first report of these populations in marine waters. Detection levels were however low (~0.5% of total bacterioplankton community) and mainly derived from the 20-180 µm size fraction (Figs. 5c and 7), which is consistent with the expected diatom host cell diameters (approximately 30-40 µm [53]). These spheroid-body like reads were detected in surface waters from the IO, SPO and SAO (Supplementary Fig. S7b). In these regions we also detected some images of pennate diatoms containing round granules without chlorophyll autofluorescence (Supplementary Fig. S7c), but further research will be required to validate if these are in fact diatoms with diazotrophic symbionts.

Insights into environmental distribution and depth partitioning of diazotrophs

Diazotroph abundance displayed a latitudinal gradient, showing as expected higher relative abundances in tropical and subtropical regions, and a decrease at the equator where upwelling and higher dissolved nutrients are expected (Fig. 4). This pattern is congruent with decades of field observations (e.g., NAO, NPO) as well as modeling efforts [23,54,55]. Correlation analyses with environmental and physico-chemical variables measured during the *Tara* Oceans cruise identified higher abundances in oligotrophic waters (regions of low nitrate and phosphate concentrations) with sea surface temperatures >20 °C (and especially >25 °C), but with variable modeled dissolved iron concentrations in the range between 0.005 and 2 nM (Fig. 9a). Temperature and nutrient availability are common factors which govern diazotroph abundances [8,23,56]. Iron should also be important for diazotrophs due to the high iron requirement for the nitrogenase enzyme [57,58], therefore it was unexpected to find a less robust relationship between diazotroph abundances and modeled dissolved iron concentrations (Fig. 9a).

We further analysed abundance and distribution patterns within the epipelagic and mesopelagic layers (0-200 m and 200-1000 m, respectively). The higher numbers of N_2-fixing cyanobacteria detected in the surface (5 m) compared to the DCM layer (17-188 m) in both the metagenomic and imaging datasets confirms expected distributions (Figs. 2b and 9c, also compare Fig. 8 and Supplementary Fig. S8). However, detection of both *Trichodesmium* and the DDA symbionts were nonetheless significant in some DCM samples from diverse regions: IO, SPO, and RS (Supplementary Fig. S8). *Richelia* is expected at depth given its reported rapid sinking, and observations in moored sediment traps (station ALOHA: [39,59]), while

Trichodesmium is considered to have a poor export capacity [60] and thus is not expected at depth. Increased abundances of C*rocosphaera* co-occurred in the DCM of IO samples, which were additionally associated with the 5-20 μm size fraction. We interpret these latter results as being indicative of the colonial and/or symbiotic life-histories previously reported for *Crocosphaera* (Fig. 1; [14,61]). Unlike the phototrophic diazotrophs, the distribution of NCDs had no apparent depth partitioning in the epipelagic layer (Fig. 9c).

A relatively high number of *nifH* reads were detected in the mesopelagic (128 out of 158 - or 81% - of mesopelagic samples, Supplementary Fig. S9). Although BNF and *nifH* expression has been previously reported at depths, most measurements have been made in oxygen minimum zones (OMZs; where low-oxygen waters are found) and oxygen-deficient zones (ODZs; where oxygen concentrations are low enough to induce anaerobic metabolisms) [19,26], while here we mapped *nifH* sequences from many samples outside of OMZs and ODZs. For example, the highest diazotroph enrichment in the mesopelagic bacterioplankton was in SPO, NPO, NAO and SAO (Supplementary Fig. S9). Although the majority of *nifH* sequences correspond to proteobacteria, sequences from diazotrophic cyanobacteria were also detected in the mesopelagic (Supplementary Fig. S9). In particular, 44% of total *nifH* reads in mesopelagic samples at TARA_78 and 6% at TARA_76 (of 0.2-3 μm size fraction) in SAO correspond exclusively to UCYN-A (Supplementary Fig. S9, Supplementary Table S1). In the surface samples of these stations we also detected high numbers of UCYN-A reads (Fig. 8; see below), suggesting a bloom at the surface. Most reports about UCYN-A have focused on their presence and activities in the sunlit layers, with the exception of a study reporting UCYN-A *nifH* sequences in shallow water sediments of the north east Atlantic ocean (seafloor 38-76 m depth) [62]. Our observation of UCYN-A at 800 m depth in the open ocean suggests that this symbiosis could contribute to carbon export despite its small size.

Global ocean biogeography of diazotrophs

We detected several regions with high densities or "hotspots" of diazotrophs (Figs. 4 and 8). For example, the Mozambique Channel between Madagascar and the African continent, where diazotrophs constitute up to 30-40% of the bacterioplankton in the larger size fraction samples (TARA_50 to TARA_62; Figs. 4 and 8). Moreover, the confocal microscopy observations confirm higher densities of both *Trichodesmium* and symbiotic diazotrophs in this region (Fig. 2a). Another example is the SAO near South America (TARA_76, TARA_78 and TARA_80), where UCYN-A reached 3-4% of the bacterioplankton population in the 0.8-5 µm size fraction (Figs. 4 and 8). These zones from IO and SAO represent previously undersampled regions for diazotrophs (Supplementary Fig. S1), which also lacks quantitative rate measurements for N_2 fixation [22].

The highest abundance of free-living single-cell NCDs (0.2-3 µm size fraction) corresponds to ~0.5% of the bacterioplankton in the wake of the Marquesas archipelago in the equatorial PO (TARA_123; Fig. 8), where a surface planktonic bloom triggered by natural iron fertilization was recently reported [63]. Other high density areas corresponded to a few stations in the SPO (TARA_98 and TARA_99 in the surface and TARA_102 at DCM), where high abundances of proteobacteria and planctomycetes (4-33% and 8-9%, respectively) were found in larger size fractions

(Fig. 8 and Supplementary Fig. S8), which likely results from association of NCDs to sinking particles. Moreover, TARA_102 is located in the Peruvian upwelling area, a region previously reported for NCDs and/or BNF activity associated with the OMZ [64–66]. These results are congruent with recent reports from the subtropical Pacific of highly diverse NCDs, some associated with sinking particles [20,66–68]. We can therefore expand the distribution of potentially particle-associated NCDs to several other ocean basins (NAO, AO, IO). Our findings emphasize the dominance and persistence of NCDs in larger size fractions of both surface and DCM, which is novel and warrants further investigation.

Overall, many regions contain a low abundance of diazotrophs. For example, the percentages of diazotrophs in the AO, the Southern Ocean (SO), and the MS reached maximum values of only 0.4, 1, and 4%, respectively (Figs. 4c and 8). The highest diazotroph abundance in the AO corresponded to NCDs found in shallow waters (20-25 m depth) of the East Siberian Sea (TARA_191; Fig. 4c and 8), a biologically undersampled region. Hence, like most plankton, diazotrophs are also largely ‘patchy’.

Cyanobacterial diazotrophs are mainly found as assemblies of abundant groups

With the exception of a few stations in IO, RS, NPO and NAO where *Trichodesmium* was the main component of the mapped reads (Figs. 8 and 9b, Supplementary Fig. S10), there was a general and consistent trend of several cyanobacterial diazotrophs that co-occurred. This pattern of co-occurrence was present in several oceanic regions (Fig. 8 and Supplementary Fig. S10). For example, in the Red Sea (RS), diazotrophs were mainly found in the oligotrophic northern part (TARA_32) (Fig. 8), and consisted of the larger diameter diazotrophs (*Trichodesmium*, *Richelia*) co-occurring with the small unicells (UCYN-A, *Crocosphaera*) (Fig. 8 and Supplementary Fig. S10). In fact, this additionally represents the first detection of UCYN-A in the RS, while the other cyanobacterial diazotrophs have been reported previously, including surface blooms of *Trichodesmium* spp. [69–72]. Co-occurring small and larger diameter diazotrophs dominated several stations in the southern IO, including two open ocean stations (TARA_50, TARA_51) where *Crocosphaera* and UCYN-A1 dominated in the small size fraction and *Trichodesmium* was predominant in the three larger size fractions (5-20 µm, 20-180 µm and 180-200 µm) (Fig. 8 and Supplementary Fig. S10). A similar pattern was observed at station ALOHA in the subtropical Pacific Ocean (TARA_131), consistent with previous observations [73,74]. On the contrary, only small diameter diazotrophs co-dominated in the SAO; for example UCYN-A1 and UCYN-A2 were very abundant at stations TARA_76, TARA_78 and TARA_80 (Fig. 8 and Supplementary Fig. S10). The numerous observations of mixed diazotrophic assemblages of different life histories (colonial, free-living, symbiotic, particle associated) highlights the need to consider how these traits enable co-occurrence.

Ultrasmall diazotrophs consist of proteobacteria and are abundant in the Arctic Ocean

Ultrasmall prokaryotes are unusual due to their reduced cell volume (these cells can pass through 0.22-µm filters, a size usually expected to exclude most microorganisms), and thus they are thought to have reduced genomes and to lack the proteins needed to carry out more complex metabolic processes. However, there is recent evidence that they do indeed participate in complex metabolisms [75]. In order to see if they also contribute to marine nitrogen fixation, we carried out the analysis of 134 metagenomes of <0.22 μm size fractionated samples of different water layers.

A total of 29 *nifH* sequences in our database mapped with at least 80% similarity to these metagenomes, retrieving a total of 42,409 mapped reads, almost all of them with high identity to proteobacterial *nifH* sequences. Of the 29 sequences, 6 retrieved only one read. Mapped *nifH* reads were detected in slightly more than half of the samples (61% or 78 of 127), which highlights an unexpected broad distribution of ultrasmall diazotrophs (blue circles in Fig. 10a; Supplementary Table S1). Notably, when *nifH* reads were normalized by *recA* reads, we found that diazotrophs comprise up to 10% of the ultrasmall bacterioplankton, with the highest abundances detected in the Arctic Ocean, and in different water layers (Fig. 10a-b). This is remarkable considering that this is the ocean with the lowest diazotroph abundance in the other size fractions (Figs. 4c and 8, Supplementary Figs. S8 and S9).

The majority (86%) of the total recruited reads mapping to our *nifH* database corresponded to two sequences assembled from the <0.22 μm size-fractionated metagenomes: OM-RGC.v2.008173703 and OM-RGC.v2.008955342. The former has 99% identity to *nifH* from the epsilon-proteobacterium *Arcobacter nitrofigilis* DSM7299 [76] and only retrieved reads from surface and DCM (Fig. 10c). The second has close similarity to sequences from gamma-proteobacteria and it retrieved reads from different water layers (Fig. 10c). Both sequences also retrieved reads from other sizes fractions (Fig. 7 and Supplementary Fig. S11). In the case of *Arcobacter*, this is in agreement with the fact that the species of this genus are either symbionts or pathogens [76], although its highest abundance is observed in the <0.22 μm size fraction: it constitutes >9% of ultrasmall bacterioplankton in the DCM waters of station TARA_158 (Supplementary Fig. S11). In addition to these two abundant sequences detected in different size ranges, we found a proteobacterium sequence that exclusively retrieved reads from the <0.22 μm size fractionated samples: OM-RGC.v2.008817394 (Supplementary Fig. S11). All in all, these results may prompt a fundamental revisit of marine nitrogen fixation and the incorporation of ultrasmall diazotrophs in ocean nitrogen cycle models.

Conclusions

This is the first attempt to assess the diversity, abundance, and distribution of diazotrophs at a global ocean scale using paired image and (PCR-free) molecular analyses. Unlike earlier studies, our work included the full biological and ecological complexity of diazotrophs: i.e. unicellular, colonial, symbiotic, cyanobacteria and NCDs. Our work also enabled estimates of total diazotrophic biovolume in several layers of the global ocean; information that is directly relevant to predicting C and N sources/sinks. Diazotrophs were found to be globally distributed and present in all size fractions, even among ultrasmall bacterioplankton (<0.22 μm), which were especially abundant in the Arctic Ocean. Unexpectedly, we detected sequences similar to obligate symbionts of freshwater diatoms nearly exclusively in the larger size fraction (20-180 μm). We interpret these results as evidence for a new symbiosis, given that their expected cell diameter is less than 5 μm. We did not find strong evidence for widespread distributions for UCYNs, which was unexpected given the results from a decade of past observations (although we cannot discount the influence of seasonal sampling biases). On the contrary, the highest abundance of UCYN-A was restricted to an area of the SAO where we found UCYN-A at depth and in surface samples, suggesting its significant contribution to carbon export, in spite of the small expected size of these symbiotic cells. A major conclusion from our work is the identification of new hotspots for diazotrophs in previously undersampled regions of the global ocean, for example, in several locations of the IO.

Both the morphological and molecular data support the canonical view of *Trichodesmium* dominance, rather than more recent propositions that have emphasized the importance of UCYNs. The numerous observations of co-occurring diazotrophs suggests the need to consider another further paradigm shift, namely that the diverse life histories of diazotrophs (colonial, free-living, symbiotic, particle associated) could enable their co-occurrence in mixed assemblages and a collective contribution to the N budget. Overall, this work provides an updated composite of diazotroph biogeography in the global ocean, providing valuable information towards modeling in the context of global change and the substantial anthropogenic perturbations to the marine nitrogen cycle [77].

Methods

Tara Oceans sampling

Tara Oceans performed a worldwide sampling of plankton between 2009 and 2013 (Supplementary Fig. S1b). Three different water depths were sampled: surface (5 m depth), deep chlorophyll maximum (hereafter DCM; 17–188 m), and mesopelagic (200–1000 m) (Supplementary Fig. S2). The plankton were separated into discrete size fractions using a serial filtration system [25]. Given the inverse logarithmic relationship between plankton size and abundance [25,78], higher seawater volumes were filtered for the larger size fractions (10-10^5 L; see Table 1 and Fig. 5 in [25]).

Taking into account that diazotrophs are less abundant than sympatric populations and have a wide size variation (Fig. 1), a comprehensive perspective requires analyses over a broad spectrum, which to date has been lacking. Five major organismal size fractions were collected: picoplankton (0.2 to 1.6 µm or 0.2 to 3 µm; named here 0.2-1.6/3 µm size fraction), piconanoplankton (0.8 to 5 µm or 0.8 to 2000 µm; named here 0.8-5 µm size fraction), nanoplankton (5 to 20 µm or 3 to 20 µm; named here 5-20 µm size fraction), microplankton (20 to 180 µm), and mesoplankton (180 to 2000 µm) (Supplementary Fig. S2) [25]. In addition, ultrasmall plankton (<0.22 µm) was also collected (Supplementary Fig. S2) [25]. The *Tara* Oceans datasets used in the present work are listed in Supplementary Fig. S2 and specific details about them and their analysis are described below.

Read recruitment of marker genes in metagenomes

The use of metagenomes avoids the biases linked to the PCR amplification steps of metabarcoding methods, and thus it is better for quantitative observations. This is especially important for protein-coding gene markers, such as *nifH*, which display high variability in the third position of most codons, and thus necessitate the use of highly degenerate primers for a broad taxonomic coverage [79]. The detection of low-abundance organisms, such as diazotrophs, is facilitated by the deep sequencing of the *Tara* Oceans samples (between ~10^8 and ~10^9 total metagenomic reads per sample) [26,28,29]. The 1,326 metagenomes generated by the expedition are derived from 147 globally distributed stations and three different water layers: 745 metagenomes from surface, 382 from DCM (17–188 m) and 41 from the bottom of the mixed layer when no DCM was observed (25-140 m), and 158 from mesopelagic (200–1000 m) (Supplementary Fig. S2).

The metagenomes were aligned against sequence catalogs of marker genes for diazotrophs (*nifH*) and bacteria (*recA*). The analysis was carried out using bwa tool version 0.7.4 [80] using the following parameters: -minReadSize 70 -identity 80 -alignment 80 -complexityPercent 75 -complexityNumber 30. The *nifH* sequence catalog (hereafter *nifH* database) was composed of 41,229 publicly available sequences from the laboratory of JP Zehr (University of California, Santa Cruz, USA; version April 2014; https://www.jzehrlab.com) complemented with 21 additional *nifH* genes with less than 95% identity to those in the Zehr database retrieved from different *Tara* Oceans datasets: OM-Reference Gene Catalog version 2 (OM-RGC-v2, [26]), assemblies [18] and clones. Although the Zehr database has some redundancy (9,048 out of the 41,251 total sequences are retained when clustered at 95% identity using CDHIT-EST tool [81]), we decided to use the whole database to maximize the number of metagenomic mapping reads. The *recA* sequences were obtained from sequenced genomes in the Integrated Microbial Genome database (IMG) [82] and from OM-RGC-v2 [26]. Homologous sequences were included in the two catalogs as outgroups to minimize false positive read alignments. They were retrieved from IMG, OM-RGC-v2 and the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP [83]) using HMMer v3.2.1 with gathering threshold option (http://hmmer.org/). The outgroups for *recA* consisted of sequences coding for the RecA Pfam domain (PF00154) different from the canonical *recA* gene, which include those coding for RADA and RADB in archaea, RAD51 and DCM1 in eukaryotes, and UvsX in viruses [84]. Outgroups for *nifH* consisted of sequences coding for the Pfam domain Fer4_NifH (PF00142) different from *nifH*, including those coding for a subunit of the pigment biosynthesis complexes protochlorophyllide reductase and chlorophyllide reductase [85].

We used the read abundance of the single-copy gene *recA* to estimate the total bacterial community in each sample (in contrast to the widely used 16S rRNA gene, which varies between one and fifteen copies among bacterial genomes; [86,87]). For simplicity, we assumed that *nifH* is also a single-copy gene, so the abundance ratio of *nifH/recA* provides an estimate for the relative contribution of diazotrophs to the total bacterial community. However, we realize that there are examples of 2-3 *nifH* copies in heterocyst-forming cyanobacteria such as *Anabaena variabilis* and *Fischerella* sp. [88,89], or in the firmicutes *Clostridium pasteurianum* [90], and that we are not taking into account the polyploidy effect observed for example in *Trichodesmium* spp. [45] and *Anabaena* spp. [91,92].

Phylogenetic analysis of recruited metagenomic reads

To support the taxonomic affiliation of metagenomic reads recruited by *nifH* sequences from 'spheroid bodies', we carried out a phylogenetic reconstruction in the following way. The translated metagenomic reads were aligned against a NifH reference alignment using the option --add of MAFFT version 6 with the G-INS-I strategy [93]. The resulting protein alignment was used for aligning the corresponding nucleotide sequences using TranslatorX [94] and phylogenetic trees were generated using the HKY85 substitution model in PhyML version 3.0 [95]. Four categories of rate variation were used. The starting tree was a BIONJ tree and the type of tree improvement was subtree pruning and regrafting. Branch support was calculated using the approximate likelihood ratio test (aLRT) with a Shimodaira–Hasegawa-like (SH-like) procedure.

Flow cytometry data and analysis

Picoplankton samples were prepared for flow cytometry from three aliquots of 1 ml of seawater (pre-filtered through 200-µm mesh), as described in [27,96]. For quantifying the densities of single-cell free-living diazotrophs, we combined the cell density measurements from flow cytometry with the relative abundances derived from molecular methods as done by Props et al. [97]. Specifically, we multiplied the bacterial concentration derived from flow cytometry by the *nifH* to *recA* ratio of metagenomic read abundances from samples of size fraction 0.22-1.6 µm or 0.22-3 µm. For biovolume estimations of single-cell free-living diazotrophs, we assumed an arbitrary average cell biovolume of 1 µm^3.

Detection of diazotrophs in the confocal laser-scanning microscopy dataset

Quantitative microscopy was performed using eHCFM [33] on 61 samples collected using a microplankton net (20-180 µm mesh size) at 48 different stations (Supplementary Fig. S2). Sample collection and preparation as well imaging acquisition is described in [33]. Briefly, samples were fixed on board *Tara* in 10% monomeric formaldehyde (1 % final concentration) buffered at pH 7.5 and 500 µl EM grade glutaraldehyde (0.25% final concentration) and kept at 4 °C until analysis. Cells were imaged by Confocal Laser Scanning Microscopy (Leica Microsystem SP8, Leica Germany), equipped with an automated high-content imaging platform and several laser lines (405 nm, 488 nm, 552 nm, 638 nm). Automated images using the HCS A module of LSAF software (Leica Microsystem) and the water immersion lens HC PL APO 40x/1,10 mot CORR CS2 objective were scanned bidirectionally at 600 Hz. Multiple fluorescent dyes were used to observe the cellular components of the organisms, including: the nuclei (blue, Hoechst, Ex405/Em420-470), cellular membranes (green, DiOC6(3), Ex488/Em500-520), cell surface (cyan, AlexaFluor 546, Ex552/Em560-590), and chlorophyll autofluorescence (red, Ex638/Em680-700).

We used the confocal microscopy data to quantify only the DDAs and *Trichodesmium* free filaments in terms of abundances and biovolume. Image detection and annotation was carried out using the Ecotaxa web platform [98] in the following way. We first manually searched for the target taxa and curated an initial training set in a few samples where molecular methods detected high abundances (i.e., high metagenomic read abundance of *nifH*), obtaining 53 images for DDAs and 80 for *Trichodesmium* filaments. This training set was then used for machine learning automated recognition (random forest) based on a collection of 480 numeric 2D/3D features [33]. The predictions were, in turn, manually curated and used as a new training set, repeating this step numerous times until no new images appeared. Other taxonomic groups were also annotated and used as outgroups to improve the predictions of our taxa of interest. Abundance estimates were normalized based on the total sample volumes as cells L^{-1}. We used the major and minor axis of every image to calculate their ellipsoidal equivalent biovolume.

Underwater Vision Profiler dataset and analysis

The Underwater Vision Profiler 5 (UVP5, Hydroptics, France) [35] is an underwater imager mounted on the Rosette Vertical Sampling System. This system allows to illuminate precisely calibrated volumes of water and capture images at a rate of 5 to 20 images s^{-1} during the descent. The UVP5 was operated *in situ* and was designed to detect and count objects of >100 µm in length and to identify those of >600 µm in size. In the current work, we used this method for the quantification of

Trichodesmium colony abundance and biovolume. The search, curation and annotation of the corresponding images and their biovolume determination were carried out as described in the previous section.

Determination of contextual physicochemical parameters

Measurements of temperature were recorded at the time of sampling using the vertical profile sampling system (CTD-rosette) and Niskin bottles following the sampling package described in [99,100] . Phosphate concentrations were determined using segmented flow analysis [101]. Nitrate concentrations were measured using a SATLANTIC ISUS nitrate sensor [100]. Iron levels were derived from a global ocean biogeochemical model [102].

Plotting and statistical analysis

All analyses were carried out in R language (http://www.r-project.org/). Graphs were plotted with R library *ggplot2* [103] and treemaps were generated with R library *treemap*. The trends between diazotroph abundance and latitude were displayed with generalized additive models using the *geom_smooth* function of *ggplot2* [103]. Metric multidimensional scaling (NMDS) analysis to visualize Bray-Curtis distances was carried out with the *metaMDS* command in the R package *vegan* [104], and the influence of environmental variables on sample ordination was evaluated with the function *envfit* in the same R package. Hierarchical agglomerative clustering of samples using average linkage was performed with the function *hclust* of the R package *stats*.

Data availability

Contextual data [25]: https://doi.org/10.1594/PANGAEA.875582; flow cytometry [27,96]: http://dx.doi.org/10.17632/p9r9wttjkm.1; high throughput confocal microscopy images [33] of 20-180 µm sized-fractionated samples: https://ecotaxa.obs-vlfr.fr/prj/2274; UVP5 images [35]: https://ecotaxa.obs-vlfr.fr/prj/579. *Tara* Oceans metagenomes [26,28,29] are archived at ENA under the accession numbers: PRJEB1787, PRJEB1788, PRJEB4352, PRJEB4419, PRJEB9691, PRJEB9740 and PRJEB9742.

Acknowledgements.

We would like to thank all colleagues from the *Tara* Oceans consortium as well as the *Tara* Ocean Foundation for their inspirational vision. We also acknowledge the laboratory of JP Zehr for public access to the *nifH* database, Daniel Lundin from Linnaeus University for his analytical support and advice, Tom Delmont from Genoscope for his useful discussions, Flora Vincent from Weizmann Institute of Science for her help with Ecotaxa resource, Eric Bapteste, Philippe Lopez and Romain Lannes from Sorbonne Université for their advice with ultrasmall bacteria analysis. This work has been supported by the FFEM - French Facility for Global Environment, French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), and PSL Research University (ANR-11-IDEX-0001-02). R.A.F. acknowledges funding from Knut and Alice Wallenberg foundation. C.B. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Diatomic; grant agreement No. 835067) and Agence Nationale de la Recherche "Phytomet"

(ANR-16-CE01-0008) projects. J.J.P.K. acknowledges postdoctoral funding from the Fonds Français pour l'Environnement Mondial. This article is contribution number ** of *Tara* Oceans.

Author contributions

RAF and CB designed the study and supervised the project. RAF, CB and JJPK wrote the paper with substantial input from CdV, FL, PW, EP and MP. EP performed the metagenomic mapping. RP and EK set up the imaging platform for the e-HFCM data generation and processing. JJPK, MC, ED, FL, SC performed the taxonomic annotation of the e-HFCM dataset of 20-180 µm size-fractionated samples. MP performed the collection and taxonomic annotation of UVP5 dataset. JJPK performed the formal analysis and visualization.

Competing financial interests: The authors declare no competing financial interests.

References

1. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
2. Moore, C. M. *et al.* Processes and patterns of oceanic nutrient limitation. *Nature Geoscience* vol. 6 701–710 (2013).
3. Tyrrell, T. The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature* vol. 400 525–531 (1999).
4. Falkowski, P. G. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature* vol. 387 272–275 (1997).
5. Karl, D. *et al.* The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature* vol. 388 533–538 (1997).
6. Capone, D. G. *et al.* Nitrogen fixation byTrichodesmiumspp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochemical Cycles* vol. 19 (2005).
7. Capone, D. G. Trichodesmium, a Globally Significant Marine Cyanobacterium. *Science* **276**, 1221–1229 (1997).
8. Moisander, P. H. *et al.* Unicellular Cyanobacterial Distributions Broaden the Oceanic N2 Fixation Domain. *Science* vol. 327 1512–1514 (2010).
9. Zehr, J. P., Shilova, I. N., Farnelid, H. M., Muñoz-Marín, M. D. C. & Turk-Kubo, K. A. Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. *Nat Microbiol* **2**, 16214 (2016).
10. Hagino, K., Onuma, R., Kawachi, M. & Horiguchi, T. Discovery of an endosymbiotic nitrogen-fixing cyanobacterium UCYN-A in Braarudosphaera

bigelowii (Prymnesiophyceae). *PLoS One* **8**, e81749 (2013).

11. Thompson, A. W. *et al.* Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science* vol. 337 1546–1550 (2012).
12. Farnelid, H., Turk-Kubo, K., Muñoz-Marín, M. C. & Zehr, J. P. New insights into the ecology of the globally significant uncultured nitrogen-fixing symbiont UCYN-A. *Aquatic Microbial Ecology* vol. 77 125–138 (2016).
13. Cornejo-Castillo, F. M. *et al.* UCYN-A3, a newly characterized open ocean sublineage of the symbiotic N2 -fixing cyanobacterium Candidatus Atelocyanobacterium thalassa. *Environmental Microbiology* vol. 21 111–124 (2019).
14. Carpenter, E. J. & Janson, S. INTRACELLULAR CYANOBACTERIAL SYMBIONTS IN THE MARINE DIATOM CLIMACODIUM FRAUENFELDIANUM (BACILLARIOPHYCEAE). *J. Phycol.* **36**, 540–544 (2000).
15. Webb, E. A., Ehrenreich, I. M., Brown, S. L., Valois, F. W. & Waterbury, J. B. Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, Crocosphaera watsonii, isolated from the open ocean. *Environ. Microbiol.* **11**, 338–348 (2009).
16. Caputo, A., Nylander, J. A. A. & Foster, R. A. The genetic diversity and evolution of diatom-diazotroph associations highlights traits favoring symbiont integration. *FEMS Microbiol. Lett.* **366**, (2019).
17. Foster, R. A. *et al.* Influence of the Amazon River plume on distributions of free-living and symbiotic cyanobacteria in the western tropical north Atlantic Ocean. *Limnology and Oceanography* vol. 52 517–532 (2007).
18. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol* **3**,

804–813 (2018).

19. Moisander, P. H. *et al.* Chasing after Non-cyanobacterial Nitrogen Fixation in Marine Pelagic Environments. *Front. Microbiol.* **8**, 1736 (2017).

20. Farnelid, H. *et al.* Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* **6**, e19223 (2011).

21. Halm, H. *et al.* Heterotrophic organisms dominate nitrogen fixation in the South Pacific Gyre. *The ISME Journal* vol. 6 1238–1249 (2012).

22. Luo, Y.-W. *et al.* Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. (2012) doi:10.5194/essd-4-47-2012.

23. Tang, W. & Cassar, N. Data-Driven Modeling of the Distribution of Diazotrophs in the Global Ocean. *Geophysical Research Letters* vol. 46 12258–12269 (2019).

24. Bork, P. *et al.* Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science* **348**, 873 (2015).

25. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**, 150023 (2015).

26. Salazar, G. *et al.* Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell* **179**, 1068–1083.e21 (2019).

27. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).

28. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* **4**, 170093 (2017).

29. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**,

373 (2018).

30. Cornejo-Castillo, F. M. *et al.* Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat. Commun.* **7**, 11071 (2016).

31. Cornejo-Castillo, F. M. & Zehr, J. P. Intriguing size distribution of the uncultured and globally widespread marine non-cyanobacterial diazotroph Gamma-A. *The ISME Journal* (2020) doi:10.1038/s41396-020-00765-1.

32. Vorobev, A. *et al.* Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Res.* (2020) doi:10.1101/gr.253070.119.

33. Colin, S. *et al.* Quantitative 3D-imaging for cell biology and ecology of environmental microbial eukaryotes. *Elife* **6**, (2017).

34. Cabello, A. M. *et al.* Global distribution and vertical patterns of a prymnesiophyte-cyanobacteria obligate symbiosis. *ISME J.* **10**, 693–706 (2016).

35. Picheral, M. *et al.* The Underwater Vision Profiler 5: An advanced instrument for high spatial resolution studies of particle size spectra and zooplankton. *Limnology and Oceanography: Methods* vol. 8 462–473 (2010).

36. Gómez, F., Furuya, K. & Takeda, S. Distribution of the cyanobacterium Richelia intracellularis as an epiphyte of the diatom Chaetoceros compressus in the western Pacific Ocean. *Journal of Plankton Research* vol. 27 323–330 (2005).

37. Villareal, T. A. Laboratory Culture and Preliminary Characterization of the Nitrogen-Fixing Rhizosolenia-Richelia Symbiosis. *Marine Ecology* vol. 11 117–132 (1990).

38. Foster, R. A., Goebel, N. L. & Zehr, J. P. ISOLATION OF CALOTHRIX RHIZOSOLENIAE (CYANOBACTERIA) STRAIN SC01 FROM CHAETOCEROS

(BACILLARIOPHYTA) SPP. DIATOMS OF THE SUBTROPICAL NORTH PACIFIC OCEAN1. *Journal of Phycology* vol. 46 1028–1037 (2010).

39. Karl, D. M., Church, M. J., Dore, J. E., Letelier, R. M. & Mahaffey, C. Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proceedings of the National Academy of Sciences* vol. 109 1842–1849 (2012).
40. Scharek, R., Tupas, L. M. & Karl, D. M. Diatom fluxes to the deep sea in the oligotrophic North Pacific gyre at Station ALOHA. *Mar. Ecol. Prog. Ser.* **182**, 55–67 (1999).
41. Scharek, R., Latasa, M., Karl, D. M. & Bidigare, R. R. Temporal variations in diatom abundance and downward vertical flux in the oligotrophic North Pacific gyre. *Deep Sea Research Part I: Oceanographic Research Papers* vol. 46 1051–1075 (1999).
42. Ratten, J.-M. *et al.* Sources of iron and phosphate affect the distribution of diazotrophs in the North Atlantic. *Deep Sea Research Part II: Topical Studies in Oceanography* vol. 116 332–341 (2015).
43. Tzubari, Y., Magnezi, L., Be'er, A. & Berman-Frank, I. Iron and phosphorus deprivation induce sociality in the marine bloom-forming cyanobacterium Trichodesmium. *ISME J.* **12**, 1682–1693 (2018).
44. Geisler, E., Bogler, A., Rahav, E. & Bar-Zeev, E. Direct Detection of Heterotrophic Diazotrophs Associated with Planktonic Aggregates. *Scientific Reports* vol. 9 (2019).
45. Sargent, E. C. *et al.* Evidence for polyploidy in the globally important diazotrophTrichodesmium. *FEMS Microbiology Letters* vol. 363 fnw244 (2016).
46. Foster, R. A. & Zehr, J. P. Diversity, Genomics, and Distribution of

Phytoplankton-Cyanobacterium Single-Cell Symbiotic Associations. *Annu. Rev. Microbiol.* **73**, 435–456 (2019).

47. Tripp, H. J. *et al.* Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* vol. 464 90–94 (2010).

48. Scavotto, R. E., Dziallas, C., Bentzon-Tilia, M., Riemann, L. & Moisander, P. H. Nitrogen-fixing bacteria associated with copepods in coastal waters of the North Atlantic Ocean. *Environ. Microbiol.* **17**, 3754–3765 (2015).

49. Nakayama, T. *et al.* Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. *Proceedings of the National Academy of Sciences* vol. 111 11407–11412 (2014).

50. Drum, R. W. & Pankratz, S. Fine structure of an unusual cytoplasmic inclusion in the diatom genus,Rhopalodia. *Protoplasma* vol. 60 141–149 (1965).

51. Nakayama, T. & Inagaki, Y. Genomic divergence within non-photosynthetic cyanobacterial endosymbionts in rhopalodiacean diatoms. *Sci. Rep.* **7**, 13075 (2017).

52. Prechtl, J., Kneip, C., Lockhart, P., Wenderoth, K. & Maier, U.-G. Intracellular Spheroid Bodies of Rhopalodia gibba Have Nitrogen-Fixing Apparatus of Cyanobacterial Origin. *Molecular Biology and Evolution* vol. 21 1477–1481 (2004).

53. Patrick, R. & Reimer, C. W. *The Diatoms of the United States: Exclusive of Alaska and Hawaii. Volume 2, Part 1, Entomoneidaceae, Cymbellaceae, Gomphonemaceae, Epithemiaceae*. (1975).

54. Dutkiewicz, S., Ward, B. A., Monteiro, F. & Follows, M. J. Interconnection of nitrogen fixers and iron in the Pacific Ocean: Theory and numerical simulations. *Global Biogeochemical Cycles* vol. 26 (2012).

55. Deutsch, C., Sarmiento, J. L., Sigman, D. M., Gruber, N. & Dunne, J. P. Spatial coupling of nitrogen inputs and losses in the ocean. *Nature* vol. 445 163–167 (2007).

56. Mills, M. M., Ridame, C., Davey, M., La Roche, J. & Geider, R. J. Iron and phosphorus co-limit nitrogen fixation in the eastern tropical North Atlantic. *Nature* vol. 429 292–294 (2004).

57. Raven, J. A. The iron and molybdenum use efficiencies of plant growth with different energy, carbon and nitrogen sources. *New Phytologist* vol. 109 279–287 (1988).

58. Kustka, A. B. *et al.* Iron requirements for dinitrogen- and ammonium-supported growth in cultures ofTrichodesmium(IMS 101): Comparison with nitrogen fixation rates and iron: carbon ratios of field populations. *Limnology and Oceanography* vol. 48 1869–1884 (2003).

59. Subramaniam, A. *et al.* Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10460–10465 (2008).

60. Bar-Zeev, E., Avishay, I., Bidle, K. D. & Berman-Frank, I. Programmed cell death in the marine cyanobacterium Trichodesmium mediates carbon and nitrogen export. *ISME J.* **7**, 2340–2348 (2013).

61. Foster, R. A., Sztejrenszus, S. & Kuypers, M. M. M. Measuring carbon and N2fixation in field populations of colonial and free-living unicellular cyanobacteria using nanometer-scale secondary ion mass spectrometry1. *Journal of Phycology* vol. 49 502–516 (2013).

62. Brown, S. M. & Jenkins, B. D. Profiling gene expression to distinguish the likely active diazotrophs from a sea of genetic potential in marine sediments.

Environmental Microbiology vol. 16 3128–3142 (2014).

63. Caputi, L. *et al.* Community-Level Responses to Iron Availability in Open Ocean Plankton Ecosystems. *Global Biogeochem. Cycles* **33**, 391–419 (2019).
64. Loescher, C. R. *et al.* Facets of diazotrophy in the oxygen minimum zone waters off Peru. *The ISME Journal* vol. 8 2180–2192 (2014).
65. Fernandez, C., Farías, L. & Ulloa, O. Nitrogen fixation in denitrified marine waters. *PLoS One* **6**, e20539 (2011).
66. Jayakumar, A., Al-Rshaidat, M. M. D., Ward, B. B. & Mulholland, M. R. Diversity, distribution, and expression of diazotroph nifH genes in oxygen-deficient waters of the Arabian Sea. *FEMS Microbiol. Ecol.* **82**, 597–606 (2012).
67. Turk-Kubo, K. A., Karamchandani, M., Capone, D. G. & Zehr, J. P. The paradox of marine heterotrophic nitrogen fixation: abundances of heterotrophic diazotrophs do not account for nitrogen fixation rates in the Eastern Tropical South Pacific. *Environ. Microbiol.* **16**, 3095–3114 (2014).
68. Jayakumar, A. *et al.* Biological nitrogen fixation in the oxygen-minimum region of the eastern tropical North Pacific ocean. *ISME J.* **11**, 2356–2367 (2017).
69. Kimor, B. & Golandsky, B. Microplankton of the Gulf of Elat: Aspects of seasonal and bathymetric distribution. *Marine Biology* vol. 42 55–67 (1977).
70. Gordon, N., Angel, D. L., Neorl, A., Kress, N. & Kimor, B. Heterotrophic dinoflagellates with symbiotic cyanobacteria and nitrogen limitation in the Gulf of Aqaba. *Marine Ecology Progress Series* vol. 107 83–88 (1994).
71. Post, A. F. *et al.* Spatial and temporal distribution of Trichodesmium spp. in the stratified Gulf of Aqaba, Red Sea. *Marine Ecology Progress Series* vol. 239 241–250 (2002).
72. Foster, R. A., Paytan, A. & Zehr, J. P. Seasonality of N2 fixation andnifHgene

diversity in the Gulf of Aqaba (Red Sea). *Limnology and Oceanography* vol. 54 219–233 (2009).

73. Church, M. J., Short, C. M., Jenkins, B. D., Karl, D. M. & Zehr, J. P. Temporal patterns of nitrogenase gene (nifH) expression in the oligotrophic North Pacific Ocean. *Appl. Environ. Microbiol.* **71**, 5362–5370 (2005).
74. Zehr, J. P. & Capone, D. G. Changing perspectives in marine nitrogen fixation. *Science* **368**, (2020).
75. Lannes, R., Olsson-Francis, K., Lopez, P. & Bapteste, E. Carbon Fixation by Marine Ultrasmall Prokaryotes. *Genome Biol. Evol.* **11**, 1166–1177 (2019).
76. Pati, A. *et al.* Complete genome sequence of Arcobacter nitrofigilis type strain (CI). *Stand. Genomic Sci.* **2**, 300–308 (2010).
77. Kim, I.-N. *et al.* Chemical oceanography. Increasing anthropogenic nitrogen in the North Pacific Ocean. *Science* **346**, 1102–1106 (2014).
78. Belgrano, A., Allen, A. P., Enquist, B. J. & Gillooly, J. F. Allometric scaling of maximum population density: a common rule for marine phytoplankton and terrestrial plants. *Ecology Letters* vol. 5 611–613 (2002).
79. Gaby, J. C. & Buckley, D. H. The Use of Degenerate Primers in qPCR Analysis of Functional Genes Can Cause Dramatic Quantification Bias as Revealed by Investigation of nifH Primer Performance. *Microbial Ecology* vol. 74 701–708 (2017).
80. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* vol. 25 1754–1760 (2009).
81. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
82. Chen, I.-M. A. *et al.* IMG/M v.5.0: an integrated data management and

comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res.* **47**, D666–D677 (2019).

83. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
84. Lin, Z., Kong, H., Nei, M. & Ma, H. Origins and evolution of the recA/RAD51 gene family: evidence for ancient gene duplication and endosymbiotic gene transfer. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 10328–10333 (2006).
85. Fujita, Y. & Bauer, C. E. Reconstitution of light-independent protochlorophyllide reductase from purified bchl and BchN-BchB subunits. In vitro confirmation of nitrogenase-like features of a bacteriochlorophyll biosynthesis enzyme. *J. Biol. Chem.* **275**, 23583–23588 (2000).
86. Kembel, S. W., Wu, M., Eisen, J. A. & Green, J. L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* **8**, e1002743 (2012).
87. Angly, F. E. *et al.* CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome* **2**, 11 (2014).
88. Zehr, J. P., Mellon, M. T. & Hiorns, W. D. Phylogeny of cyanobacterial nifH genes: evolutionary implications and potential applications to natural assemblages. *Microbiology* **143 (Pt 4)**, 1443–1450 (1997).
89. Thiel, T. & Pratte, B. S. Regulation of Three Nitrogenase Gene Clusters in the Cyanobacterium Anabaena variabilis ATCC 29413. *Life* **4**, 944–967 (2014).
90. Langlois, R. J., Hümmer, D. & LaRoche, J. Abundances and distributions of the dominant nifH phylotypes in the Northern Atlantic Ocean. *Appl. Environ.*

Microbiol. **74**, 1922–1931 (2008).

91. Simon, R. D. Macromolecular Composition of Spores from the Filamentous Cyanobacterium Anabaena cylindrica. *Journal of Bacteriology* vol. 129 1154–1155 (1977).
92. Simon, R. D. DNA content of heterocysts and spores of the filamentous cyanobacteriumAnabaena variabilis. *FEMS Microbiology Letters* vol. 8 241–245 (1980).
93. Katoh, K. & Toh, H. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. *BMC Bioinformatics* **9**, 212 (2008).
94. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* **38**, W7–13 (2010).
95. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
96. Hingamp, P. *et al.* Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**, 1678–1695 (2013).
97. Props, R. *et al.* Absolute quantification of microbial taxon abundances. *ISME J.* **11**, 584–587 (2017).
98. Picheral, M., Colin, S. & Irisson, J.-O. EcoTaxa, a tool for the taxonomic classification of images. http://ecotaxa.obs-vlfr.fr (2017).
99. Website. Picheral, Marc; Searson, Sarah; Taillandier, Vincent; Bricaud, Annick; Boss, Emmanuel; Ras, Josephine; Claustre, Hervé; Ouhssain, Mustapha; Morin, Pascal; Tremblay, Jean-Éric; Coppola, Laurent; Gattuso, Jean-Pierre; Metzl,

Nicolas; Thuillier, Doris; Gorsky, Gabriel; Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants (2014): Vertical profiles of environmental parameters measured on discrete water samples collected with Niskin bottles during the Tara Oceans expedition 2009-2013. PANGAEA, https://doi.org/10.1594/PANGAEA.836319.

100. Website. Picheral, Marc; Searson, Sarah; Taillandier, Vincent; Bricaud, Annick; Boss, Emmanuel; Stemmann, Lars; Gorsky, Gabriel; Tara Oceans Consortium, Coordinators; Tara Oceans Expedition, Participants (2014): Vertical profiles of environmental parameters measured from physical, optical and imaging sensors during Tara Oceans expedition 2009-2013. PANGAEA, https://doi.org/10.1594/PANGAEA.836321.

101. Practical Guidelines for the Analysis of Seawater. (2009) doi:10.1201/9781420073072.

102. Aumont, O., Ethé, C., Tagliabue, A., Bopp, L. & Gehlen, M. PISCES-v2: an ocean biogeochemical model for carbon and ecosystem studies. *Geoscientific Model Development* vol. 8 2465–2513 (2015).

103. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer Science & Business Media, 2009).

104. Oksanen, J. *et al.* vegan: Community Ecology Package. https://cran.r-project.org/web/packages/vegan/index.html (2019).

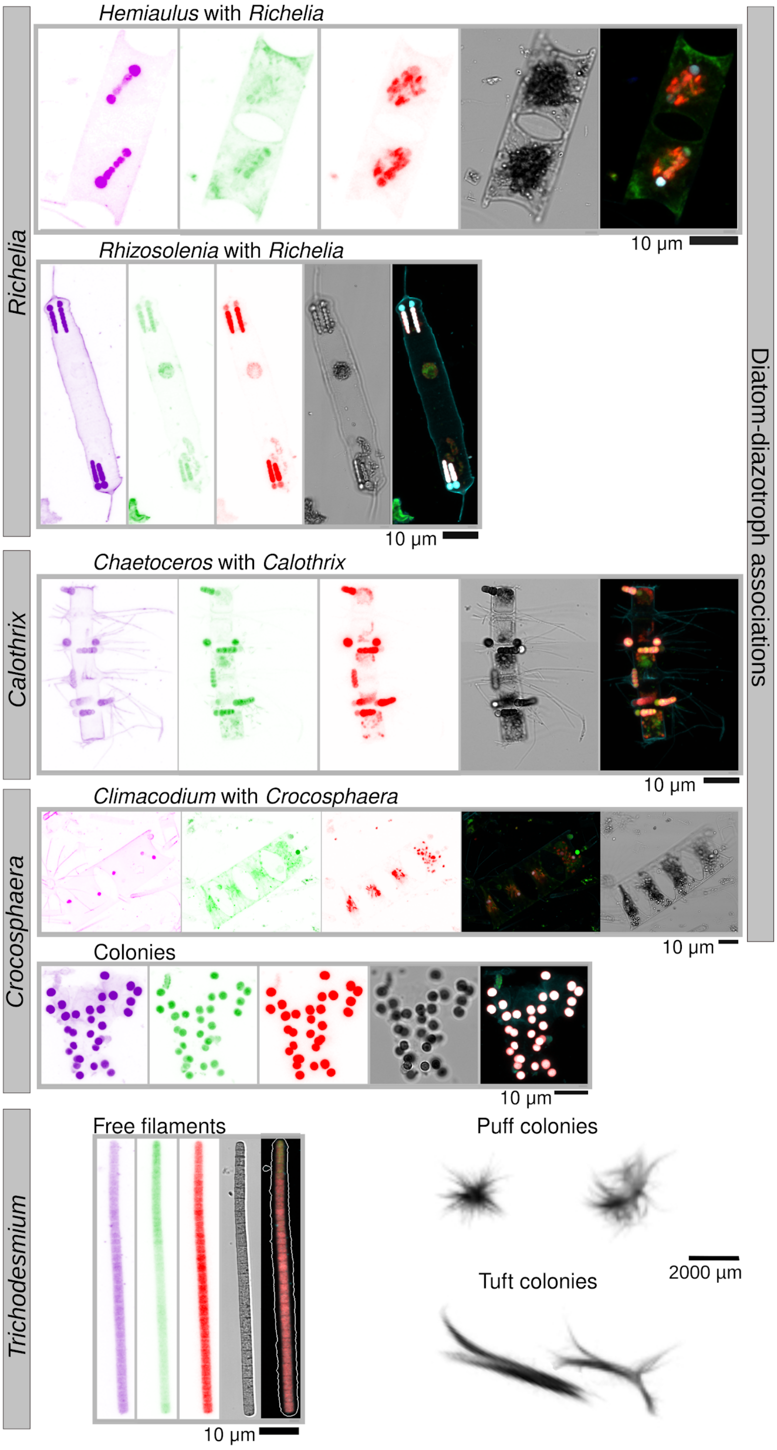

Hemiaulus with *Richelia*
10 µm
Rhizosolenia with *Richelia*
10 µm
Richelia
Chaetoceros with *Calothrix*
10 µm
Calothrix
Climacodium with *Crocosphaera*
10 µm
Colonies
10 µm
Crocosphaera
Diatom-diazotroph associations
Free filaments
10 µm
Puff colonies
Tuft colonies
2000 µm
Trichodesmium

Figure 1: Imaging observations of diazotrophs in *Tara* Oceans samples. Images were obtained by environmental High Content Fluorescence Microscopy (eHCFM; Colin et al., 2017), with the exception of *Trichodesmium* colonies , which were detected in situ using an Underwater Vision Profiler 5 (UVP5; Picheral et al., 2010). From left to right, the displayed channels for each micrograph correspond to cell surface (cyan, AlexaFluor 546), cellular membranes (green, DiOC6), chlorophyll autofluorescence (red), the bright field, and the merged channels. The displayed *Hemiaulus-Richelia* association was detected at station TARA_80 in the South Atlantic Ocean, *Rhizosolenia-Richelia* at TARA_53 in the Indian Ocean, *Chaetoceros-Calothrix* at TARA_131 (ALOHA) in the North Pacific Ocean, *Climacodium-Croscophaera* at TARA_140 in the North Pacific Ocean, the *Croscophaera*-like colony at TARA_53 in the Indian Ocean, the *Trichodesmium* filament at TARA_42 in the Indian Ocean, and the *Trichodesmium* colonies at TARA_141 and TARA_142 in the North Atlantic Ocean.

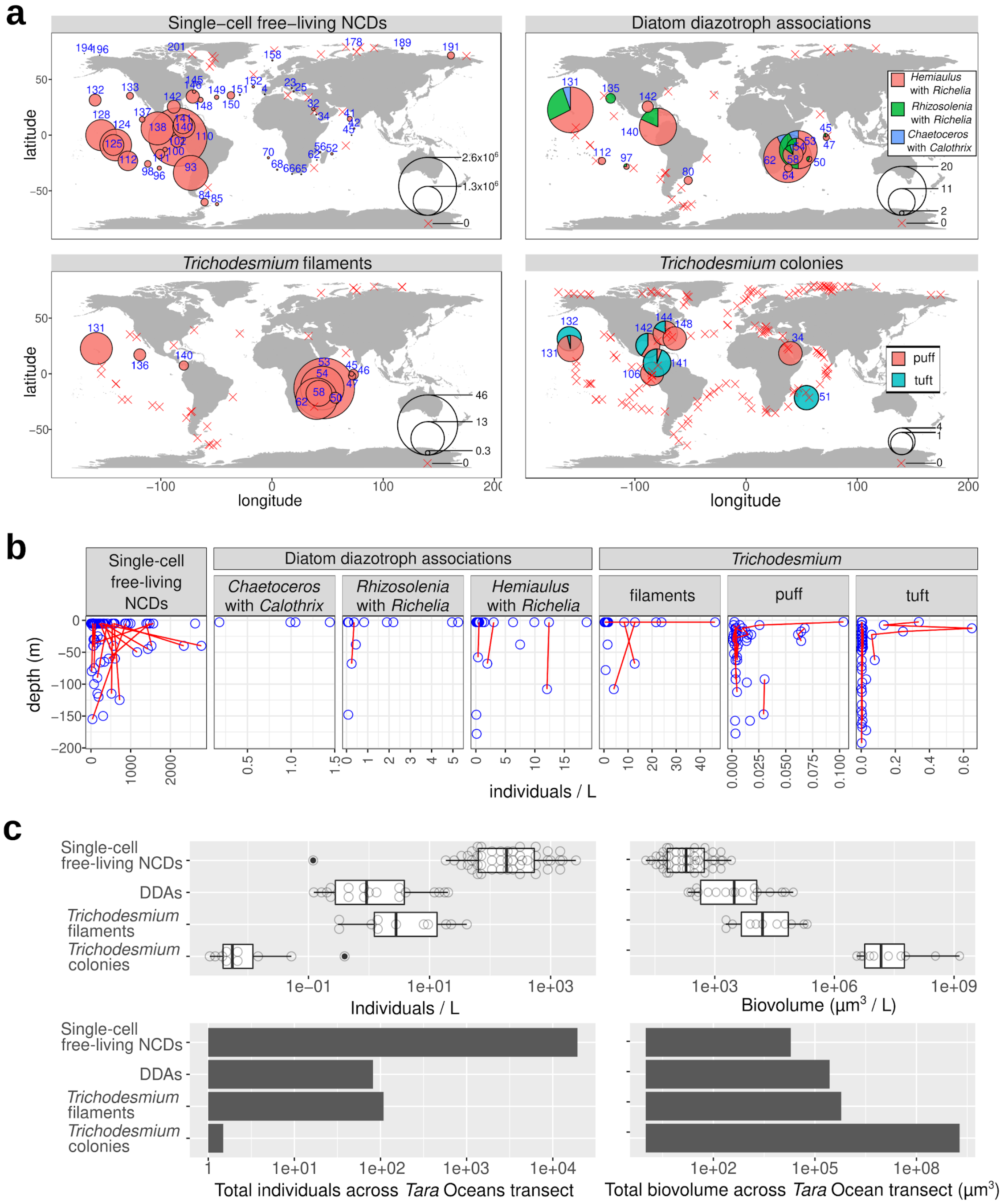


Figure 2: Abundance and distribution of diazotrophs by quantitative imaging methods. (**a**) Biogeography in surface waters. Bubble size varies according to the corresponding diazotroph concentration (individuals/L), while crosses indicate their absence. Station labels with detection of diazotrophs are indicated in blue. (**b**) Depth partition. Samples from the same geographical site are connected by lines. (**c**) Distribution of individual abundances and biomass in surface waters. Single-cell free-living non-cyanobacterial diazotrophs (NCDs) were quantified by merging flow cytometry counts with *nifH*/*recA* ratio from metagenomes from size fraction 0.22-1.6/3 µm and assuming an arbitrary average cellular biovolume of 1 µm³. The detection and biovolume determinations of diatom-diazotroph associations (DDAs) and *Trichodesmium* free filaments were carried out by high-throughput confocal microscopy in samples from the 20-180 µm size fraction. In the case of *Trichodesmium* colonies, it was determined using images from the UVP5 .

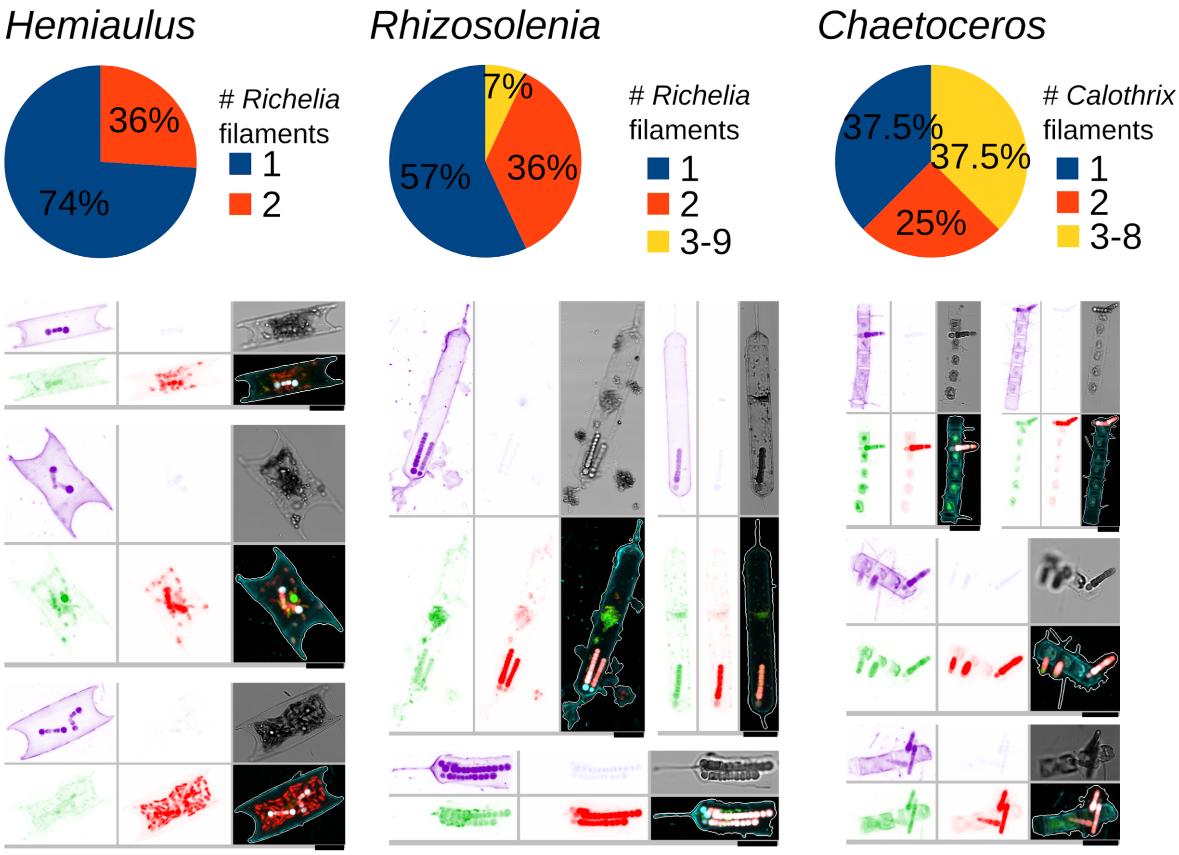


Figure 3: Variation in the number of *Richelia*/*Calothrix* filaments among the diatom-diazotroph associations observed by high-throughput confocal microscopy. Examples of images are shown. From top left to bottom right, the displayed channels for each micrograph correspond to cell surface (cyan, AlexaFluor 546 dye), DNA (blue, Hoechst dye), cellular membranes (green, DiOC6 dye), chlorophyll autofluorescence (red), the bright field, and the merged channels. The size bar at the bottom left of each microscopy image corresponds to 10 μm.

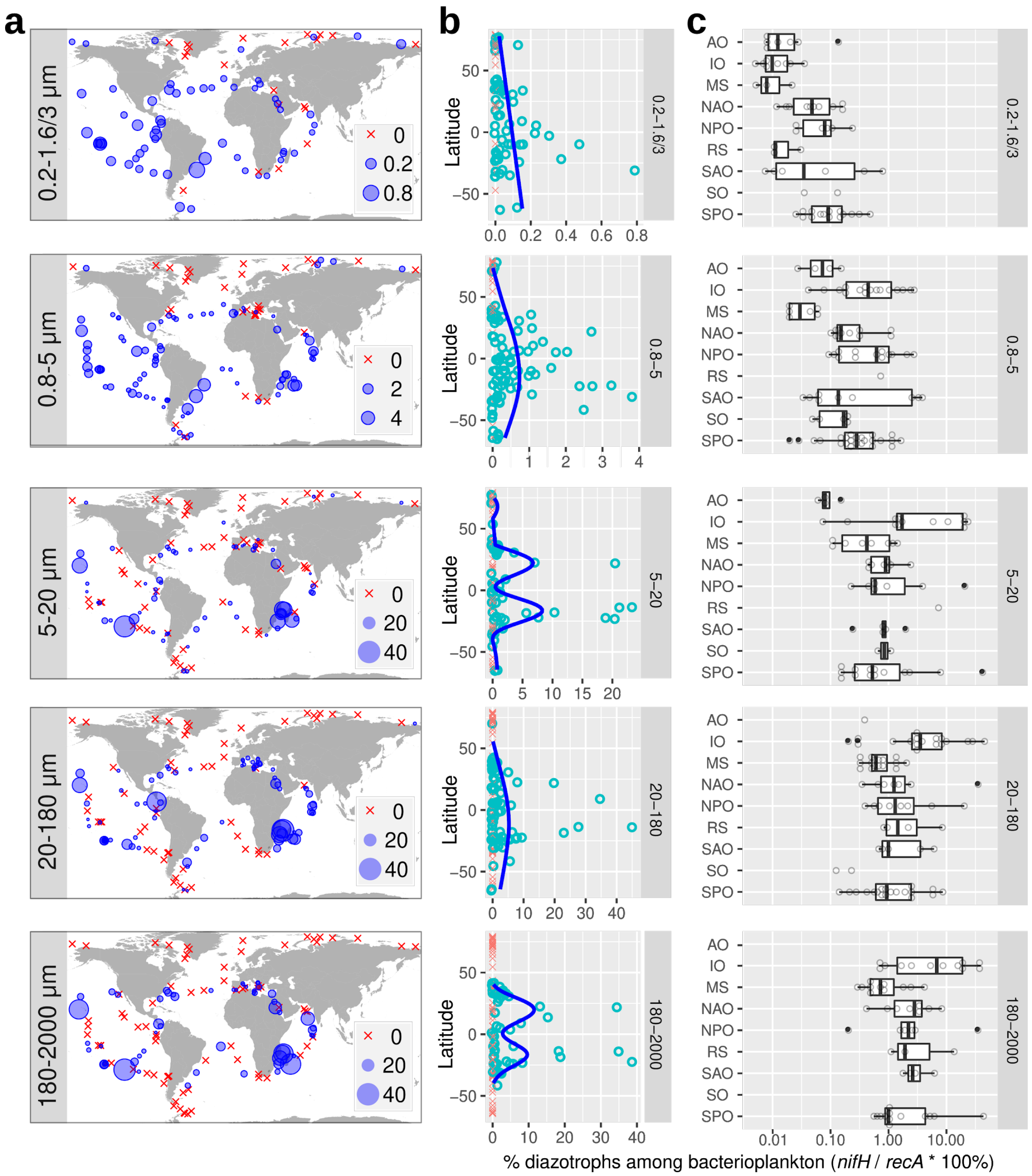


Figure 4: Biogeography of diazotrophs in surface waters using metagenomes obtained from different size-fractionated samples. The percentage of diazotrophs in the bacterioplankton community was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. (**a**) Biogeography. The bubble size varies according to the percentage of diazotrophs, while crosses indicate absence (i.e., no detection of *nifH* reads). (**b**) Latitudinal abundance gradient. The blue lines correspond to generalized additive model smoothings. (**c**) Ocean distribution. Abbreviations: MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; AO, Arctic Ocean.

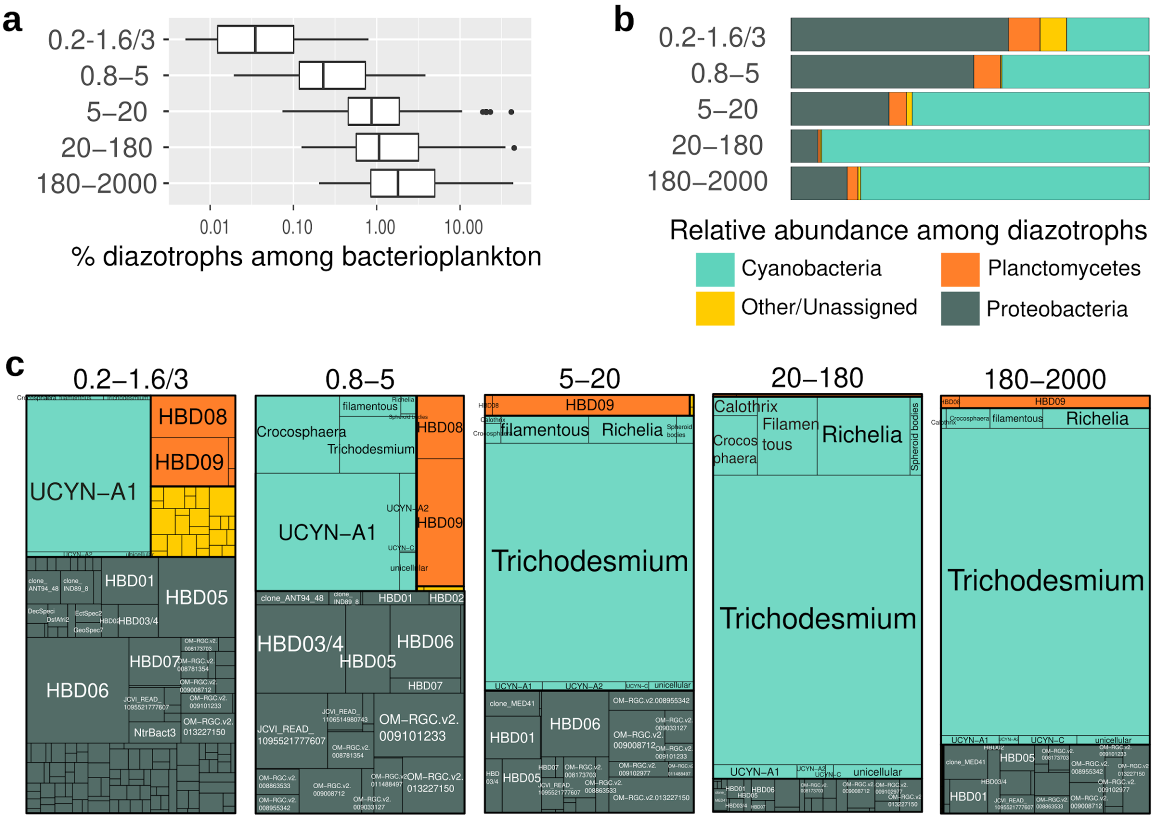


Figure 5: Abundance of diazotrophs in surface waters using metagenomes obtained from different size-fractionated samples. The percentage of diazotrophs in the bacterioplankton community was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. (**a**) Diazotroph abundance. (**b**) Taxonomic distribution. (**c**) Taxonomic distribution at deeper resolution.

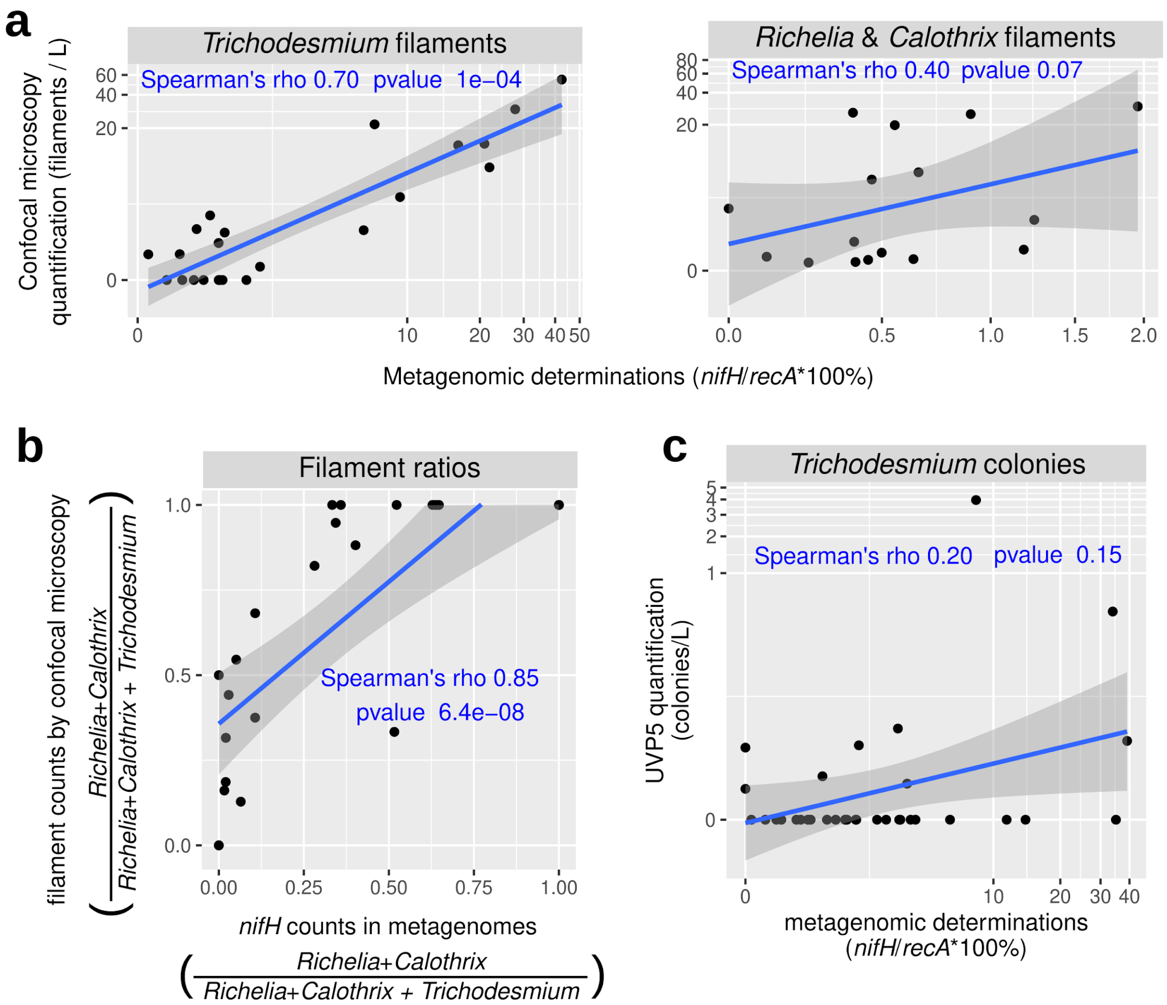


Figure 6: Correlation analysis between diazotroph quantifications by imaging and molecular methods. (**a-b**) Comparison between high-throughput confocal microscopy and metagenomics. *Calothrix*, *Richelia* and *Trichodesmium* in samples from size fraction 20-180 µm were measured by quantification of high-throughput confocal microscopy images (filaments L^{-1}) and by metagenomic counts (% of diazotrophs in the bacterioplankton community by the ratio between the marker genes *nifH* and *recA*). (**a**) Correlation of relative abundances in metagenomes and absolute abundances by confocal microscopy for the three taxa. (**b**) Correlation between the ratio of abundances between taxa. (**c**) Comparison between UVP5 and metagenomics. *Trichodesmium* colonies were measured by UVP5 quantification (colonies L^{-1}) and by metagenomic counts in the 180-2000 µm size-fractionated samples. Spearman rho correlation coefficients and p-values are displayed in blue.

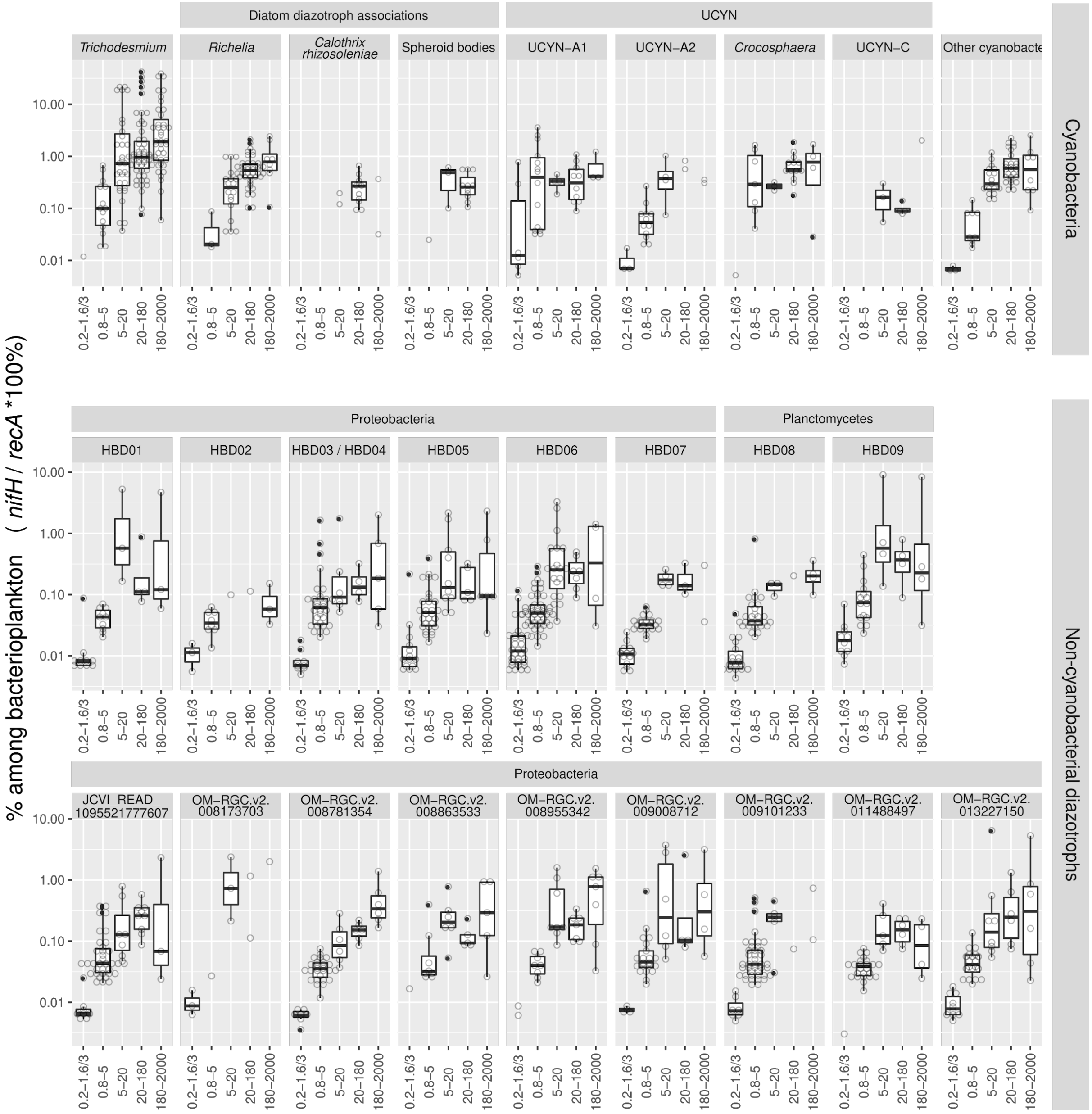


Figure 7: Distribution of the main diazotroph taxa across metagenomes obtained in different size-fractionated samples from surface waters. For each taxon, the percentage in the bacterioplankton community is estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. The lineages grouped into 'Other cyanobacteria' are displayed in Supplementary Table S1. The 'OM-RGC.v2' prefix indicates the the *nifH* sequences assembled from the metagenomes of <3 µm size fractions (Salazar al., 2019), while HBD01 to HBD09 corresponds to the metagenome-assembled genomes from the same samples (Delmont et al 2018).

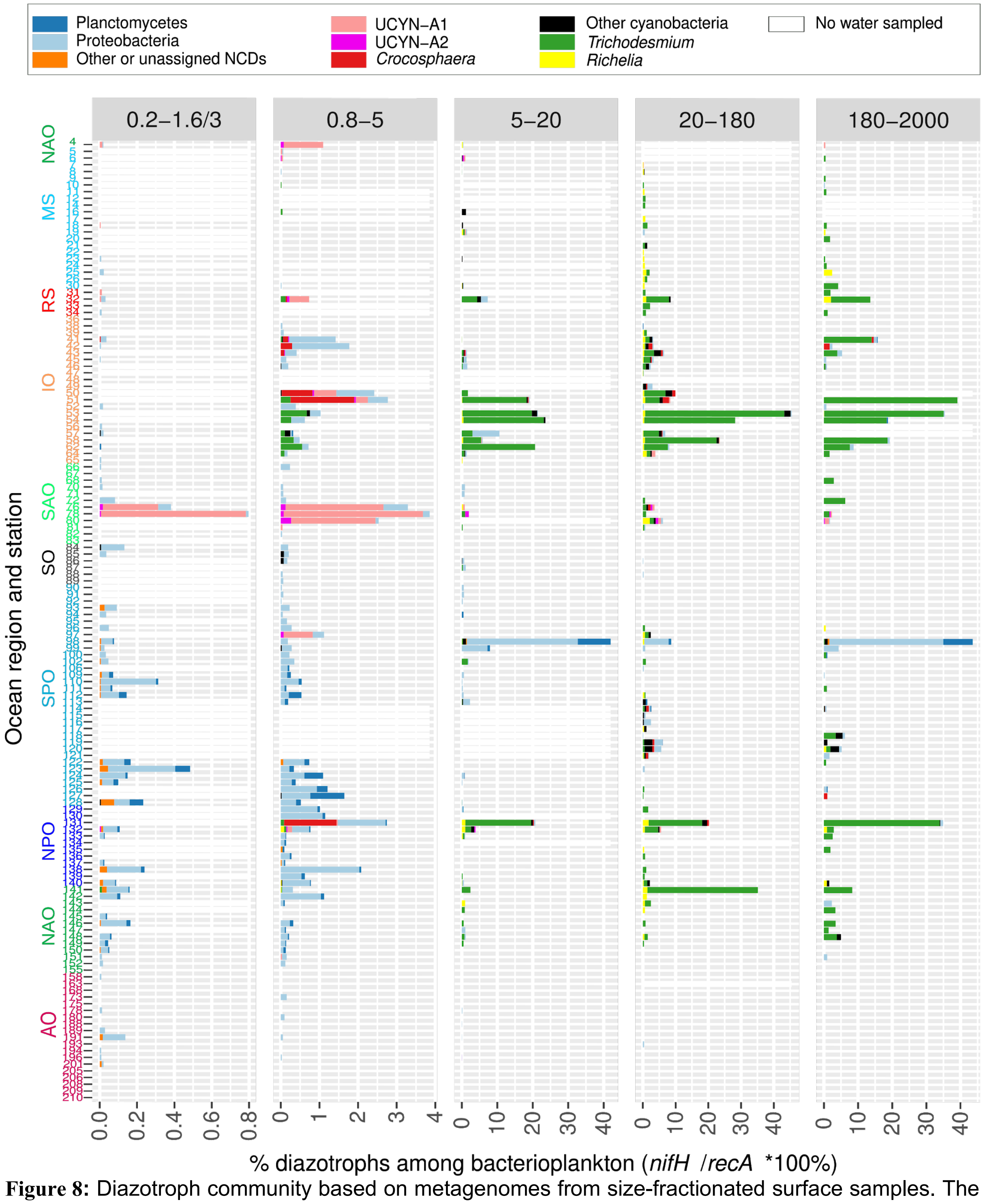


Figure 8: Diazotroph community based on metagenomes from size-fractionated surface samples. The percentage of diazotrophs in the bacterioplankton community was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. The bar color code shows the taxonomic annotation, while the absence of water sample is indicated by a white bar. The Y axis shows the *Tara* Oceans stations and the ocean regions. Abbreviations: MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; AO, Arctic Ocean. The equivalent figure showing the DCM water layer is shown in Figure S2 (note the differences in scales between both figures, showing the higher relative abundance of diazotrophs in the surface layer).

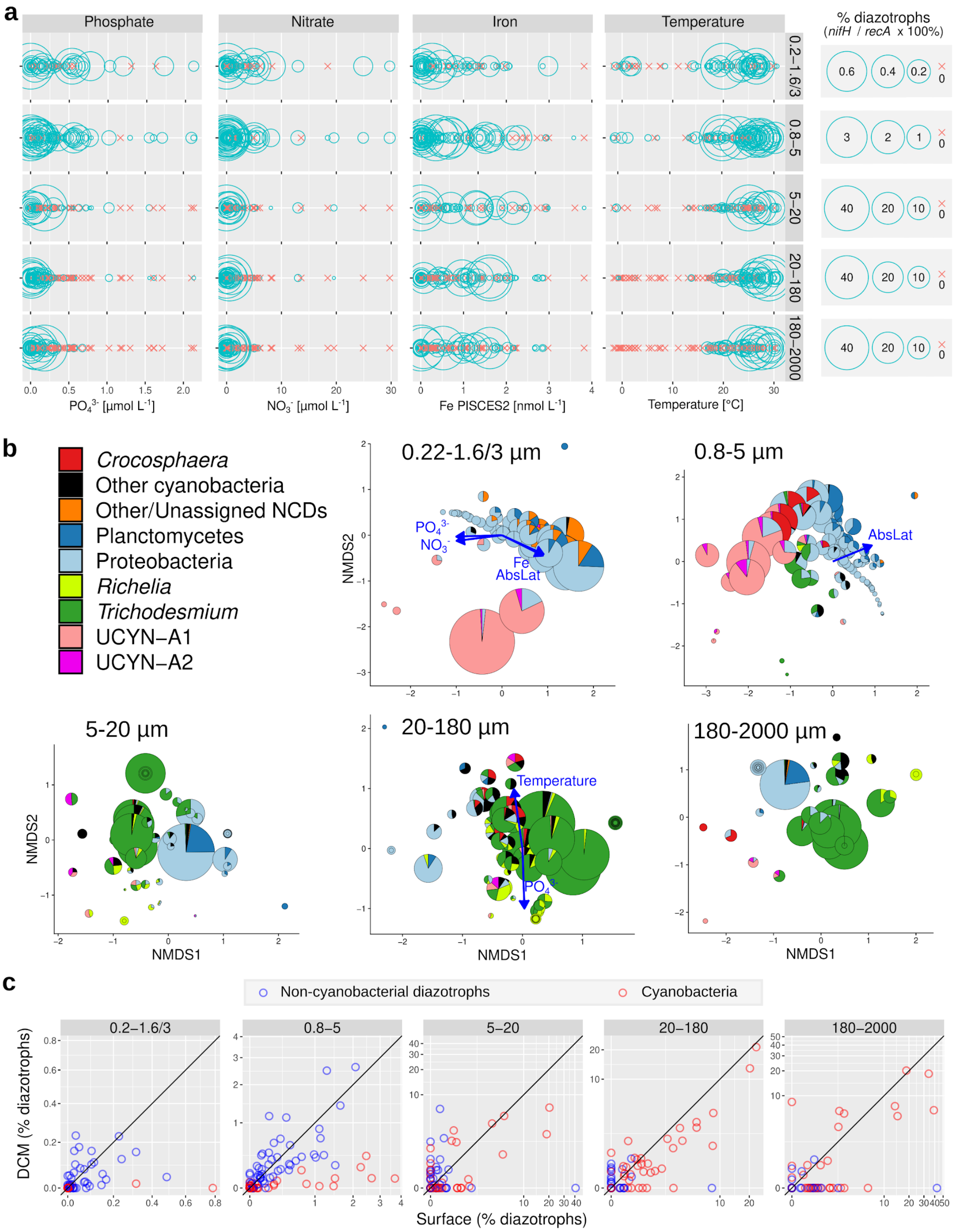


Figure 9: Environmental parameters and diazotroph distributions. (**a**) Distribution across gradients of nutrients and temperature in surface waters. Circles correspond to samples with diazotrophs, while crosses indicate absence (i.e., no detection of *nifH* reads). (**b**) NMDS analysis of stations according to Bray–Curtis distance between diazotroph communities of size-fractionated surface samples. Fitted statistically significant physico-chemical parameters are displayed (adjusted *P* value < 0.05). NMDS stress values: 0.07276045, 0.1122258, 0.1452893, 0.09693721, and 0.07969211. (**c**) Depth distribution. The scatter plots compare the diazotroph abundances between surface (5 m) and deep chlorophyll maximum (DCM; 17-180 m) for cyanobacteria (red points) and non-cyanobacterial diazotrophs (NCDs, blue points). Axes are in the same scale and the diagonal line corresponds to a 1:1 slope.

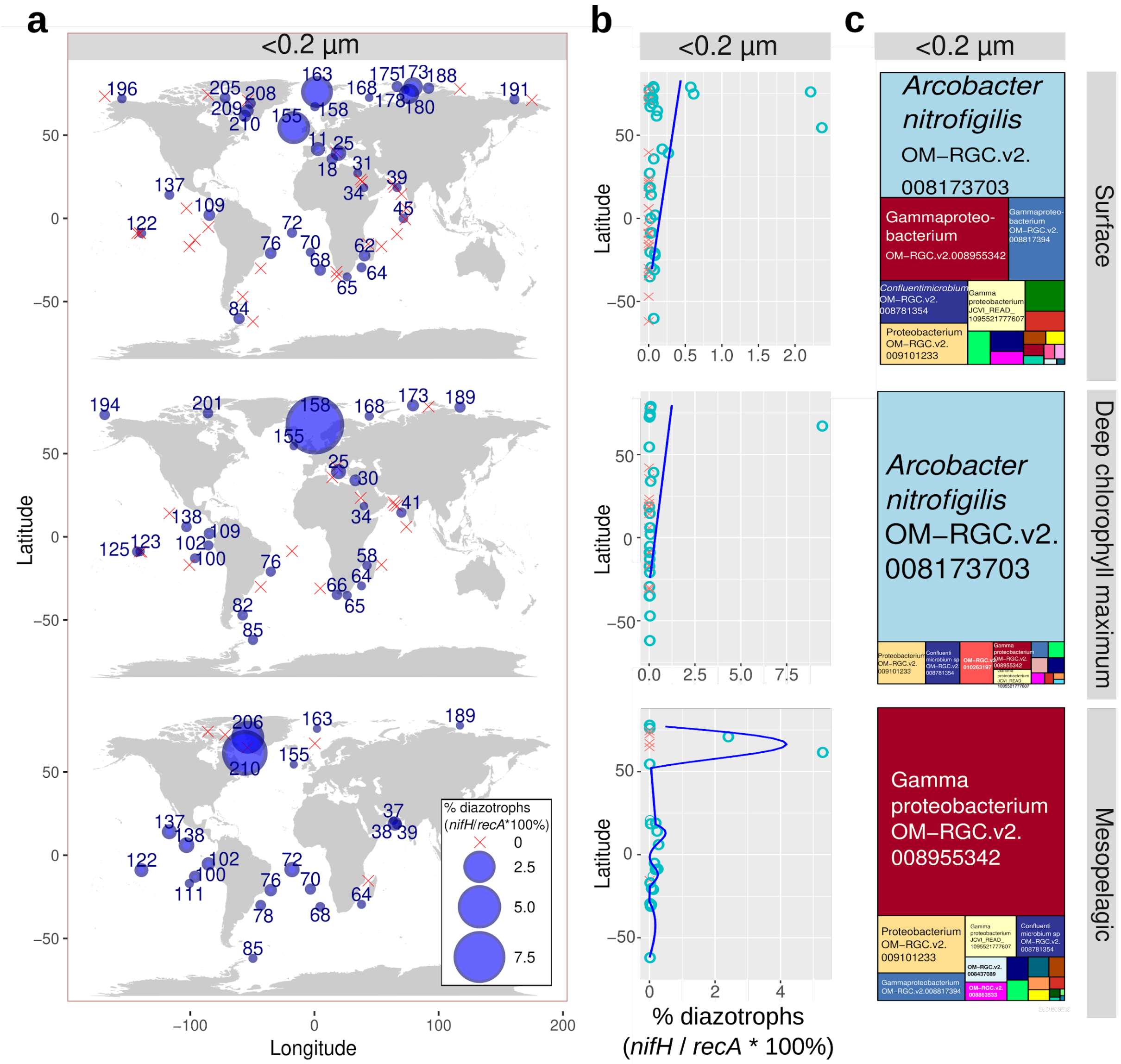
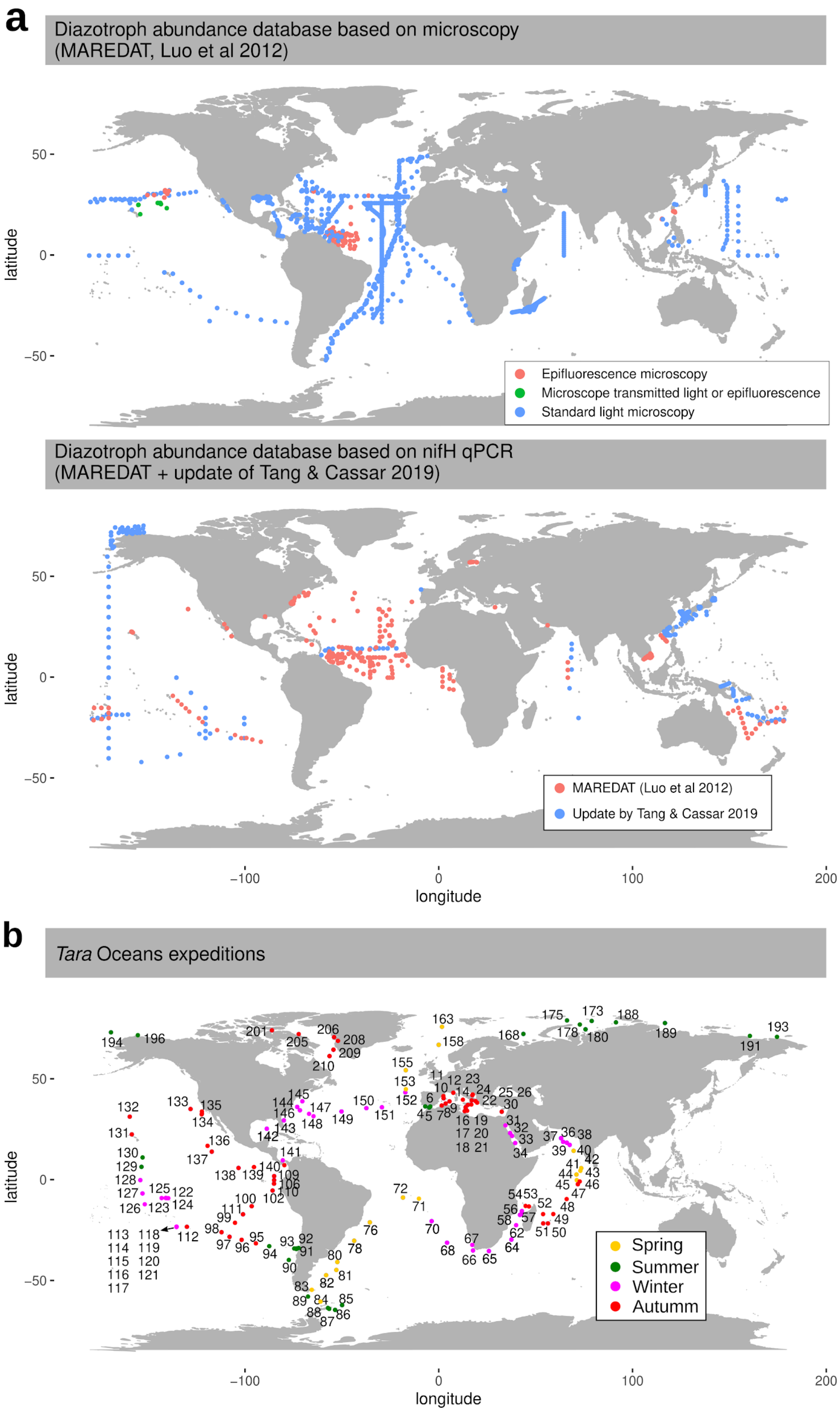
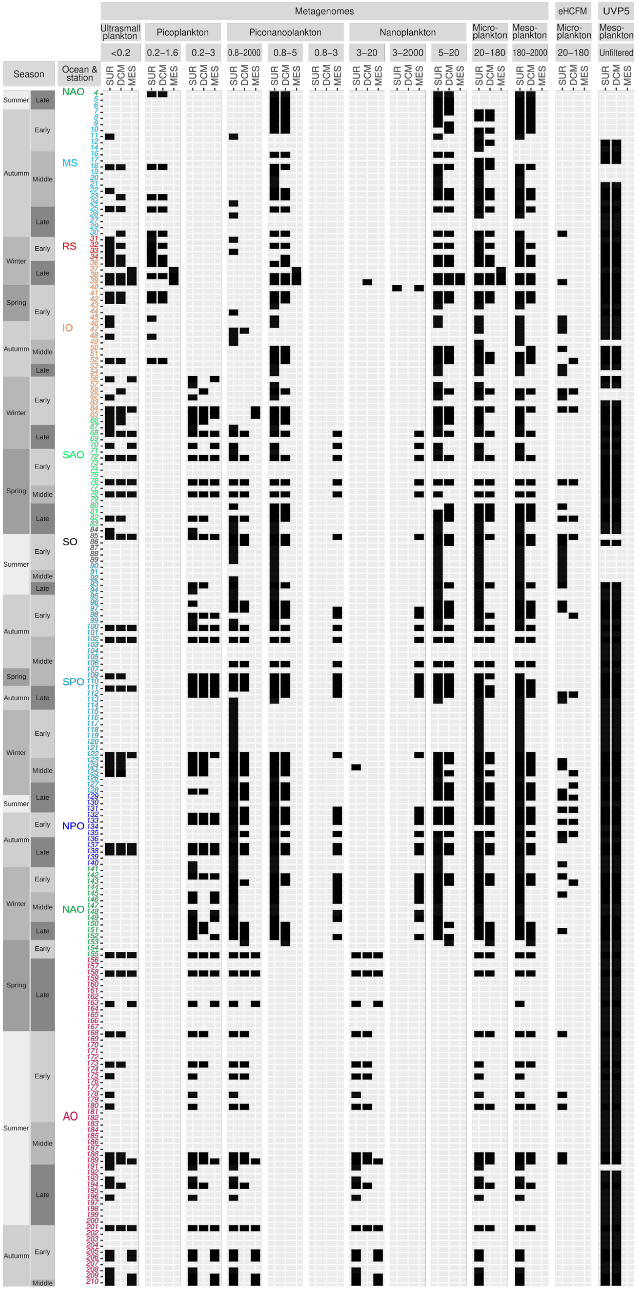


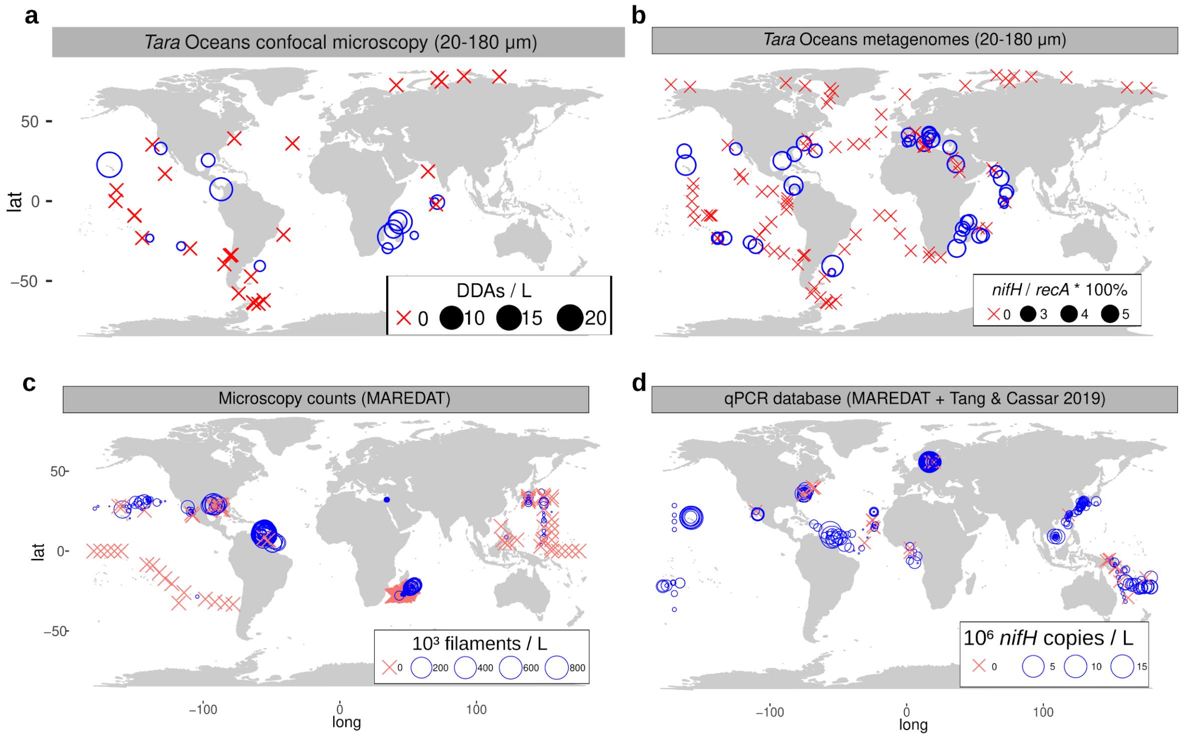
Figure 10: Detection of ultrasmall diazotrophs in metagenomes obtained from <0.22 µm size-fractionated samples of different water layers. The percentage of diazotrophs among ultrasmall bacterioplankton was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. (**a**) Biogeography. The bubble size varies according to the percentage of diazotrophs, while crosses indicate absence (i.e., no detection of *nifH* reads). Station labels with diazotrophs detection are indicated in blue. (**b**) Latitudinal abundance gradient. Circles correspond to samples with diazotrophs, while crosses indicate absence. The blue lines correspond to generalized additive model smoothings. (**c**) Taxonomic distribution. The 'OM-RGC.v2' prefix indicates the *nifH* sequences assembled from metagenomes of <3 µm size fractions (Salazar al., 2019), including <0.22 µm.



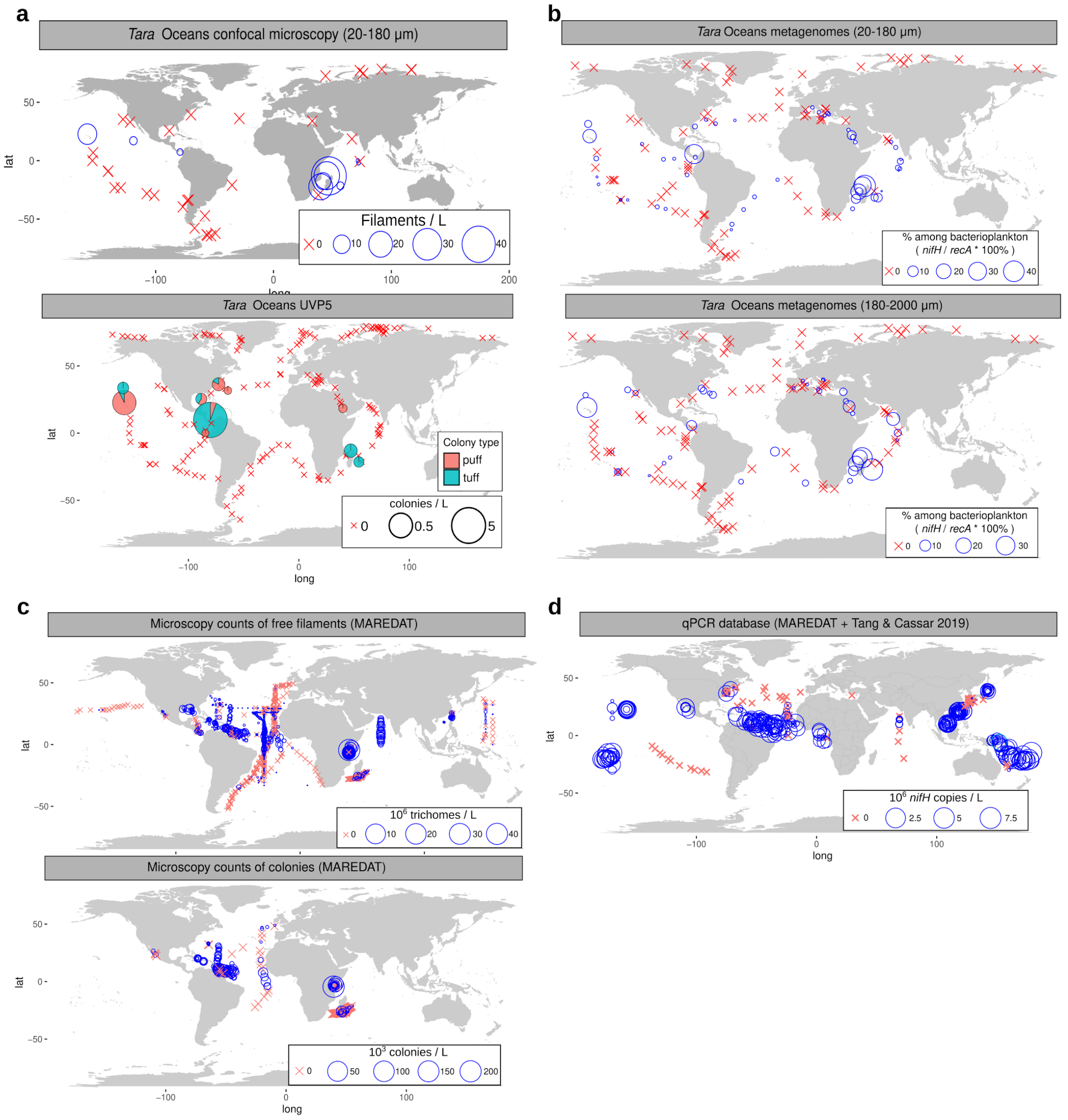
Supplementary Figure S1: Comparison between the databases of diazotroph distribution and the *Tara* Oceans expeditions. (**a**) The MARine Ecosystem DATa (MAREDAT) includes a database for microscopy counts (upper map) and for quantitative PCR targeting the *nifH* gene (lower map). The microscopy only covers *Trichodesmium* and diatom-diazotroph associations and it is a compilation of 44 different publications between 1966 and 2011 (Luo et al., 2012). The *nifH* dataset includes *Trichodesmium*, diatom-diazotroph associations, UCYN-A, *Crocosphaera* (UCYN-B) and UCYN-C and it is the result of 19 publications between 2005 and 2011 (red points; Luo et al., 2012). This later dataset has been recently updated by Tang and Cassar (2019) with measurements from 17 new publications between 2012 and 2018 (blue points). (**b**) Sampling route of the *Tara* Oceans expeditions (2009-2013), showing station labels and sampling season.



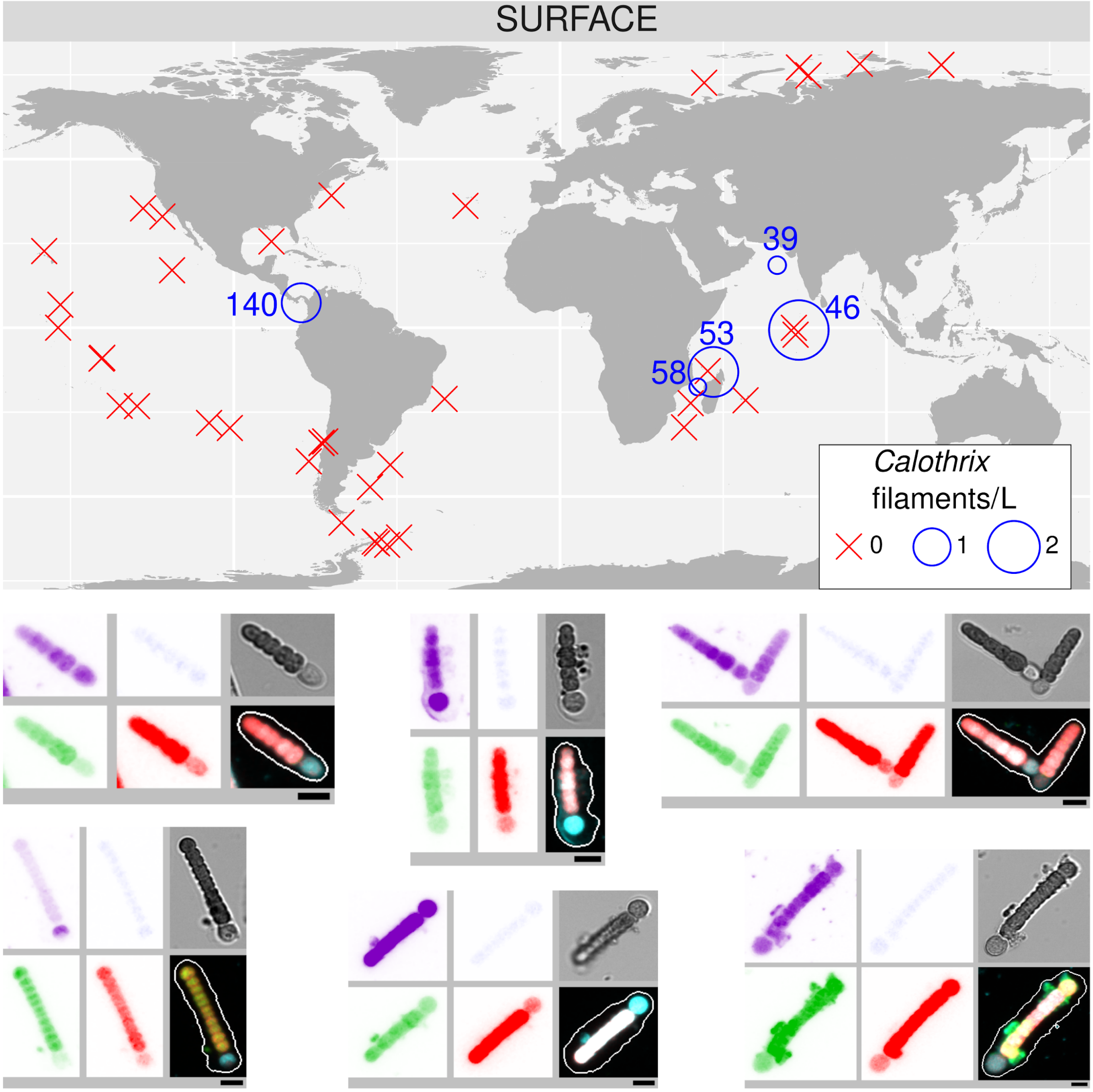
Supplementary Figure S2: Samples and methods used in this study. The current analysis of global diversity and abundance of diazotrophs was carried out across 197 *Tara* Oceans stations where samples where taken for metagenomic sequencing and/or for environmental High Content Fluorescence Microscopy (eHCFM) and/or images were taken *in situ* using an Underwater Vision Profiler 5 (UVP5). The analyzed samples are indicated as filled boxes. A complete sampling station consisted of collecting plankton from three distinct depth layers: surface (SUR), deep chlorophyll maximum (DCM), and mesopelagic (MES). The data from the bottom of the mixed layer was collected when no deep chlorophyll maximum was observed (stations TARA_123, TARA_124, TARA_125, TARA_152 and TARA_153). Plankton communities from SUR and DCM were fractionated into six main size classes: ultrasmall plankton (<0.22 µm), picoplankton (0.2 to 1.6 µm or 0.2 to 3 µm), piconanoplankton (0.8 to 5 µm or 0.8 to 2000 µm), nanoplankton (5 to 20 µm or 3 to 20 µm), microplankton (20 to 180 µm), and mesoplankton (180 to 2000 µm). For MES samples, size fractions were more heterogeneous (<0.22 µm, 0.2 to 1.6 µm, 0.2 to 3 µm, 0.8 to 3 µm, 0.8 to 5 µm, 0.8 to 200 µm, 0.8 to 2000 µm, 3-20 µm, 3-2000 µm, 5-20 µm). Season and moment of the season (early, middle, late) are displayed to the left of the panel. Station labels are coloured according to the ocean region: IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic Ocean; RS, Red Sea; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean.


a
Tara Oceans confocal microscopy (20-180 µm)
DDAs / L
0
10
15
20
b
Tara Oceans metagenomes (20-180 µm)
nifH / recA * 100%
0
3
4
5
c
Microscopy counts (MAREDAT)
10³ filaments / L
0
200
400
600
800
d
qPCR database (MAREDAT + Tang & Cassar 2019)
10⁶ nifH copies / L
0
5
10
15
lat
50
0
−50
long
−100
0
100

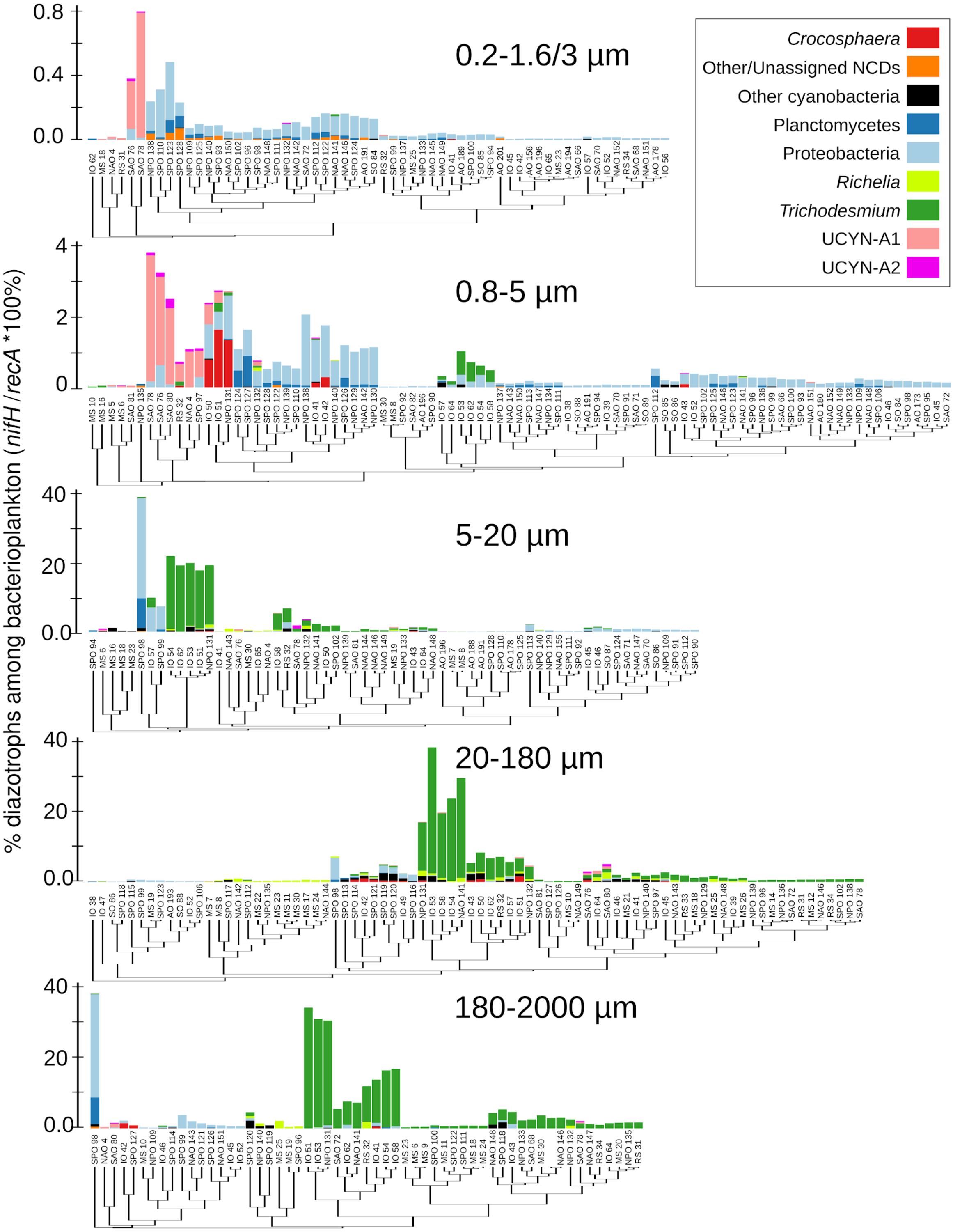
Supplementary Figure S3: Biogeography of diatom-diazotroph associations (DDAs) in surface waters. (**a-b**) Abundance across the *Tara* Oceans transect based on quantification of high-throughput confocal microscopy determinations (a) and from metagenomic read abundance of *nifH* gene (b). (**c-d**) Abundance in the MARine Ecosystem DATa (MAREDAT) database for microscopy counts (c) and for quantitative PCR targeting the *nifH* gene (d). This latter includes the recent compilation update by Tang and Cassar 2019. Bubble size varies according to the corresponding concentration, while crosses indicate their absence.


a
Tara Oceans confocal microscopy (20-180 µm)
Filaments / L
0
10
20
30
40
lat
long
Tara Oceans UVP5
Colony type
puff
tuff
colonies / L
0
0.5
5
b
Tara Oceans metagenomes (20-180 µm)
% among bacterioplankton
(nifH / recA * 100%)
Tara Oceans metagenomes (180-2000 µm)
c
Microscopy counts of free filaments (MAREDAT)
10⁶ trichomes / L
Microscopy counts of colonies (MAREDAT)
10³ colonies / L
50
100
150
200
d
qPCR database (MAREDAT + Tang & Cassar 2019)
10⁶ nifH copies / L
2.5
5
7.5
50
0
−50
−100
100
200

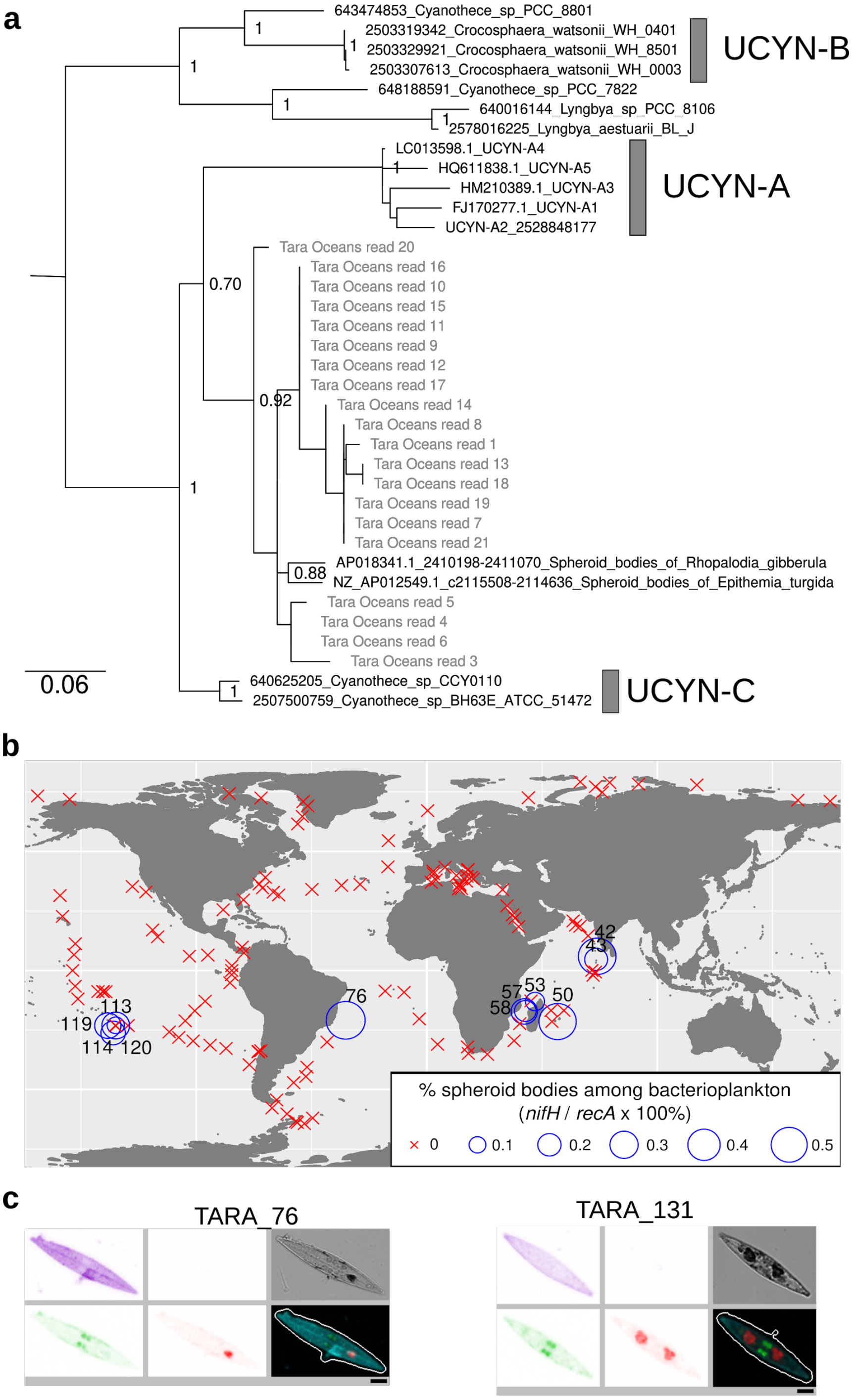
Supplementary Figure S4: Biogeography of *Trichodesmium* aggregates in surface waters. (**a**) Abundance across the *Tara* Oceans transect of free-filaments by high-throughput confocal microscopy in 20-180-µm size-fractionated samples (upper map) and colonies by Underwater Vision Profiler 5 (lower map). (**b**) Abundance across the *Tara* Oceans transect based on the metagenomic read abundance of the *nifH* marked gene in 20-180-µm and 180-2000-µm size-fractionated samples. (**c-d**) Abundance in the MARine Ecosystem DATa (MAREDAT) database for microscopy counts of free-filaments (c; upper map) and colonies (c; lower map) and for quantitative PCR targeting the *nifH* gene (d). This latter includes the recent compilation update by Tang and Cassar 2019. Bubble size varies according to the corresponding concentration, while crosses indicate their absence.



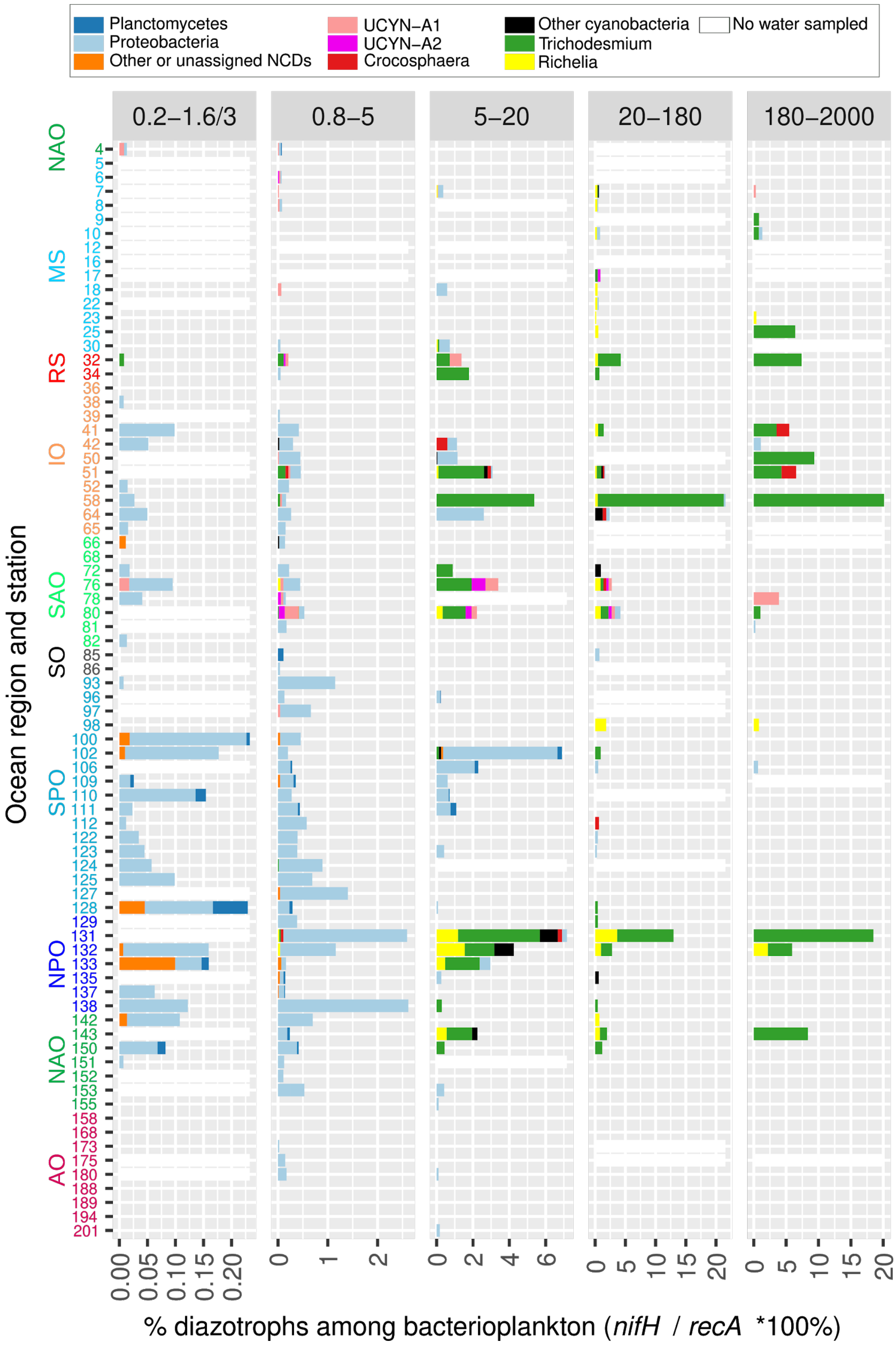
Supplementary Figure S5: Abundance and distribution of free filaments of *Richelia*/*Calothrix* in surface waters by quantification of high-throughput confocal microscopy images in samples from size fraction 20-180 µm. Maps show the biogeographical distribution. Bubble size varies according to the corresponding filament concentration, while red crosses indicate their absence. Examples of images are shown. From up left to bottom right, the displayed channels for each micrograph correspond to cell surface (cyan, AlexaFluor 546 dye), DNA (blue, Hoechst dye), cellular membranes (green, DiOC6 dye), chlorophyll autofluorescence (red), the bright field, and the merged channels. The size bar at the bottom left of each microscopy image corresponds to 2.5 µm.



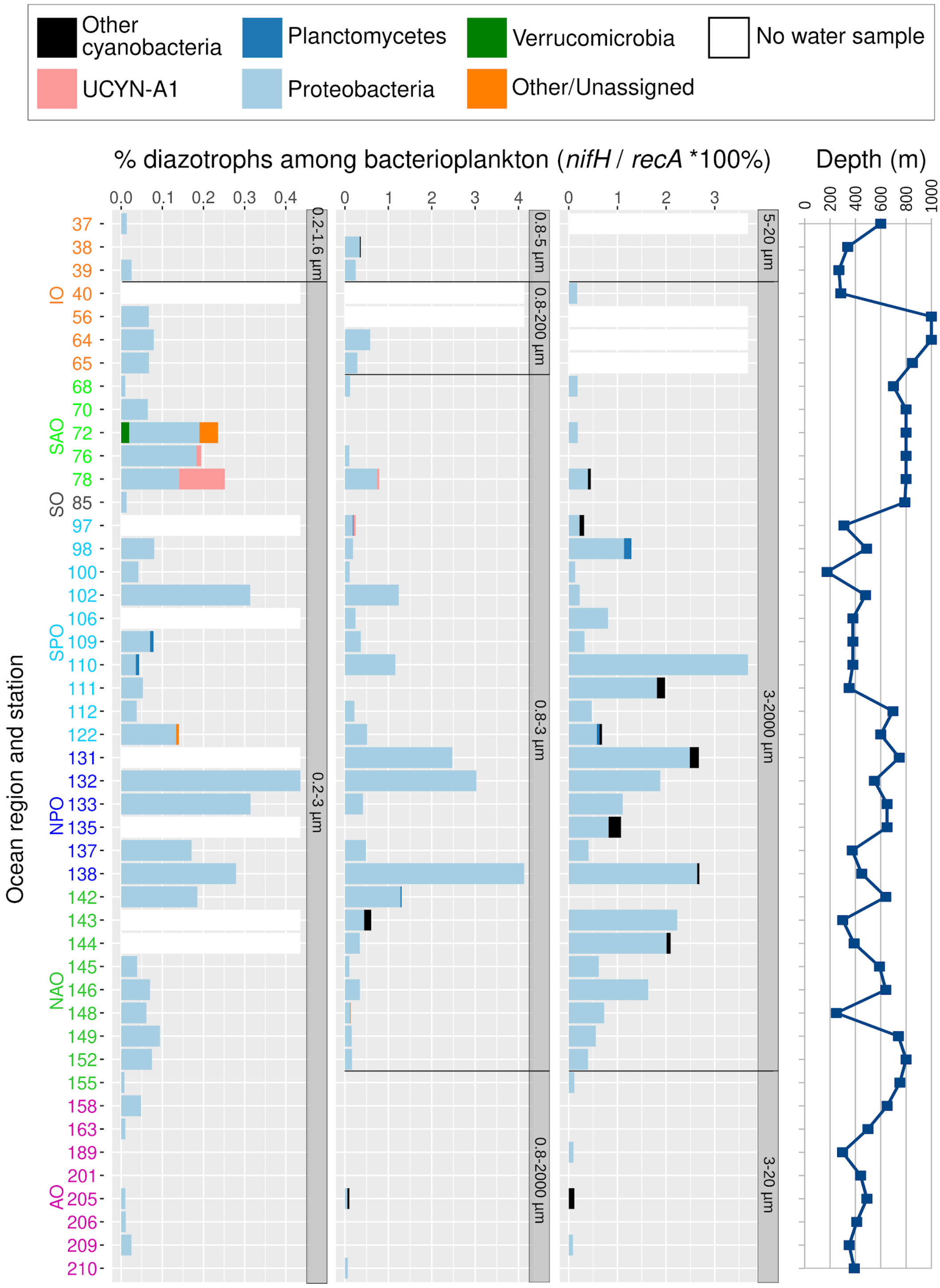
Supplementary Figure S6: Clusters of diazotroph communities based on metagenomes from size-fractionated surface samples. For each size fraction, the samples are sorted by similarity using hierarchical clustering (Bray–Curtis distance) and the corresponding diazotroph relative abundances are displayed as bar plots, with the color code according to the taxonomic annotation. The percentage of diazotrophs in the bacterioplankton community was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. The dendrogram tip labels show the *Tara* Oceans stations and the ocean regions. Abbreviations: MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; AO, Arctic Ocean.



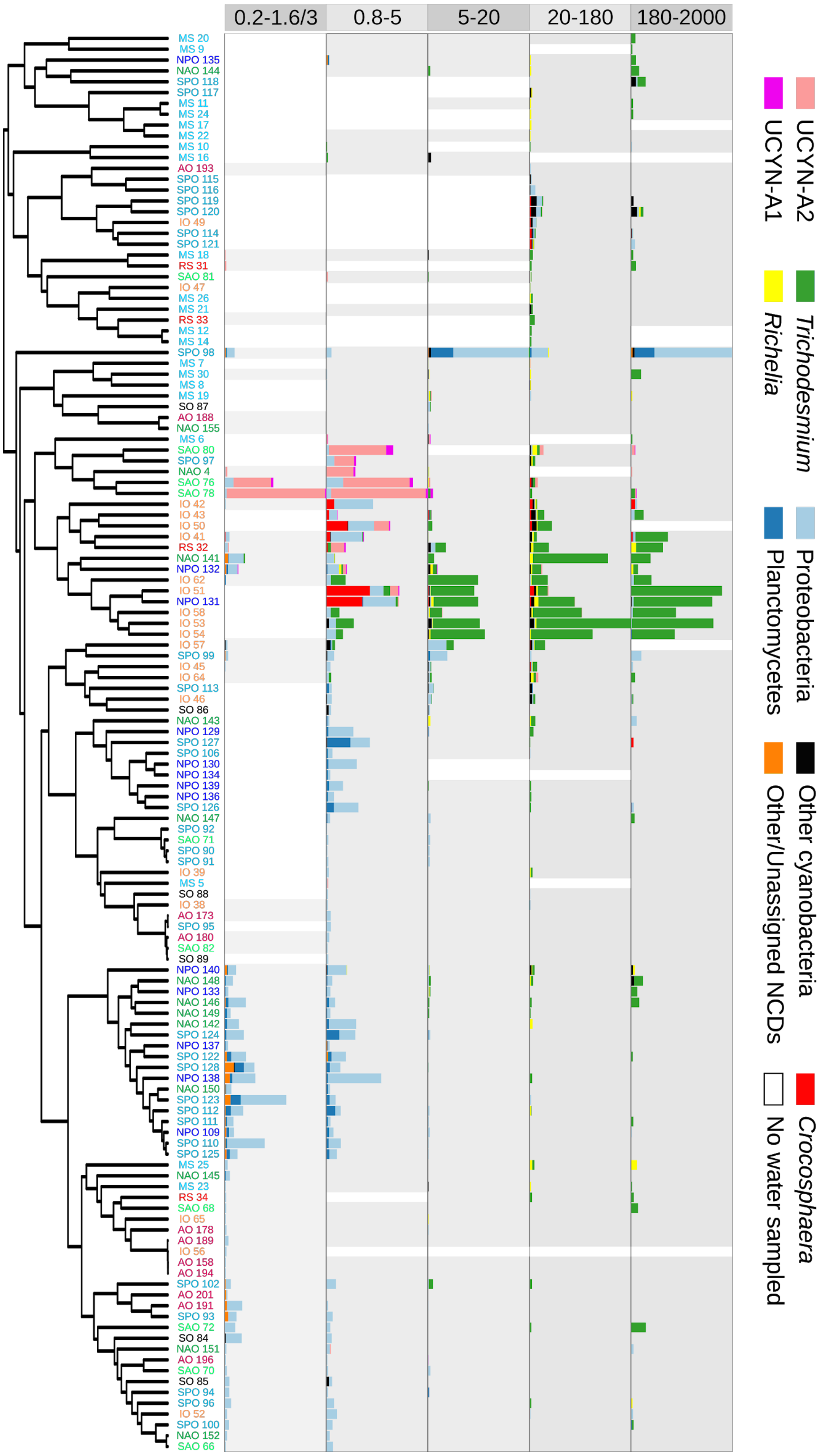
Supplementary Figure S7: Putative spheroid bodies in *Tara* Oceans samples. (**a**) Phylogeny of metagenomic reads with sequence similarity to the *nifH* gene from spheroid bodies. NCBI or IMG accession numbers of reference nucleotide sequences and the species names are indicated in the tip labels. The aLRT values are shown for the main clades. (**b**) Biogeography in surface waters of 20-180 µm size fractionated samples. The bubble size varies according to the percentage of reads of potential spheroid bodies, while crosses indicate absence (i.e., no detection of *nifH* reads). Station labels with read detection are indicated. (**c**) Images of pennate diatoms containing round granules that lack chlorophyll autofluorescence that were observed in the same samples where putative metagenomic sequences from spheroid-bodies were detected. From up left to bottom right, the displayed channels for each micrograph correspond to cell surface (cyan, AlexaFluor 546 dye), DNA (blue, Hoechst dye), cellular membranes (green, DiOC6 dye), chlorophyll autofluorescence (red), the bright field, and the merged channels. The size bar at the bottom left of each microscopy image corresponds to 2.5 µm.



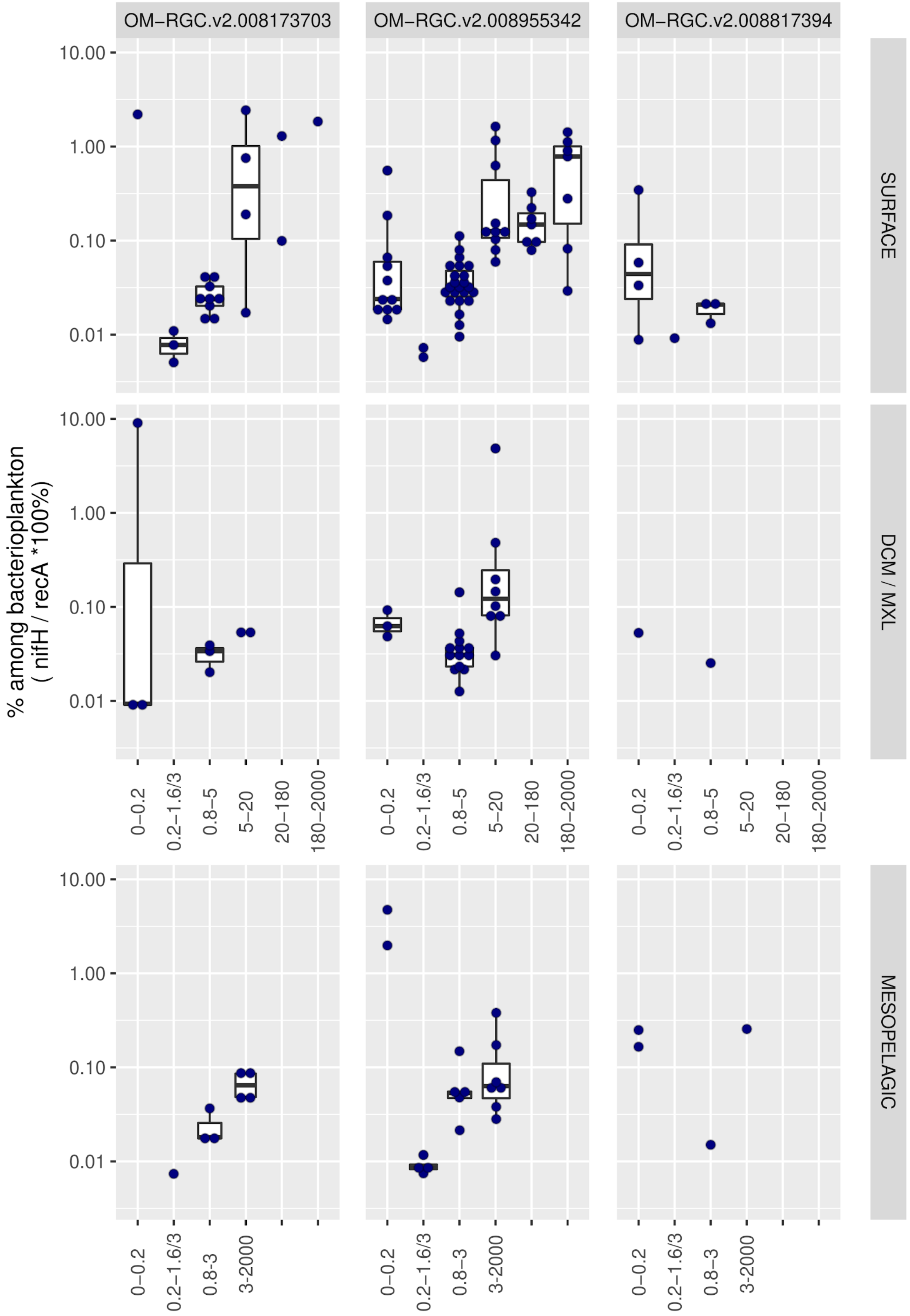
Supplementary Figure S8: Diazotroph community based on metagenomes from size-fractionated samples derived from deep-chlorophyll maxima. The percentage of diazotrophs in the bacterioplankton community was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. The bar color code shows the taxonomic annotation, and the absence of water sample is indicated by a white bar. The Y axis shows the *Tara* Oceans stations and the ocean regions. Abbreviations: MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; AO, Arctic Ocean. The equivalent figure showing the surface layer is shown in Figure 7 (note the differences in scales between both figures, showing the higher relative abundance of diazotrophs in the surface layer). The data from the bottom of the mixed layer is displayed when no deep chlorophyll maximum was observed (stations TARA_123, TARA_124, TARA_125, TARA_152 and TARA_153).



Supplementary Figure S9: Diazotroph community based on metagenomes from size-fractionated samples from mesopelagic depths. The percentage of diazotrophs in the bacterioplankton community was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. The bar color code shows the taxonomic annotation, and the absence of water sample is indicated by a white bar. Size fractions are also indicated (they are more heterogeneous than those from surface and deep chlorophyll maximum samples). The Y axis shows the *Tara* Oceans stations and the ocean regions. Abbreviations: MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; AO, Arctic Ocean. Sampling depth is indicated in the right panel.



Supplementary Figure S10: Clusters of diazotroph communities based on metagenomes from size-fractionated surface samples. For each size fraction, the samples are sorted by similarity using hierarchical clustering (Bray–Curtis distance) and the corresponding diazotroph relative abundances are displayed as bar plots, with the color code according to the taxonomic annotation. The percentage of diazotrophs in the bacterioplankton community was estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. The dendrogram tip labels show the *Tara* Oceans stations and the ocean regions. Abbreviations: MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean; AO, Arctic Ocean.



Supplementary Figure S11: Distribution of potential ultrasmall diazotrophs across metagenomes obtained in different size-fractionated samples. For each taxon, the percentage in the bacterioplankton community is estimated by the ratio of metagenomic read abundance between the marker genes *nifH* and *recA*. The 'OM-RGC.v2' prefix indicates the *nifH* sequences assembled from the metagenomes of <0.22 µm size fraction (Salazar al., 2019).