Diversity, ecology and evolution of marine diazotrophic microorganisms

Francisco Miguel Cornejo-Castillo

Barcelona, December 2017

Universitat Politècnica de Catalunya

Institut de Ciències del Mar (ICM-CSIC)

Diversity, ecology and evolution of marine diazotrophic microorganisms

Francisco Miguel Cornejo Castillo

Directora: Dra. Silvia González Acinas
Depto. Biología Marina y Oceanografía
Instituto de Ciencias del Mar (ICM-CSIC)

Noviembre 2017

Tesis doctoral presentada para la obtención del título de Doctor
por la Universitat Politècnica de Catalunya
Programa de Doctorado de Ciencias del Mar

A las mujeres que han hecho posible esta tesis,
a Isabel, mi madre,
a Silvia, mi jefa,
a Ana María
y a Clara.

Contents

RESUMEN

La fijación biológica de nitrógeno, es decir, la reducción del nitrógeno (N_2) a amonio, es un proceso fundamental ya que representa una fuente de nitrógeno para la vida marina en áreas donde este elemento es limitante, posibilitando la producción primaria y por tanto la exportación de carbono al océano profundo. Este proceso se lleva a cabo por microorganismos procariotas, los llamados diazotrofos. Sin embargo, aún sabemos muy poco sobre la identidad y la ecología de estos microorganismos, lo que limita enormemente nuestra comprensión de la importancia global de este proceso, y nuestra capacidad de predecir cambios en la fijación de N_2 ligados a cambios en el ambiente. El objetivo de esta tesis, por tanto, fue ahondar en el conocimiento de la diversidad, ecología y evolución de los microorganismos diazotrofos en el océano.

La mayoría del conocimiento actual sobre la diversidad de diazotrofos se deriva del gen marcador *nifH*, que codifica una proteína estructural del complejo enzimático responsable de la fijación de nitrógeno. Por tanto, en el Capítulo 1 realizamos una exploración global del gen *nifH* usando datos metagenómicos de 68 estaciones muestreadas durante la campaña oceanográfica *Tara* Oceans. Nuestra aproximación se diferencia de los estudios anteriores ya que no se basa en el uso de cebadores para detectar el *nifH* y posibilita por tanto una cuantificación más precisa de la diversidad real. Este estudio representa el primer mapa global (no basado en cebadores) de la distribución de diazotrofos en el océano desde superficie hasta el mesopelágico. Aunque la abundancia de diazotrofos fue muy baja en general, era significativamente mayor en el océano profundo. Asimismo, descubrimos nuevos diazotrofos que habían pasado desapercibidos en los estudios basados en cebadores: más de la mitad de los diazotrofos detectados no se capturan por los cebadores para el *nifH*. Esto sugiere que la mayoría de estudios previos pueden haber obviado una fracción importante de las comunidades de fijadores de nitrógeno.

Entre los diazotrofos detectados en el Capítulo 1, el más abundante fue la cianobacteria unicelular *Candidatus* Atelocyanobacterium thalassa (UCYN-A), que vive en simbiosis con un alga primnesiofita y que juega un papel importante en la fijación de nitrógeno. En los capítulos 2 y 3 nos dedicamos a estudiar en detalle los

aspectos relacionados con la ecología, diversidad y evolución de este diazotrofo. Mediante el análisis de metagenomas y de técnicas de visualización microscópicas como el CARD-FISH pudimos detectar UCYN-A en el atlántico sur, revelando que UCYN-A1 y UCYN-A2, dos linajes diferentes de UCYN-A, viven en simbiosis con dos hospedadores diferentes, dos primnesiofitas de tamaños distintos. Además, el análisis del perfil de expresión del genoma de ambos linajes mostró una dedicación optimizada a la fijación de nitrógeno. La edad de divergencia de UCYN-A se estimó en unos 100 millones de años, y presumiblemente ocurrió bajo presiones evolutivas de tipo estabilizadora. Por último, en el Capítulo 3, nos centramos en el estudio de UCYN-A3, otro linaje del que se sabe muy poco. Mediante el uso de varios métodos (PCR, qPCR, CARD-FISH y metagenomas) se logró visualizar e identificar por primera vez el linaje UCYN-A3 asociado con una alga de tamaño diferente, lo que sugiere que los distintos linajes de UCYN-A ocupan diferentes compartimentos planctónicos que no siempre se consideran en estudios de diversidad de *nifH* o de fijación de nitrógeno. Finalmente, pudimos reconstruir una fracción importante del genoma de UCYN-A3, estableciendo que representa una especie genómica diferente a las anteriores.

En definitiva, esta tesis ha contribuido significativamente al conocimiento de los diazotrofos en el océano mediante el descubrimiento de nueva diversidad como de nuevos compartimentos del plancton donde puede darse la fijación de nitrógeno y que podrían ayudar a entender mejor el ciclo marino del nitrógeno.

SUMMARY

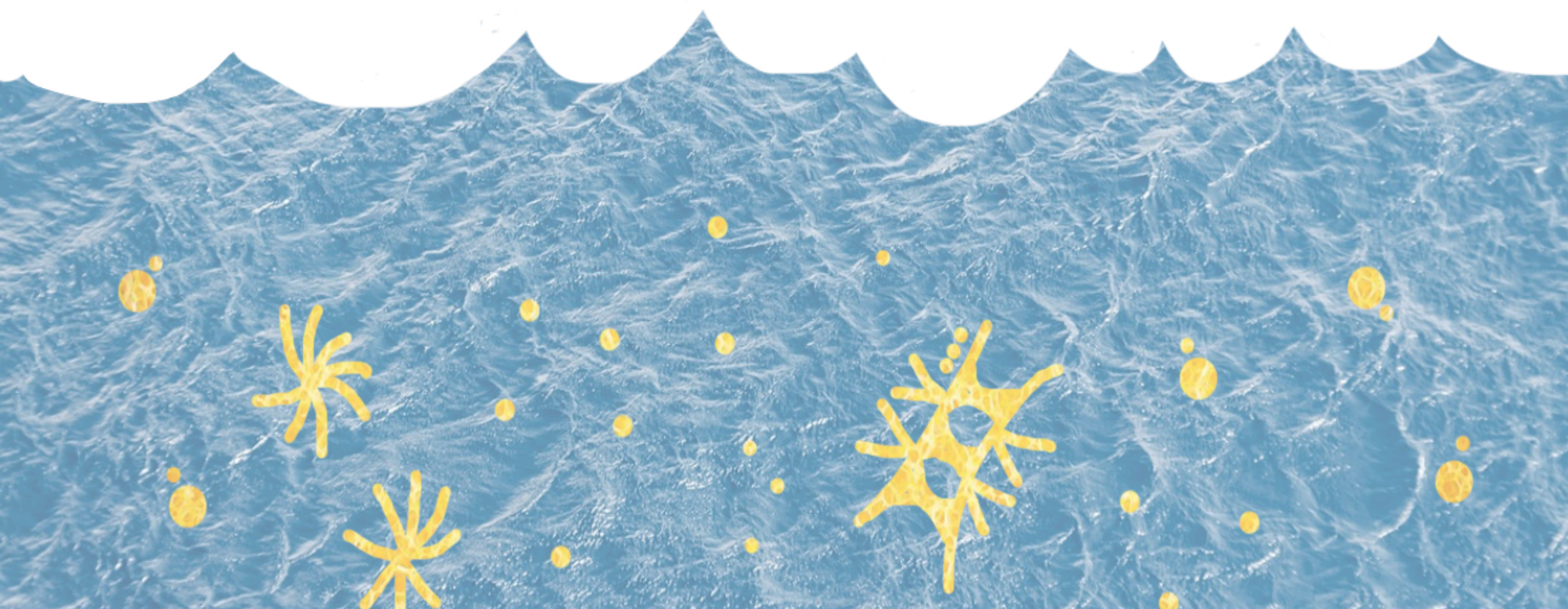
Biological N_2 fixation, the reduction of dinitrogen (N_2) gas to biologically available nitrogen, is a fundamental process since it represents a source of new nitrogen for marine life in areas where this important element can be limiting, supporting primary productivity and thus biological carbon export to the deep ocean. This process is performed by the nitrogen-fixing prokaryotic microorganisms (the so-called diazotrophs). However, very little is still known about the identity and ecology of diazotrophs, which largely limits our capacity to understand the global significance of this process, and to predict potential variations in nitrogen fixation upon changes in environmental conditions. In this thesis, we aimed at improving the knowledge on the diversity, ecology and evolution of the marine nitrogen-fixing microorganisms in the open ocean.

Most current knowledge on diazotrophic diversity has been obtained using the *nifH* marker gene, which encodes for a structural protein of the enzymatic complex that performs the N_2 fixation reaction. Thus, in Chapter 1 we first conducted a global exploration of the *nifH* gene extracted from metagenomic data derived from 68 globally distributed stations collected during the *Tara* Oceans expedition. This approach differs from previous studies in that it does not rely on primers to detect the *nifH* genes, and thus allows a more quantitative estimation of the contribution of these microorganisms and a more realistic view of their diversity. This study provides a first 'primer-free' global map of the distribution of open ocean diazotrophic communities across ocean basins and throughout the water column, showing that diazotrophs often occurred at very low abundances, and that in general they were significantly more abundant in the mesopelagic than in photic waters. Likewise, we uncovered novel diversity that had remained unnoticed in all previous primer-based studies, since we demonstrate that more than half of the detected *nifH* variants cannot be captured by the primers used. This suggests that most diazotroph diversity studies may be disregarding an important fraction of the nitrogen-fixing community members.

Among the diazotrophs detected in Chapter 1, the most abundant was the unicellular cyanobacterium *C. Atelocyanobacterium thalassa* (UCYN-A), which lives in symbiosis with a prymnesiophyte alga and has been shown to be a relevant player in nitrogen fixation. Thus, in Chapter 2 and Chapter 3, we explored aspects related to the ecology, diversity and evolution of this remarkable microorganism. We detected UCYN-A in the South Atlantic Ocean using not only metagenomic approaches but also microscopic visualization techniques (CARD-FISH). This allowed us to unveil that different UCYN-A lineages, UCYN-A1 and UCYN-A2, live in symbiosis with two distinct prymnesiophyte partners of different sizes. Both UCYN-A lineages showed a streamlined genome expression towards nitrogen fixation. We estimated that these two lineages diverged almost 100 Mya under a strong purifying selection process. Finally, in Chapter 3 we focused on the study of UCYN-A3, another lineage of which very little was known, to gain insight into its ecology. Using an array of methods (PCR, qPCR, CARD-FISH and metagenomes) we could visualize and identify for the first time UCYN-A3 and its association with an alga of different size, which suggests that different UCYN-A lineages occupy different planktonic compartments that are not always considered when nitrogen fixation of nifH diversity are studied. Finally, we manage to reconstruct a significant fraction of its genome, establishing that this lineage constitutes a new UCYN-A genomic species.

Overall, this thesis has significantly contributed to expand the knowledge on marine diazotrophic organisms, unveiling new diversity and new planktonic compartments that could potentially lead to a better understanding of the marine nitrogen cycle.

General Introduction



General Introduction

The nitrogen cycle and the relevance of nitrogen fixation in the ocean

Global nutrient cycling cannot be understood without microbes, which act as microscopic agents of change by transforming carbon and nutrients, such us nitrogen and phosphorous, throughout the Earth`s biomes (Arrigo, 2005). Within the Earth`s biomes, the oceans occupy a central feature of the biosphere, with biogeochemical connections to the land and atmosphere. Therefore, because the oceans represent the largest biome on Earth covering almost three quarters of the Earth`s surface, those chemical transformations happening in the ocean, both biotic and abiotic, are of fundamental importance for the functioning of the whole system.

Of the Earth`s elements, nitrogen comprises the majority of the atmosphere (78%) and is the fourth most abundant element in cellular biomass and, thus, an essential element for all life forms. Likewise, in marine environments, more than 95% of nitrogen mostly occurs as inert dissolved N_2 gas, but this form of nitrogen is unavailable for the majority of living organisms. The nitrogen cycle in the ocean is driven by complex microbial transformations and consists of five accepted nitrogen-transformation flows (Figure 1): (i) ammonification, including nitrogen fixation, and assimilatory and dissimilatory reduction of nitrite; (ii) nitrification; (iii) denitrification, including canonical, nitrifier-dependent and methane-oxidation-dependent denitrification; (iv) anammox, as a form of coupled nitrification-denitrification; and (v) nitrite–nitrate interconversion. The general processes of organic matter mineralization and assimilation by living organisms complete the movement of reactive nitrogen throughout the water column (Figure 1)(Stein and Klotz, 2017). All these nitrogen transformations are coupled to the marine cycling of oxygen, phosphorus, and carbon, and thus the nitrogen cycle is considered as a critical component of the biogeochemical processes in the ocean (Zehr and Kudela, 2011).

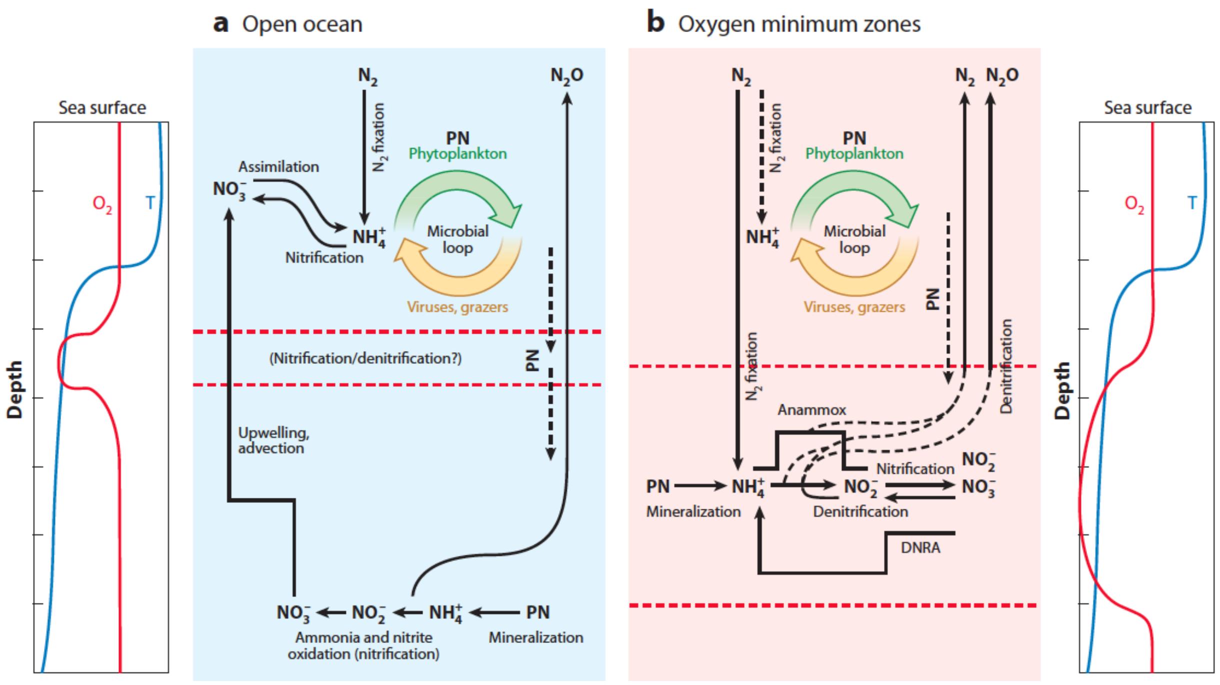


Figure 1. Conceptual diagram highlighting and comparing the major nitrogen-cycle processes in (a) the typical oceanic water column to that in (b) oxygen minimum zones. The oxidation of ammonium to nitrate is called nitrification but includes the processes of ammonia oxidation and nitrite oxidation, catalysed by different microorganisms. DNRA, dissimilatory nitrate reduction to ammonia; PN, particulate nitrogen. From Zehr and Kudela, 2011.

Due to the general unavailability of nitrogen for marine organisms, two biological processes are of fundamental importance in oceanic environments: nitrogen fixation, which is the conversion of atmospheric N_2 gas to ammonia, and denitrification, i.e. the conversion of nitrate to N_2. The balance of these two processes ultimately determines the size of the oceanic inventory of bioavailable nitrogen, and consequently, marine productivity (Capone *et al.*, 2008). In particular, biological nitrogen fixation is responsible for approximately 50% of the total nitrogen sources in the ocean (Karl *et al.*, 2002), supporting primary productivity in sites where there would be no available nitrogen otherwise, highlighting the need of understanding all aspects related to this process. Although there have been some estimates of global ocean N_2 fixation, which are in the range of 100–200 Tg N yr^{-1} (Karl *et al.*, 2002), the current methodologies used to measure nitrogen fixation, i.e., the $^{15}N_2$-tracer addition method (Montoya *et al.*, 1996; Capone and Montoya, 2001) or the acetylene reduction assay (ARA) (Hardy *et al.*, 1968), have been shown to underestimate nitrogen fixation rates (Mohr *et al.*, 2010; Grokopf *et al.*, 2012), which makes it very difficult to accurately estimate the magnitude of the bioavailable nitrogen budget. But more importantly, a major limitation for our capacity to understand this component of the nitrogen cycle relies on our current lack of knowledge on the identity and ecology of the organisms involved in N_2 fixation (Horner-Devine and Martiny, 2008; Monteiro *et al.*, 2010). There is therefore an urgent need for gaining insight into the diversity, ecology and distribution of the organisms responsible for nitrogen fixation before we can accurately understand the global significance of this process, and in order to eventually predict potential variations in nitrogen fixation rates upon changes in environmental conditions.

Diversity of nitrogen-fixing microorganisms

The biochemical process of fixing nitrogen is confined to a diverse but limited number of bacterial and archaeal lineages, the so-called diazotrophs, which are found widely, though paraphyletically, distributed across both the bacterial and archaeal domains (Figure 2) (Raymond *et al.*, 2004). The capacity of diazotrophs for nitrogen fixation relies solely on their oxygen-sensitive nitrogenase enzyme system, which, at 16 ATPs hydrolysed per N_2 fixed, carries out one of the most metabolically expensive processes in nature (Simpson and Burris, 1984). These diazotrophs contribute up to 50% of the total nitrogen sources in the global marine nitrogen budget and thus comprise the dominant source of new nitrogen to the open oligotrophic ocean (Gruber and Sarmiento, 1997; Karl *et al.*, 2002).

The first marine microorganisms identified as diazotrophs were heterotrophic bacterial isolates growing on (N)-free media (Waksman *et al.*, 1933). However, after the observation that some heterocystous cyanobacteria were diazotrophs (Fogg, 1942), most efforts were devoted to find marine representatives of such cyanobacteria, mainly in near shore ecosystems (Fogg, 1978; Stewart, 1965). The N_2 fixation performed by these microorganisms relies on the heterocysts, which are specialized cells where a micro-anaerobic environment is created to prevent the inactivation of the oxygen-sensitive nitrogenase enzymatic complex. However, with the exception of some cyanobacterial symbionts of various open ocean diatoms (Villareal, 1992), the distribution of heterocystous cyanobacteria is mostly confined to fresh and brackish waters and benthic environments (Stal, 2009).

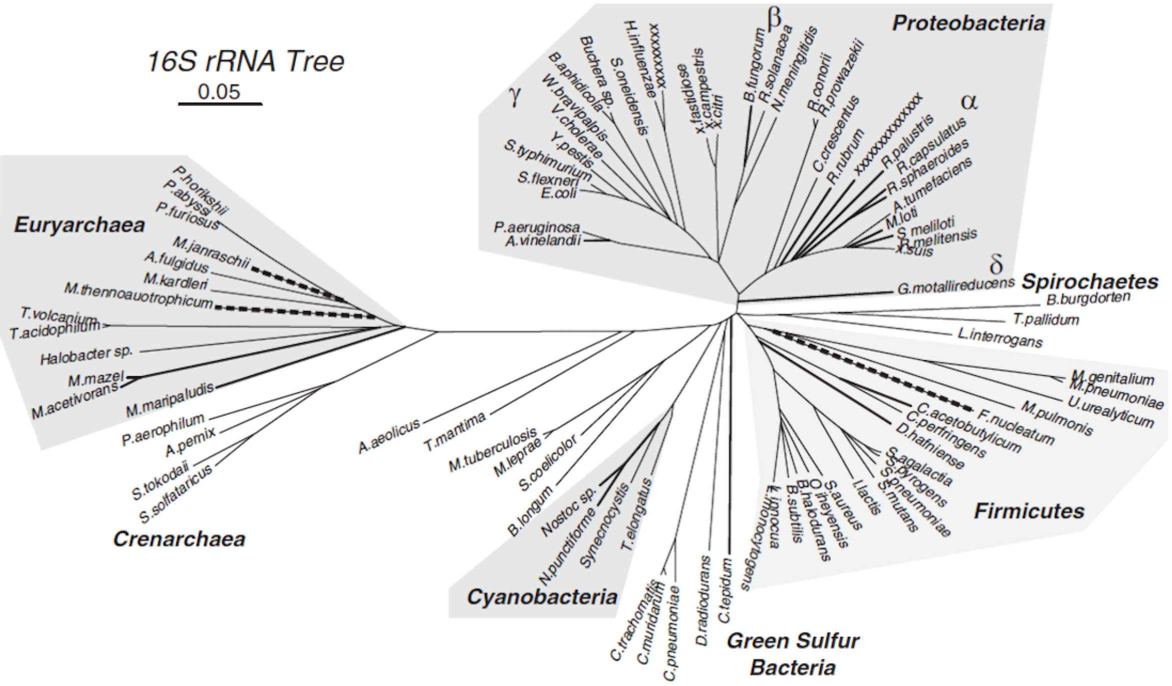


Figure 2. Phylogenetic tree showing the diversity of diazotrophs. 16S rRNA gene tree indicating position of diazotrophs within the major clades. Bold lines indicate lineages with the nif operon. Dashed lines have nif homologues. Shaded major phyla indicated those containing diazotrophs. From Raymond *et al.*, 2004.

Later, in 1961, it was discovered that the planktonic non-heterocystous cyanobacterium *Trichodesmium* was also a diazotrophic microorganism (Dugdale *et al.*, 1961). Although *Trichodesmium* had been studied for decades and its cosmopolitan distribution through the global oligotrophic ocean was well known (Carpenter, 1983), the absence of heterocystous in their trichomes and its distribution along aerobic environments made it very difficult to understand how this organism was able to fix nitrogen (Carpenter and Price, 1976; Bergman *et al.*, 2013). Indeed, with the exception of the filamentous heterocyst-forming cyanobacteria, it is actually impossible to identify N_2-fixing cyanobacteria strictly based on morphology. Therefore, this motivated the search for marker genes that allowed estimating the diversity of N_2-fixing microorganisms. The application of molecular biology approaches, specifically the sequencing of *nifH* genes amplified by the polymerase chain reaction (PCR), was the crucial step that revolutionized our knowledge of the diversity of diazotrophs.

The nitrogenase itself is an ATP-hydrolyzing, redox-active complex of two component proteins, the dinitrogenase $\alpha_2\beta_2$ heterotetramer (where α = NifD and β = NifK proteins) and the dinitrogenase reductase γ_2 homodimer (NifH protein). Initially, the sequences of the *nifH* gene as well as those of related genes (*hetR*, *nifD*, *nifK*) composing this nitrogenase enzyme system were used to build phylogenetic trees. The resulting *nifH* phylogenetic trees were very congruent with analyses based on 16S rRNA genes (Zehr *et al.*, 2003), and therefore the *nifH* gene sequences were selected as the marker gene to address the studies of biodiversity of diazotrophs. Since then, sequencing of this marker has provided a large rapidly expanding database of diazotroph sequences from diverse terrestrial and aquatic environments (Heller *et al.*, 2014).

The sequencing of *nifH* genes amplified by PCR was initially used to study nitrogenase gene sequences from the cyanobacterium *Trichodesmium thiebautii* in the Caribbean Sea (Zehr and McReynolds, 1989). However, this also resulted in the discovery of previously unknown diversity of uncultivated marine prokaryotes capable of diazotrophy, including heterotrophic (non-cyanobacterial prokaryotes) and autotrophic prokaryotes (mainly cyanobacteria) (Figure 3) (Zehr *et al.*, 1998, 2003). Currently, five major clusters can be defined based on the phylogenetic relationships established between the *nifH* diversity (Zehr *et al.*, 2003; Raymond *et al.*, 2004). Only three of them, however, clusters I, II and III, code for true functional nitrogenase (Zehr *et al.*, 2003; Raymond *et al.*, 2004). The remaining two clusters

(clusters IV and V) are formed by paralogs of *nifH* not involved in nitrogen fixation and include genes of various functions, such as some involved in photopigment biosynthesis and certain electron transport reactions (Raymond *et al.*, 2004; Staples *et al.*, 2007; Young, 2005).

Among the three functional clusters, cluster I is composed by Cyanobacteria, alpha-, beta- and gamma-Proteobacteria, and also by certain Firmicutes and Actinobacteria. This cluster I is formed by genes encoding conventional Fe-Mo nitrogenases, and mostly contains sequences retrieved by cloning and high-throughput sequencing from marine surveys (Farnelid *et al.*, 2011). Indeed, most of the known diversity of marine diazotrophs so far is contained within this cluster I. Cluster II is composed by relatively few *nifH* genes, among them those belonging to methanogenic archaea, coding for 'alternative' nitrogenases (Fe-Fe and Fe-V cofactor nitrogenases). Finally, cluster III is basically formed by obligate anaerobic bacteria and archaea, including Firmicutes, Spirochetes, sulfate-reducing delta-Proteobacteria, and methanogens.

The heterotrophic players of marine nitrogen fixation

Heterotrophic diazotrophs are believed to be active members of the bacterioplankton community. For example, several recent studies have shown that the abundance of heterotrophic diazotrophs or nitrogen fixation increase after nutrient additions (Moisander et al., 2012; Bonnet *et al.*, 2013). Moreover, other studies have found that this group of diazotrophs includes populations thriving in environments such as aphotic, N-rich oxygenated waters (Rahav *et al.*, 2013; Benavides *et al.*, 2015), oxygen-deficient waters such as the oxygen minimum zones (OMZ) (Loescher *et al.*, 2014), or the ultra-oligotrophic South Pacific Gyre (Halm *et al.*, 2012), pointing to their widespread distribution.

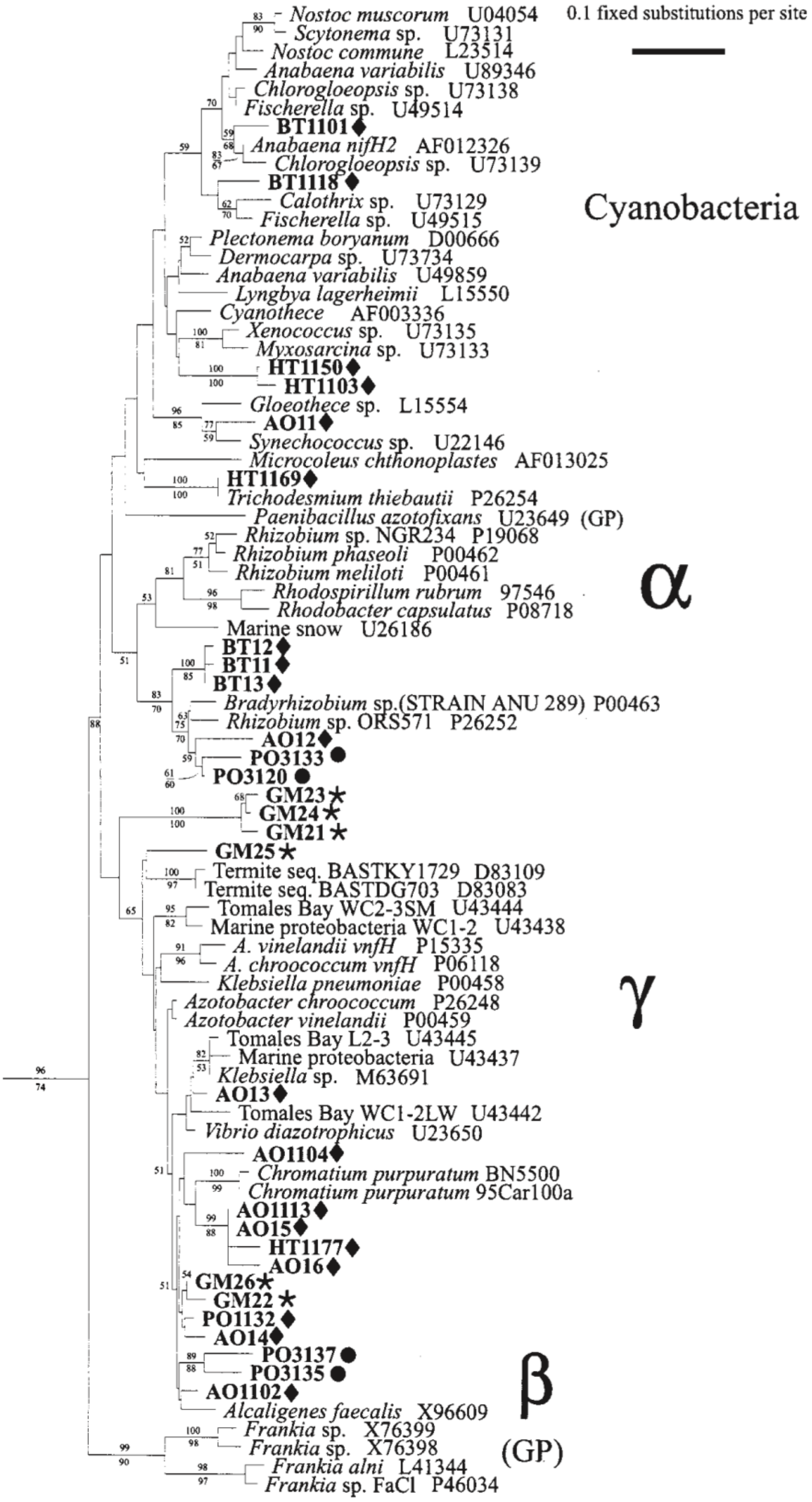

0.1 fixed substitutions per site
Nostoc muscorum U04054
Scytonema sp. U73131
Nostoc commune L23514
Anabaena variabilis U89346
Chlorogloeopsis sp. U73138
Fischerella sp. U49514
BT1101◆
Anabaena nifH2 AF012326
Chlorogloeopsis sp. U73139
BT1118◆
Calothrix sp. U73129
Fischerella sp. U49515
Plectonema boryanum D00666
Dermocarpa sp. U73734
Anabaena variabilis U49859
Lyngbya lagerheimii L15550
Cyanothece AF003336
Xenococcus sp. U73135
Myxosarcina sp. U73133
HT1150◆
HT1103◆
Gloeothece sp. L15554
AO11◆
Synechococcus sp. U22146
Microcoleus chthonoplastes AF013025
HT1169◆
Trichodesmium thiebautii P26254
Paenibacillus azotofixans U23649 (GP)
Rhizobium sp. NGR234 P19068
Rhizobium phaseoli P00462
Rhizobium meliloti P00461
Rhodospirillum rubrum 97546
Rhodobacter capsulatus P08718
Marine snow U26186
BT12◆
BT11◆
BT13◆
Bradyrhizobium sp.(STRAIN ANU 289) P00463
Rhizobium sp. ORS571 P26252
AO12◆
PO3133 ●
PO3120 ●
GM23★
GM24★
GM21★
GM25★
Termite seq. BASTKY1729 D83109
Termite seq. BASTDG703 D83083
Tomales Bay WC2-3SM U43444
Marine proteobacteria WC1-2 U43438
A. vinelandii vnfH P15335
A. chroococcum vnfH P06118
Klebsiella pneumoniae P00458
Azotobacter chroococcum P26248
Azotobacter vinelandii P00459
Tomales Bay L2-3 U43445
Marine proteobacteria U43437
Klebsiella sp. M63691
AO13◆
Tomales Bay WC1-2LW U43442
Vibrio diazotrophicus U23650
AO1104◆
Chromatium purpuratum BN5500
Chromatium purpuratum 95Car100a
AO1113◆
AO15◆
HT1177◆
AO16◆
GM26★
GM22★
PO1132◆
AO14◆
PO3137●
PO3135●
AO1102◆
Alcaligenes faecalis X96609
Frankia sp. X76399
Frankia sp. X76398
Frankia alni L41344
Frankia sp. FaCl P46034
Cyanobacteria
α
γ
β
(GP)

Figure 3. Phylogenetic analysis of *nifH* genes obtained from oceanic picoplankton and zooplankton in 1998. The phylogeny includes *nifH* sequences from cultivated organisms (genus and species) and uncultivated organims (e.g, from termites and from marine [e.g., Tomales Bay] samples). GenBank sequence numbers are indicated). The data set was bootstrapped 100 times, and bootstrap values greater than 50% are indicated at the relevant methods (distance and parsimony methods are represented by values above and below the nodes, respectively). GP, gram positive; , diatom associated; ◆, picoplankton; ★, zooplankton associated. From Zehr *et al.*, 1998.

Within the heterotrophic diazotrophs, the most studied case corresponds to the phylotype Gamma-A (Bird *et al.*, 2005). Although Gamma-A have been found in all major ocean basins along the water column up to mesopelagic depths, the highest abundances have been observed within the warm, oligotrophic, fully oxygenated surface waters of tropical and subtropical latitudes (Langlois *et al.*, 2015). In particular, in the Eastern Tropical South Pacific (ETSP), the *nifH* sequences detected belonged to heterotrophic groups within Clusters I and III, whereas almost no cyanobacterial phylotypes were found (Bonnet *et al.*, 2013; Turk-Kubo *et al.*, 2014; Loescher *et al.*, 2014), and detectable rates of nitrogen fixation have been detected in ETSP area throughout the water column (Fernandez *et al.*, 2011; Bonnet *et al.*, 2013; Loescher *et al.*, 2014). However, it has not been possible to demonstrate a direct link between nitrogen fixation rates and heterotrophic diazotroph phylotypes in this particular area (Turk-Kubo *et al.*, 2014). Therefore, despite the huge diversity and widespread distribution reported for heterotrophic diazotrophs, which have been detected in all the major oceanic regions (Bombar *et al.*, 2016), there is still not enough evidence supporting a significant role of heterotrophic diazotrophs in marine nitrogen fixation (Turk-Kubo *et al.*, 2014).

In terms of abundance contribution to the bacterioplankton community, phylotypes associated with Gamma A populations have shown maximal abundances of 10^4 *nifH* gene copies per liter, likewise several other heterotrophic diazotrophic phylotypes (Moisander *et al.*, 2014; Shiozaki *et al.*, 2014), with only sporadic exceptions of higher abundances that can reach up to 10^7 *nifH* gene copies per liter (Halm *et al.*, 2011; Church, Jenkins, *et al.*, 2005; Zhang *et al.*, 2011). However, even at such 'high' abundances, these diazotrophs only account for up to about 1% of the local bacterial community. In any case, low abundances of a particular species do not necessarily mean a low contribution or significance of these species in the functioning of the ecosystem, since rare populations can act as keystone microbial species with disproportionately large effects on ecosystem services (Giovannoni and Stingl, 2005). In general, however, the commonly found low abundance of marine heterotrophic diazotrophs, together with a supposed lack of pigment fluorescence, makes their study through techniques such as flow cytometry cell sorting coupled with metagenomics or fluorescence in situ hybridization (FISH) much more difficult than in the case cyanobacterial diazotrophs. In consequence, the significance of heterotrophic diazotrophs for oceanic nitrogen fixation remains poorly understood, even in regions where they dominate the diazotrophic community, and further efforts are need to better understand the ecology of these diazotrophs.

Cyanobacterial diazotrophs in the ocean

Unlike their heterotrophic counterparts, the different cyanobacterial diazotrophs inhabiting the open ocean are much more diverse in terms of morphology (they can be coccoid, filamentous or can develop specialized cells, i.e., heterocysts). Some cyanobacterial diazotrophs are free-living, but they can also occur associated with particles or with other organisms in a symbiosis. They are distributed widely throughout tropical and subtropical waters (Figure 4) (Zehr, 2011) and rarely found in high polar latitudes (Blais *et al.*, 2012; Díez *et al.*, 2012; Shiozaki *et al.*, 2017). Cyanobacterial diazotrophs can be classified into four major groups differing in their morphologic characteristics and main lifestyles: (1) the free-living filamentous non-heterocyst-forming *Trichodesmium*, (2) the filamentous heterocyst-forming symbionts with unicellular eukaryotic algae like for instance *Richelia* or *Calothrix*, (3) free-living unicellular cyanobacteria such as UCYN-B, (4) and finally unicellular symbiotic cyanobacteria such as *Candidatus* Atelocyanobacterium thalassa (UCYN-A) (Figure 5) (Thompson and Zehr, 2013).

Free-living filamentous non-heterocyst-forming Trichodesmium

Trichodesmium spp. had initially been assumed to represent the most important nitrogen-fixing cyanobacterium in the ocean. However, this ‘ranking’ is currently under debate because a comparable importance has been attributed to unicellular nitrogen-fixing cyanobacteria (Martínez-Pérez *et al.*, 2016). *Trichodesmium* representatives have consistently been shown to be stable components of tropical and subtropical segments of the Atlantic, Pacific, and Indian Oceans (Figure 4a). Genetic and morphological markers argue that the different *Trichodesmium* strains form only two major distinct clades, and it has been suggested that they display different ecological distributions (Orcutt *et al.*, 2002; Hynes *et al.*, 2012; Bergman *et al.*, 2013). An important characteristic of *Trichodesmium* spp. is that they can control buoyancy and their depth in the water column through its ability to generate gas vacuoles and, consequently, allowing a vertical migration to optimize light intensity or to obtain nutrients from deeper waters (Villareal and Carpenter, 2003). They are well-known contributors to sustaining marine life via active release of key nutrients, hence making this fully photoautotrophic genus a vital player in the biogeochemical cycling in the oceans (Carpenter and Capone, 2008).

Filamentous heterocyst-forming symbionts

Heterocyst-forming filamentous cyanobacteria are mainly found in estuaries and in the Baltic Sea, but they are not extremely abundant in oligotrophic oceans for several reasons. For example, they are known to be sensitive to turbulence (Howurth *et al.*, 1993), and nitrogen fixation by heterocysts has been suggested to be limited by oxygen concentration and flux into the cell in tropical areas (Stal, 2009; Staal *et al.*, 2007), being outcompeted by non-heterocyst-forming diazotrophic cyanobacteria. However, the heterocyst might be of advantage in the microenvironment of a photosynthetic symbiotic partner cell, since heterocyst-forming cyanobacterial symbionts are commonly found associated with several oceanic diatom genera (Figure 4c) (Carpenter, 2002; Foster *et al.*, 2009). Among these, the heterocyst-forming cyanobacteria of the genera *Richelia* and *Calothrix* are symbiotically associated with different genera of diatoms, including *Rhizosolenia* and *Hemiaulus* (living within the diatom frustule but outside the cell wall), and *Chaetoceros* (living externally associated to the frustule) (Zehr, 2011), all diatoms that can form blooms or occur in transient high concentrations after river discharge, mesoscale features, or mixing (Dore *et al.*, 2008; Fong *et al.*, 2008). Their rapid sink after blooming and the subsequently export of carbon to the deep ocean (Scharek *et al.*, 1999; Subramaniam *et al.*, 2008) makes these heterocyst-forming symbiotic cyanobacteria some of the most ecologically important diazotrophic cyanobacteria in the oceans.

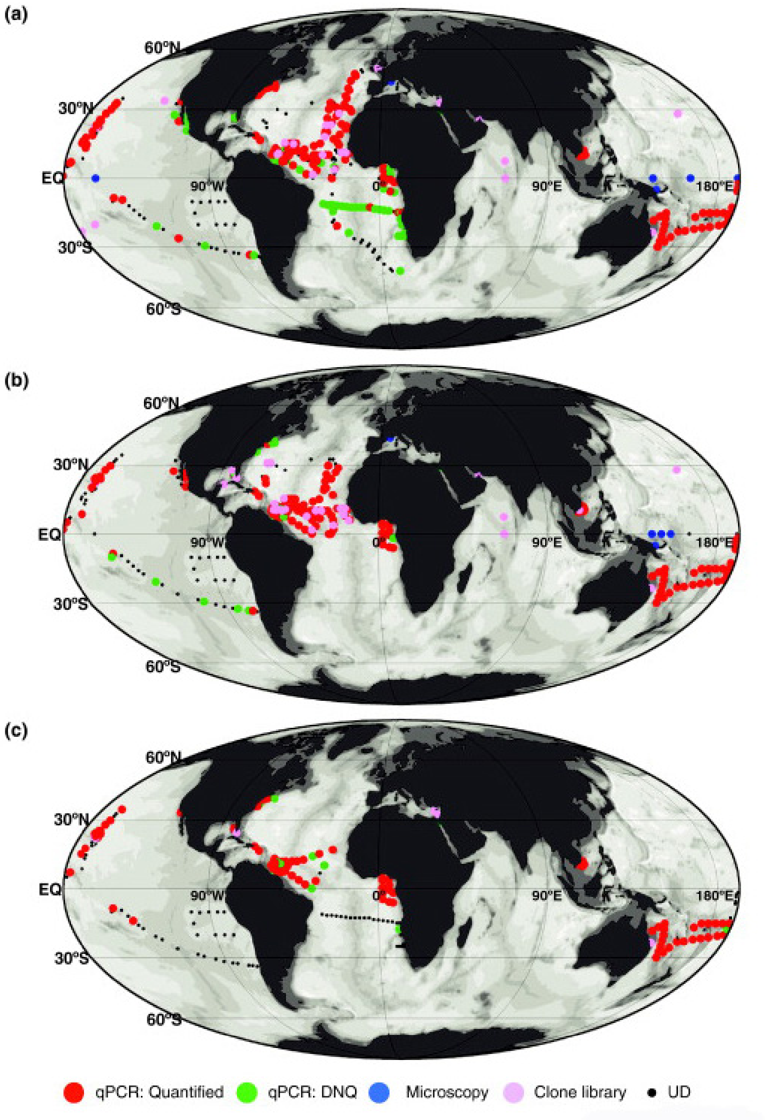


Figure 4. Distribution of major groups of oceanic N_2-fixing cyanobacteria targeted by quantitative PCR assays. The occurrence of these phylotypes in clone libraries and FISH-based observations of the UCYN clades are also included. (a) *Trichodesmium*, (b) unicellular N_2-fixing cyanobacteria, (c) heterocyst-forming cyanobacteria symbionts of diatoms. Colours: blue, cyanobacteria were observed by microscopy; green, detection by qPCR but at low concentrations and were not quantifiable (DNQ); red, cyanobacteria were quantified by qPCR; pink, detected by PCR amplification, cloning and sequencing; black, samples collected for qPCR, but cyanobacteria were not detected (UD). Abbreviation: EQ, equator. From Zehr, 2011.

Free-living unicellular cyanobacteria

The third group of cyanobaterial diazotrophs is formed by the unicellular free-living (although they can also interact indirectly with other microbes) clades called UCYN-B and UCYN-C (Zehr, 2011; Taniuchi *et al.*, 2012). UCYN-B, or *Crocosphaera* spp., although originally isolated from the South Atlantic Ocean, have been successfully isolated from multiple ocean basins and found widely distributed in tropical and subtropical oceans (Figure 5, 4b). These cyanobacteria can be easily observed by microscopy because they contain the photosynthetic pigment phycoerythrin, and were detected for the first time in very oligotrophic waters of the central southern tropical Pacific by fluorescence activated cell sorting (FACS)(Neveux *et al.*, 1999). Differently to *Trichodesmium*, to the heterocyst-forming *Richelia* and *Calothrix* cyanobacteria or to UCYN-A, *Crocosphaera* fixes the nitrogen during the night hours. *Crocosphaera* strains exhibit phenotypic differences, but genetic comparisons have found high sequence conservation among cultivated strains and environmental sequences. In this context, it seems that *Crocosphaera* strains diverge and maintain genetic diversity through genetic rearrangements and by incorporating strain-specific sequences (Zehr *et al.*, 2007; Bench *et al.*, 2011). In addition to its free-living state, *Crocosphaera*-like cells can be associated with diatoms (Carpenter and Janson, 2000; Foster *et al.*, 2011), or with aggregates (Sohm, Webb, *et al.*, 2011; Foster *et al.*, 2013). Cultivated strains can be divided into two broad phenotypic and genomic categories: (i) those that have larger cell diameters (>4 µm) and larger genomes characterized by some genetic redundancy and potentially increased adaptations to iron and phosphorous limitations,and (ii) those that show the opposite pattern, i.e. smaller cell diameters(<4 µm) and smaller genomes, and a relative loss of genetic capabilities (Webb *et al.*, 2009; Sohm, Edwards, *et al.*, 2011; Bench *et al.*, 2013). These two *Crocosphaera* categories have shown to present distinct distributions in the North and South Pacific Ocean (Bench *et al.*, 2016), which supports the adaptation of *Crocosphaera* to diverse oceanic habitats, with presumably different contribution to nitrogen fixation.

The unicellular *Cyanothece,* or UCYN-C group, also belongs to the lineages of unicellular diazotrophic bacteria with representatives in the open ocean, mainly in tropical oceans, but they are much less abundant that UCYN-A and UCYN-B (Falcon *et al.*, 2002; Foster *et al.*, 2007; Langlois *et al.*, 2008). As it happens with *Crocosphaera*, *Cyanothece* fixes the nitrogen during the night hours separating nitrogen fixation temporally from photosynthesis (Sherman *et al.*, 1998). In general, although few, the available estimates of abundances and geographic distribution of the free-living unicellular nitrogen-fixers (particularly those of *Crocosphaera*) suggest that these microorganisms are also significant contributors to global nitrogen fixation.

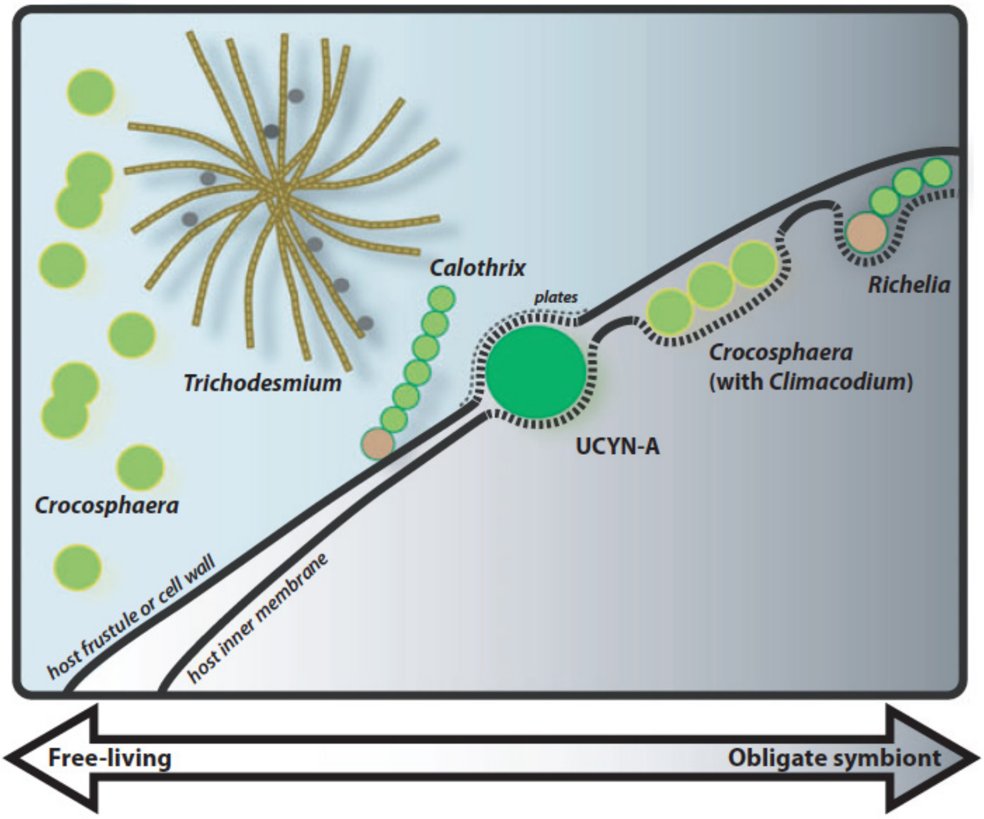


Figure 5. The spectrum of cellular interactions engaged in by marine diazotrophic cyanobacteria. Dashed lines show features that are uncertain, such us the location of symbionts relative to host inner or outer membranes and the presence of calcareous plates on the *Candidatus* Atelocyanobacterium thalassa (UCYN-A) host. From left to right, other features include: free-living *Crocosphaera* cells (double cells are dividing); *Trichodesmium* with associated microbiota (gray); *Calothrix* with terminal heterocyst (brown) and vegetative cells (green); UCYN-A in relation to its host membranes; colonial aggregates of *Crocosphaera*-like cells in association with a diatom; and *Richelia* with terminal heterocyst (brown) and vegetative cells (green). From Thompson and Zehr, 2013.

Unicellular symbiotic cyanobacteria: The case of Candidatus Atelocyanobacterium thalassa (UCYN-A)

The discovery of UCYN-A, formally *C.* Atelocyanobacterium thalassa, is completely rooted in the molecular era, and was based on a short *nifH* gene sequence in 1998 (Zehr *et al.*, 1998). Since then, much knowledge has been gained about this organism, and now it is well known that UCYN-A is a widely distributed key nitrogen-fixer in the ocean, at least as important as *Trichodesmium* (Zehr *et al.*, 2016). A wide variety of cultivation-independent approaches, including PCR-based techniques, fluorescence *in situ* hybridization (FISH), genome sequencing coupled to flow cytometry, nanoscale secondary ion mass spectrometry (nanoSIMS), metagenomics and metatranscriptomics have been employed to gain insight into the distribution and ecology of UCYN-A, confirming its global importance in the nitrogen cycle (Figure 6) (Zehr *et al.*, 2016).

Two key molecular approaches have been crucial to detect UCYN-A in a large number of ocean basins: the PCR or quantitative PCR (qPCR) of the *nifH* gene and 16S rRNA gene (Thompson and Zehr, 2013; Zehr *et al.*, 2016), and the visual quantification of their abundances through the FISH assay (Biegala and Raimbault, 2008; Bonnet *et al.*, 2009; Le Moal and Biegala, 2009; Krupke *et al.*, 2013). The more it was learned about this diazotroph, however, the more intriguing it seemed. For example, the first attempts to evaluate the abundance and the patterns of nitrogen fixation by quantitative reverse-transcriptase PCR (qRT-PCR) in UCYN-A gave enigmatic results: Generally, diazotrophic cyanobacteria separate photosynthesis and nitrogen fixation either spatially or temporally by fixing nitrogen at night, when the oxygen-evolving photosystem II (PSII) apparatus is not active and therefore the inactivation of the nitrogenase complex can be avoided (Fay, 1992; Berman-Frank *et al.*, 2003). However, contrary to the pattern shown by UCYN-B with the expected night-time nitrogen fixation for unicellular cyanobacteria, UCYN-A was shown to display the highest level of *nifH* transcriptional activity during the light-hours (Church, Short, *et al.*, 2005). This controversial nitrogen fixation pattern was clarified thanks to the sorting of an environmental UCYN-A population by flow cytometry and its subsequent whole-genome amplification and high-throughput DNA sequencing. By these means, it was shown that UCYN-A lacked all the genes for PSII and, consequently, did not evolve O_2, allowing the expression of the nitrogenase complex during the light hours (Zehr *et al.*, 2008). Moreover, the UCYN-A genome was shown to be so reduced that it was missing cyanobacterial defining features such us the RuBisCO (ribulose-1,5-bisphosphate carboxylase/oxygenase) for carbon fixation or the entire tricarboxylic acid (TCA) cycle among a variety of other metabolic pathways (Tripp *et al.*, 2010).

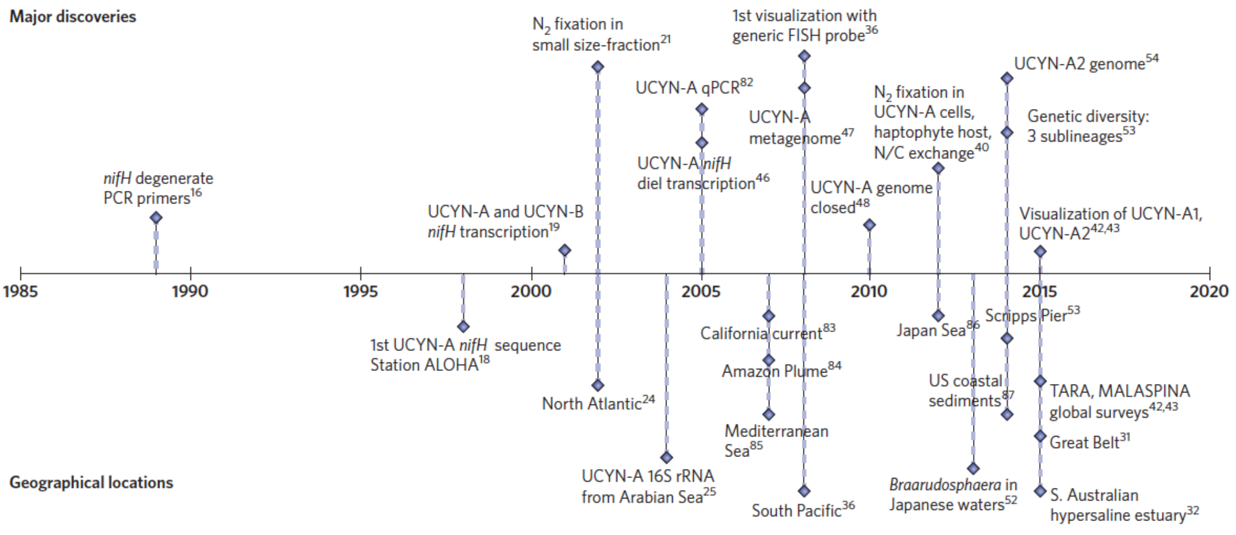

Major discoveries
nifH degenerate PCR primers[16]
UCYN-A and UCYN-B *nifH* transcription[19]
N_2 fixation in small size-fraction[21]
UCYN-A qPCR[82]
UCYN-A *nifH* diel transcription[46]
1st visualization with generic FISH probe[36]
UCYN-A metagenome[47]
UCYN-A genome closed[48]
N_2 fixation in UCYN-A cells, haptophyte host, N/C exchange[40]
UCYN-A2 genome[54]
Genetic diversity: 3 sublineages[53]
Visualization of UCYN-A1, UCYN-A2[42,43]
1985
1990
1995
2000
2005
2010
2015
2020
1st UCYN-A *nifH* sequence Station ALOHA[18]
North Atlantic[24]
UCYN-A 16S rRNA from Arabian Sea[25]
California current[83]
Amazon Plume[84]
Mediterranean Sea[85]
South Pacific[36]
Japan Sea[86]
US coastal sediments[87]
Braarudosphaera in Japanese waters[52]
Scripps Pier[53]
TARA, MALASPINA global surveys[42,43]
Great Belt[31]
S. Australian hypersaline estuary[32]
Geographical locations

Figure 6. A timeline detailing major publications leading from the discovery of UCYN-A to genome sequencing and visualization of cells, and the detection in different regions of the world`s oceans. Superscript numbers indicate references in Zehr et al, 2016. References numbers 42 and 43 are contributions derived from this thesis. From Zehr et al, 2016.

These unexpected findings raised the question of how UCYN-A could thrive in oligotrophic environments lacking such important biosynthetic pathways. Although Tripp *et al.,* (2010) had proposed that the reason might be a symbiotic lifestyle, Thompson and Foster et al. were the ones answering this question in 2012, unveiling the symbiotic association of UCYN-A with a prymnesiophyte single-celled alga closely related to *Braarudopshaera bigelowii*. This was possible because availability of the UCYN-A genome allowed to design a UCYN-A-specific 16S rRNA gene FISH probe that permitted to visualize this symbiosis for the first time (Figure 7) (Thompson *et al.*, 2012). In this study, they also showed, by means of stable isotope experiments and nanoscale secondary ion mass spectrometry (nanoSIMS), that UCYN-A fixed N_2, and rapidly exchanged N with the prymnesiophyte partner. In exchange, the fixed carbon through the photosynthetic activity of the prymnesiophyte alga was transferred to UCYN-A, establishing the mutualistic basis of this symbiosis (Figure 7) (Thompson *et al.*, 2012).

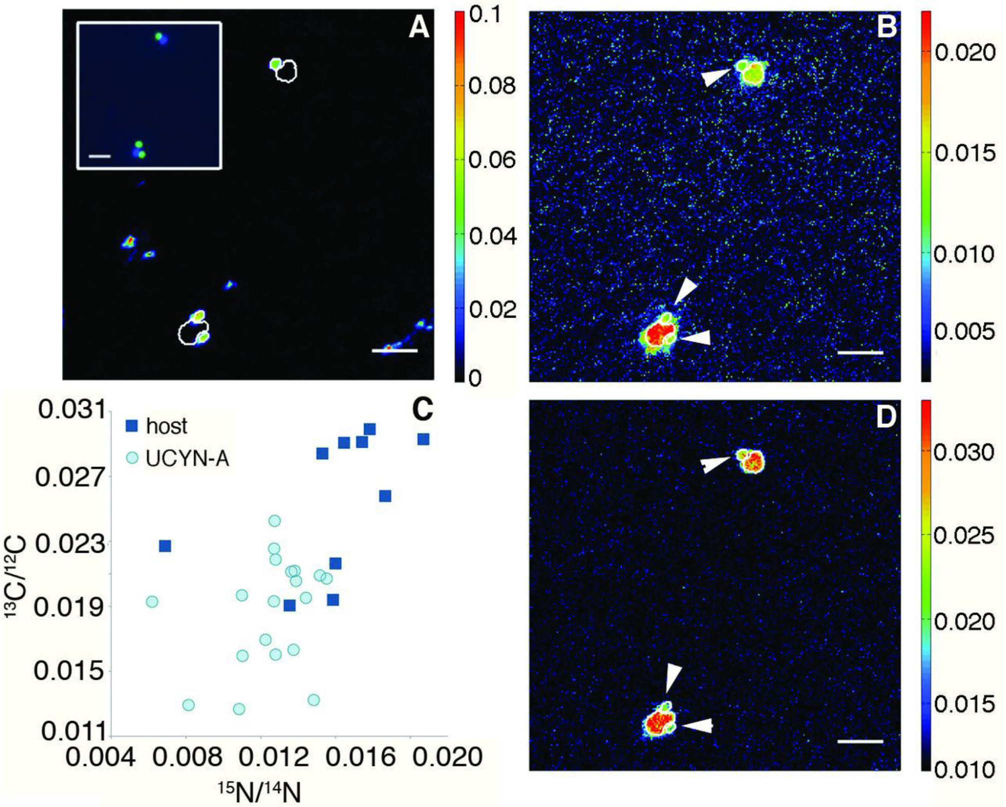


Figure 7. Microscopy and elemental composition of two UCYN-A partner cells and their associated UCYN-A cells detected in samples from sorted picoeukaryotes analysed by HISH-SIMS. (**A**) $^{19}F/^{12}C$ (HISH) labelling of UCYN-A. Inset displays labelling of the same UCYN-A cells by catalysed reporter deposition–fluorescence in situ hybridization (green) and DAPI (4′,6-diamidino-2-phenylindole) staining of partner cell nucleus (blue). (**B**) The $^{13}C/^{12}C$ ratio image of UCYN-A and partner cell. (**C**) The $^{13}C/^{12}C$ and $^{15}N/^{14}N$ in 10 selected partner cells and their associated UCYN-A cells (table S6). (**D**) The $^{15}N/^{14}N$ image ratio of UCYN-A and partner cell. The white lines define regions of interest that were used for calculating $^{13}C/^{12}C$ and $^{15}N/^{14}N$ ratios. UCYN-A cells are indicated by white arrows in (B) and (D). Scale bar, 3 µm. From Thompson and Foster et al, 2012.

Recently, the genome of a second UCYN-A genome lineage, called UCYN-A2 was also obtained (Bombar *et al.*, 2014). This new lineage was detected off Scripps Pier in San Diego, California, and allowed to open new questions about the UCYN-A symbiotic system. The two genomes of UCYN-A1 and UCYN-A2 lineages were extremely similar, lacking the same metabolic genes, which suggested a similar symbiotic strategy (Bombar *et al.*, 2014). The host of UCYN-A2 was later identified by the 18S rRNA gene sequence, and was closely related to the UCYN-A1 host but closer to a prymnesiophyte strain detected in Japanese coastal waters (Hagino *et al.*, 2013; Thompson *et al.*, 2014). It is now clear that there are several distinct hosts and UCYN-A lineages, although the specificity of these associations or whether more unexplored diversity is involved in this symbiotic system remains to be resolved. It is also unknown whether the number of symbiont cells per host is constant or whether it is specific to each lineage, as well as the consequences on the physiology of the host. Therefore, due to its global importance in the marine nitrogen cycle and the particularity of the system, the deep exploration of these symbionts at a global scale can shed light into the ecology and evolution of this remarkable symbiotic system. In addition, because of the particular symbiotic life-style of UCYN-A, this organism has been considered a possible model for understanding organelle evolution (Zehr *et al.*, 2016), which highlights the relevance of gaining insight into the evolutionary and ecological processes leading to the establishment of these unusual relationships.

Who contributes more to nitrogen fixation in the ocean?

Cyanobacterial diazotrophs have long been considered the most important diazotrophs in warm, oligotrophic, surface ocean waters (Zehr, 2011; Karl *et al.*, 2002), but this paradigm is not so clear now. Heterotrophic diazotrophs appear to be almost ubiquitous in marine waters, even in high-latitude, deep, cold, or coastal waters where cyanobacteria, with some exceptions, are few or absent (Fernandez *et al.*, 2011; Blais *et al.*, 2012; Díez *et al.*, 2012; Bonnet *et al.*, 2013; Farnelid *et al.*, 2013; Bombar *et al.*, 2017). A recent global analysis done using *nifH* PCR amplicon sequences from samples taken around the surface ocean unveiled that phylotypes of heterotrophic diazotrophs, particularly proteobacterial phylotypes from cluster I, dominated the diazotrophic community (Figure 8) (Farnelid *et al.*, 2011). However, two important points have to be taken into account about this particular study; first, important filamentous diazotrophic cyanobacteria such as *Trichodesmium* were excluded from the analysis because samples were pre-filtered by 10 µm size-pore filters and thus their contribution to the diazotrophic community could not be assessed; and second, it was shown that the relative contribution of *nifH* transcripts, which represent the expressed *nifH* genes, was higher in the case of cyanobacterial diazotrophs than in the case of heterotrophic diazotrophs, suggesting that in terms of activity the cyanobacteria might be more important contributors to nitrogen fixation.

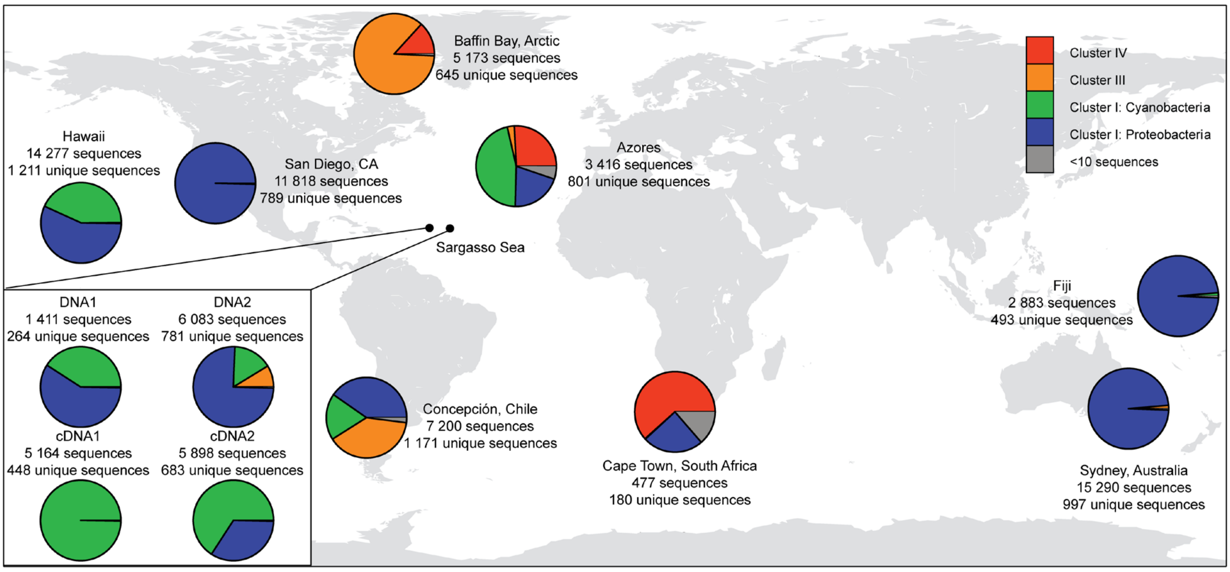
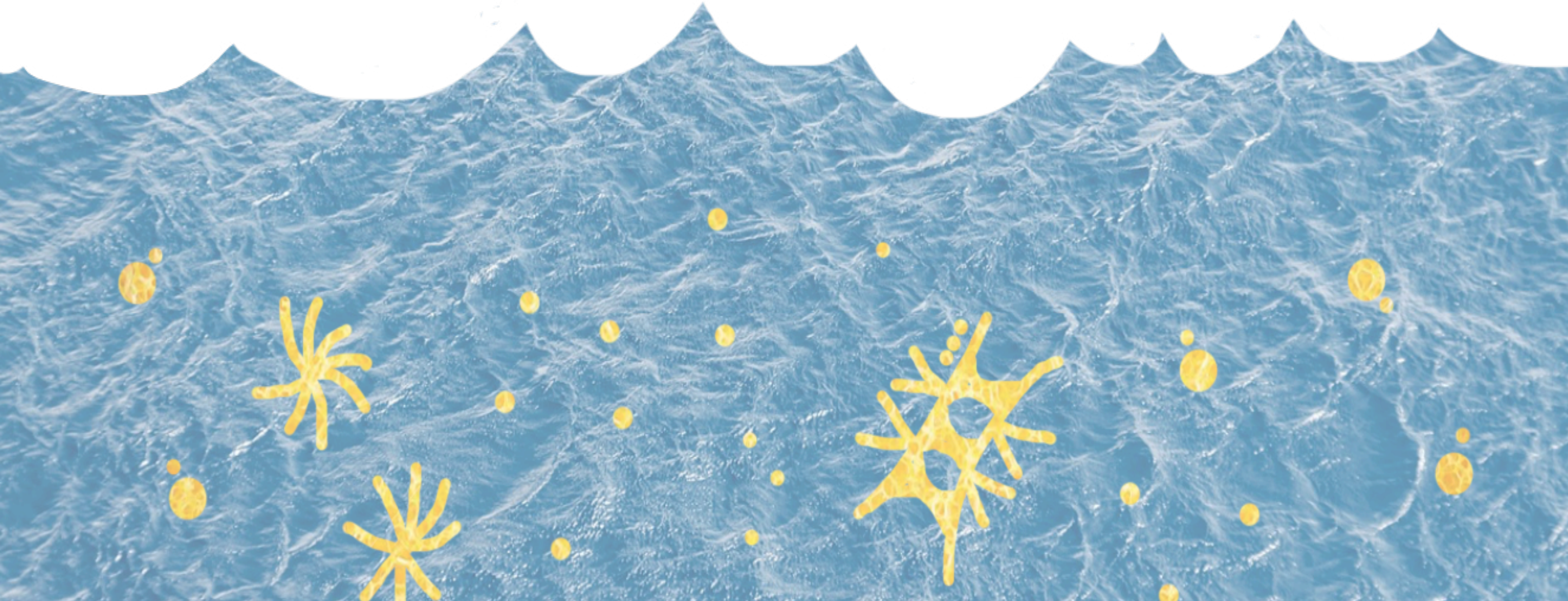


Figure 8. World map of sampling locations showing the distribution of *nifH* Clusters. Pie charts display the distribution of *nifH* Clusters within each sample. Clusters containing <10 sequences (shown in grey) were not phylogenetically designated. Note that Cluster I is split into *Proteobacteria* and Cyanobacteria, but that Cluster III also contains some *Proteobacteria*. For the Sargasso Sea samples, which were prefiltered (10 µm) to avoid filamentous cyanobacteria, the pie charts for DNA and cDNA samples are shown in the bottom left corner. From Farnelid et al. 2011.

Another important aspect to consider is that studies using PCR-based techniques hinder the comparison between samples in terms of the contribution of diazotrophs to total microbial community because the diversity obtained from the amplified gene (or gene fragment) only considers the diazotrophic community but does not allow knowing which fraction of the total prokaryotic community they represent. Furthermore, it is well known that PCR-based approaches can impact the diversity and relative abundance of the amplified gene (Acinas *et al.*, 2005). Such biases have been shown to either exclude the discovery of certain diazotrophic lineages (Bürgmann *et al.*, 2004), or to alter the ratios of sequence abundance in the products relative to the templates of PCR (Suzuki and Giovannoni, 1996; Sipos *et al.*, 2007). It is thus possible that the use of PCR-based techniques for the detection of the diazotrophic communities is giving a biased view of the actual contribution of these organisms within prokaryotic communities, which thus precludes linking changes in diazotrophic diversity to the measured nitrogen fixation rates. For example, a recent study comparing nitrogen fixation rates with the diazotrophic diversity (assessed via PCR-based *nifH* sequencing) in different regions found that although the diazotrophs detected were nearly omnipresent in marine waters, the nitrogen fixation process was regionally restricted, i.e., occurred at very few stations (Gradoville *et al.*, 2017). One possible explanation is that the aforementioned PCR biases cause the detection of extremely rare diazotrophic taxa that are not actually contributing to the process because their abundances are negligible at the community level, or that nitrogen fixation occurs only under particular conditions and the detection of diazotrophic taxa does not imply that they are active. Therefore, there is not yet a clear answer concerning the relative contribution of different phylogenetic groups of diazotrophs in the marine environment, and there is a need for conducting alternative quantitative studies (such as metagenomic approaches) to have a more accurate description of the actual distribution of different diazotrophic species.

Aims of the Thesis



Aims and outline of the thesis

The main goal of this thesis is **to gain insight into the diversity, ecology and evolution of the marine nitrogen-fixing microorganisms in the open ocean**. The contribution of this thesis to the knowledge on this topic has been structured in three chapters. In the first chapter, we analyzed the metagenomic dataset from the global circumnavigation expedition *Tara* Oceans in order to explore the biogeography of different groups of nitrogen-fixing bacteria across oceanic basins and throughout the water column, from surface to mesopelagic layers. The recruitment of *nifH* gene sequences from the metagenomic dataset to explore the diversity of diazotrophs avoids the primer bias of the PCR-based approaches used in most studies. By this means we could not only to uncover new nitrogen-fixing groups but also but also quantify the relative abundance of diazotrophs within bacterioplankton communities (**Chapter 1**). The results obtained in the first chapter identified and confirmed the presence of the uncultivated symbiotic cyanobacteria UCYN-A as one of the key nitrogen-fixing players in the ocean. Consequently, the next two chapters were devoted to explore in detail the UCYN-A diversity, ecology and evolution as well as its relationship with its host. This was carried out through the combination of different visualization techniques (CARD-FISH) and metagenome and metatranscriptome analyses (**Chapter 2**). In the last chapter, we obtained different views of the UCYN-A lineage composition in particular environmental samples depending on the approach used for the identification (PCR, qPCR, CARD-FISH, metagenomes), which led to new information on the diversity of the UCYN-A symbiosis. We uncovered a new UCYN-A genomic species and unveiled that the emerging novel diversity of the UCYN-A group is distributed along different size fractions of the plankton, which places the nitrogen fixation in novel planktonic compartments (**Chapter 3**).

The thesis outline can be further organized in the following two major objectives that can each be subdivided into several specific ones:

Objective 1: *To assess the contribution of diazotrophs to the total microbial community structure in the global ocean*

Understanding the contribution of heterotrophic versus cyanobacterial diazotrophs to total marine microbial nitrogen fixers has been rarely addressed and a consensus is still far to be reached because opposite results have been obtained to date.

In **Chapter 1** a global exploration of the diazotrophic community was carried out for the first time using a non-PCR biased approach. It consisted in the extraction of the *nifH* genes from the metagenomic data from the *Tara* Oceans expedition with the following aims:

- To obtain a comprehensive view of the contribution of diazotrophs across ocean basins.

- To assess the contribution of different autotrophic and heterotrophic diazotrophic groups across basins and along the water column.

- To uncover novel diazotrophic groups.

Objective 2: *To gain insight into the diversity, ecology and evolution of the uncultured nitrogen-fixing cyanobacteria UCYN-A*

UCYN-A is one of the major players in the marine nitrogen fixation process and lives in symbiosis with a single-celled prymnesiophyte. Different lineages of both symbionts and hosts have been described. However, the few environmental metagenomic sequences available to date, together with the absence of lineage-specific probes for visualizing the symbiosis under microscopy make difficult the description of the UCYN-A symbiotic system.

In **Chapter 2,** using samples from two stations of the *Tara* Oceans expedition where this symbiosis was significantly abundant, as well as a combination of microscopy and genomic methods, the next specific objectives were addressed:

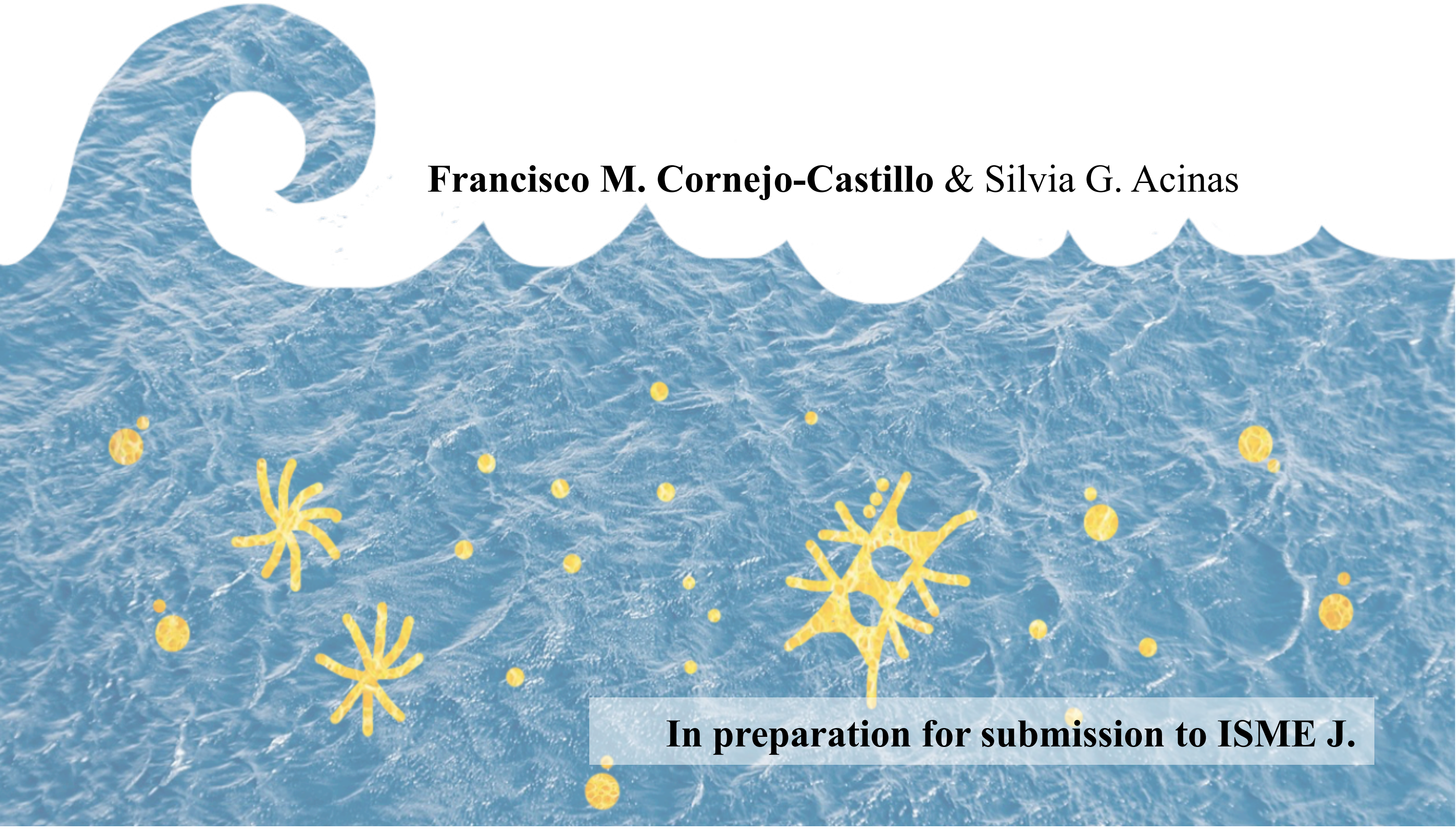
- To design CARD-FISH probes to explore microscopically the symbiotic association involving UCYN-A and prymnesiophytes.
- To explore whether different UCYN-A genomic species are in symbiosis with different hosts.
- To decipher whether distinct lineages, in association with distinct partners, exhibit different expression patterns.
- To evaluate the evolutionary forces acting on UCYN-A.

In **Chapter 3,** we focused on a deep exploration of the diversity of the UCYN-A clade with the following objectives:

- To evaluate the specificity of the methods applied in the study of the UCYN-A symbiosis.
- To design a strategy to explore new divergent genomes of UCYN-A in environmental samples.
- To explore new UCYN-A diversity in different fractions of the plankton.

Chapter 1

A metagenomic exploration of the biogeography of nitrogen-fixing microorganisms unveils novel diversity across the global ocean



Francisco M. Cornejo-Castillo & Silvia G. Acinas

In preparation for submission to ISME J.

Abstract

Nitrogen-fixing microorganisms (diazotrophs) are key members of marine ecosystems because they act as suppliers of bioavailable nitrogen for primary producers in the oceans. Although numerous studies have addressed the diversity, abundance and distribution of diazotrophs in marine systems, a consensus does not exist concerning the relative importance of different groups to marine nitrogen fixation. Since most of the previous studies are founded on the amplification of the *nifH* marker gene (PCR, qPCR), the lack of consensus could come from the methodological biases associated with the use of primers. Here we explored the diversity and abundance of marine diazotrophic microorganisms across the global ocean using the *Tara* Ocean's metagenomic dataset, in an approach that does not rely on primers. Twenty-eight *nifH* gene sequences were designated using a 95% nucleotide sequence similarity cut-off from the Ocean Microbial Reference Gene Catalog (OM-RGC), mostly non-cyanobacterial variants. Among them, 18 non-cyanobacterial gene sequences presented mismatches with the 'nifH4' primer commonly used in marine diazotrophic diversity studies, and we proposed a modification of this primer. In general, diazotrophs were found in the rare biosphere (<0.1% relative abundance) with significant higher abundances in the mesopelagic than in photic surface or DCM waters. Interestingly, some groups showed contrasting habitat preferences, photic waters for cyanobacterial diazotrophs and mesopelagic waters for gamma- and alphaproteobacterial diazotrophs. Finally, for five of the *nifH* gene sequences we identified its corresponding 16S rRNA gene via correlation analysis between the abundance profiles of both markers in 135 metagenomic samples. Our results provide the first primer-free global survey of the diversity, abundance and distribution of diazotrophs in the oceans that includes the mesopelagic realm and unveils new *nifH* variants that were previously missed by the use of primers.

Introduction

The amount of bioavailable nitrogen supplied via the biological nitrogen fixation process, i.e., the reduction of dinitrogen (N_2) gas to ammonium, is a globally relevant process that often limits primary productivity in wide areas of the oligotrophic oceans and, consequently, affects the carbon export from the photic ocean (Karl *et al.*, 2002). The balance between nitrogen fixation and denitrification (i.e., the conversion of nitrate to N_2) determines the size of the oceanic inventory of bioavailable nitrogen and thus primary productivity (Capone *et al.*, 2008). However, global-scale estimates of the marine nitrogen budget suggest that the nitrogen losses via denitrification exceed the nitrogen inputs through nitrogen fixation (Mahaffey *et al.*, 2005), making it very difficult to have accurate estimation of the magnitude of the bioavailable nitrogen budget. Although the imbalance of the budget may be due to the limitations of the methods used for measuring nitrogen fixation (Mohr *et al.*, 2010; Grokopf *et al.*, 2012; Dabundo *et al.*, 2014), it is also possible that the scarce knowledge of the organisms responsible for biological nitrogen fixation, i.e. the nitrogen-fixing microorganisms or diazotrophs, is also limiting our understanding of this fundamental process (Zehr and Kudela, 2011). It is therefore of critical importance to gain insight into the diversity, distribution and ecology of the members making up the diazotrophic communities in the ocean in order to have a better understanding of the marine nitrogen cycle.

Diazotrophic organisms are exclusively found among members of the Domains *Archaea* and *Bacteria* (Zehr *et al.*, 2003). The current knowledge of the diazotrophic diversity has been obtained using the *nifH* marker gene, which encodes for a key structural protein of the nitrogenase enzymatic complex that performs the N_2 fixation reaction (Zehr and Paerl, 1998). Among the marine diazotrophic members, those belonging to the phylum Cyanobacteria have long been considered as the main players in surface waters of the warm oligotrophic ocean (Zehr, 2011). However, the relative contribution of cyanobacterial and non-cyanobacterial diazotrophs to the diazotrophic community is not so clear now; for example, it has been recently shown that the heterotrophic diazotrophs can occupy a broader variety of habitats, including high-latitude, deep, cold, or coastal waters where cyanobacterial diazotrophs are rarely found (Bombar *et al.*, 2017). Furthermore, heterotrophic diazotrophs can dominate the diazotrophic community in vast regions of the ocean, as shown by *nifH* PCR (polymerase chain reaction)-based amplicon surveys (Farnelid *et al.*, 2011; Turk-Kubo *et al.*, 2014; Gradoville *et al.*, 2017; Bombar, *et al.*, 2017). It is thus possible that these heterotrophic diazotrophs play more important roles than previously thought, and indeed numerous studies have tried to assess the contribution of cyanobacterial and non-cyanobacterial diazotrophs based on *nifH*-based abundances from qPCR assays (Luo *et al.*, 2012; Bombar *et al.*, 2017). However, all of these studies rely on the use of primers, which is known to give a biased view of the diversity and relative abundance of microorganisms (Acinas *et al.*, 2005), including the diazotrophic taxa (Gaby and Buckley, 2017). This inaccurate depiction and quantification of the diversity could explain the uncoupling between diazotroph community composition and nitrogen fixation rates often found (Turk-Kubo *et al.*, 2014), and thus there is a need for conducting alternative quantitative surveys of the diversity of these relevant microorganisms.

The first attempt to recruit marine *nifH* gene sequences from a non primer-based approach was using a single marine metagenomic sample from the Sargasso Sea (Venter *et al.*, 2004). Interestingly, only 41 genes out of the 1.2 million genes generated from the whole-shotgun genome sequencing were classified as *nifH* genes (Meyer, 2004), but a more detailed examination of these *nifH* genes showed them to be more closely related to photosynthetic electron transfer proteins in cyanobacteria (e.g. 91% identical to protochlorophyllide reductase of *Prochlorococcus marinus*) than to *nifH* genes (Johnston *et al.*, 2005). Other marine metagenomic studies have identified diazotrophs attached to plastic debris (Bryant *et al.*, 2016) or as part of the microbiome of isolated colonies of *Trichodesmium* (Gradoville, Crump, *et al.*, 2017) but none of these studies detected diazotrophs directly in seawater. This lack of success in recovering diazotrophic diversity may be due to the shallow sequencing depth used in that study, because it is believed that diazotrophic organisms exist in low abundance in natural communities (Johnston *et al.*, 2005). Fortunately, the recent global oceanographic expedition *Tara* Oceans (2009-2012)(Karsenti *et al.*, 2011) has provided the deepest sampling and metagenomic sequencing effort performed to date in open ocean (Sunagawa *et al.*, 2015), which is now publicly available for the exploration of the diversity and distribution of functionally relevant groups such as marine nitrogen-fixing microorganisms.

In the present study, we aimed to obtain a global overview of the diversity and a quantitative estimate of the abundance of marine planktonic diazotrophs in the photic and aphotic layers of the ocean. In order to do so, we explored the metagenomic data generated in the *Tara* Oceans project looking for *nifH* genes across 135 globally distributed samples collected at surface, DCM and mesopelagic waters. This study represents the first attempt to characterize the taxonomic composition and to quantify the relative abundance of nitrogen-fixing microorganisms using an approach that does not rely on the use of primers, and which is thus thought to provide a more realistic view of these diazotrophs in the global ocean.

Material and methods

Study area and sample collection

From 2009 to 2013, the *Tara* Oceans expedition collected biological samples across all the major oceanic basins (Karsenti *et al.*, 2011; Sunagawa *et al.*, 2015). Sample location for the data used in this paper is shown in Figure 1. A total of 135 samples from 68 globally distributed stations including surface (n=63), deep chlorophyll maximum (DCM) (n=42) and mesopelagic (n=30) seawater samples were collected for metagenomic sequencing during the *Tara* Oceans expedition encompassing two different size fractions (i.e., 0.2–1.6 µm from stations TARA_004 to TARA_052 and 0.2–3 µm from stations TARA_056 to TARA_152) (Sunagawa *et al.*, 2015) that mostly represent free-living prokaryotic communities. In principle, the 0.2-1.6 μm fraction was going to be used during the whole campaign, but the type of material of the 1.6 μm size-pore prefilter, i.e., glass microfiber, was not resistant enough for supporting the volume of seawater needed for metagenome sequencing (ca. 100 liters) and these prefilters were replaced by 3 μm size-pore polycarbonate filters from station TARA_052 until the end of the cruise.

Nucleic acid extractions and sequencing

The 0.2 μm filters were cut into small pieces with sterile razor blades and half of each filter was resuspended in 3 ml of lysis buffer (40 mM EDTA, 50 mM Tris-HCl, 0.75 M sucrose). Lysozyme (1 mg ml^{-1} final concentration) was added and the samples were incubated at 37ºC for 45 min with slight movement. Then, sodium dodecyl sulfate (SDS, 1% final concentration) and proteinase K (0.2 mg ml^{-1} final concentration) were added and the samples were incubated at 55ºC for 60 min under slight movement. The lysate was collected and processed with the standard phenol-chloroform extraction procedure: an equal volume of Phenol:CHCl3:IAA (25:24:1, vol:vol:vol) was added to the lysate, carefully mixed and centrifuged 10 min at 3,000 rpm. Then the aqueous phase was recovered and the procedure was repeated. Finally, an equal volume of CHCl3:IAA (24:1, vol:vol) was added to the recovered aqueous phase in order to remove residual phenol. The mixture was centrifuged and the aqueous phase was recovered for further purification. The aqueous phase was then concentrated by centrifugation with a Centricon concentrator (Millipore, Amicon Ultra-4 Centrifugal Filter Unit with Ultracel-100 membrane). Once the aqueous phase was concentrated, this step was repeated three times adding 2 ml of sterile MilliQ water each time in order to purify the DNA. After the third wash, between 100 and 200 μl of purified total genomic DNA product per sample could be recovered. The extracted DNA was quantified using a Nanodrop ND-1000 spectrophotometer (NanoDrop Technologies Inc, Wilmington, DE, USA) and the Quant_iT dsDNA HS Assay Kit with a Qubit fluorometer (Life Technologies, Paisle).

All the information concerning metagenome sequencing can be found in Sunagawa *et al.* (2015) and in Alberti *et al.* (2017). Briefly, the extracted DNA samples were sequenced using the Illumina technology as overlapping paired reads of ~100/108 bp. After quality control, reads were merged using FLASH v1.2.7 with default parameters (Magoč and Salzberg, 2011) and cleaned based on quality using CLC QualityTrim v4.10.86742 (CLC Bio), resulting in 100–215-bp fragments. The location and sequencing effort of all metagenomic samples can be found in Table W2 from Sunagawa *et al.* (2015). All metagenomes and the corresponding environmental parameters measured during the *Tara* Oceans expedition are available online (www.pangaea.de).

Identification of nifH gene sequences from metagenomic datasets

Reads coming from the *Tara* Ocean metagenomes were assembled to build contigs and to predict genes within these contigs to, afterwards, generate the Ocean Microbial Reference Gene Catalog (OM-RGC), a non-redundant set of 40,154,822 genes clustered at 95% identity (Sunagawa *et al.*, 2015). The set of non-redundant genes were functionally annotated by blasting the protein sequences of the genes against eggNOG v3 (Powell *et al.*, 2012) and KEGG v62 (Kanehisa *et al.*, 2012) using SMASH v1.6 (Arumugam *et al.*, 2010). We took advantage from the OM-RGC (Sunagawa *et al.*, 2015) to explore the presence of the marker gene used for identifying diazotrophic microorganisms, the *nifH* gene. In this sense, within the OM-RGC we screened both the gene sequences annotated using the eggNOG database as COG1348 (nitrogenase reductase subunit NifH (ATPase)) and the gene sequences annotated using the KEGG database (Kyoto Encyclopedia of Genes and Genomes) as K02588 (*nifH*, nitrogenase iron protein NifH).

Abundance patterns of nifH gene sequences in the Tara Oceans dataset

After generating the reference gene catalog, reads from each sample were mapped to the catalog to estimate functional and taxonomic abundances (Sunagawa *et* *al.*, 2015). For each sample (n=135), the abundance of each gene sequence in the OM-RGC was determined using MOCAT (Kultima *et al.*, 2012). Based on the functional annotations of the OM-RGC, these gene sequence abundances were summarized at the level of: (i) eggNOG gene families (genes annotated to the eggNOG version 3 database (Powell *et al.*, 2012)), (ii) KEGG orthologous groups and (iii) KEGG modules (Sunagawa *et al.*, 2015). We downloaded the abundance profiles based on the KEGG annotation and extracted the information concerning to KOs K02588 (*nifH*, nitrogenase iron protein NifH) and K03553 (*recA*, recombination protein RecA). We used the abundance of K03553 (*recA*–based total bacterial abundance) to normalize the abundance of K02588 (*nifH*–based diazotrophs abundance). Since both *nifH* and *recA* are single-copy genes, the gene-based abundance approach can be easily used to estimate the relative contribution of diazotrophs to total bacterioplankton community. Additionally, the abundances across samples were also obtained for each individual *nifH* gene sequence detected in the OM-RGC.

Statistical analyses

Statistically significant ($p< 0.05$) differences in relative abundance of *nifH* gene sequences between photic (surface and DCM) and aphotic layers were assessed using one-way ANOVA and post hoc analyses (Tukey's honestly significant difference test). All analyses were performed using JMP 9.0.1 (SAS Institute, NC, USA) or R 3.0.0 software (R Core Team, 2013). In order to assign a taxonomic level to the 28 *nifH* gene sequences extracted from the OM-RGC, linear regression analyses (`lm` function of the R stats package) were performed between the abundance of Operational Taxonomic Units (OTUs) based on 16S $_{\mathrm{mi}}$TAGs extracted from the 135 metagenomes (Sunagawa *et al.*, 2015) and the normalized abundances of the 28 *nifH* gene sequences.

Results and discussion

Identification of nifH gene sequences from the Ocean Microbial Reference Gene Catalog (OM-RGC)

Our first attempt to recruit *nifH* gene sequences from the OM-RGC was based on the eggNOG annotation (eggNOG version 3 database (Powell *et al.*, 2012)). We screened the gene sequences annotated within the OM-RGC with the eggNOG code 'COG1348', defined as nitrogenase reductase subunit NifH (ATPase). The eggNOG annotation has been used in previous metagenomic studies to detect diazotrophs in a variety of environments such as cave, soils, rhizosphere, groundwater, wetlands, wastewater treatment plants or oceans (Ortiz *et al.*, 2014; Wang *et al.*, 2014; He *et al.*, 2015; Hemme *et al.*, 2015). In the OM-RGC, we retrieved 701 gene sequences annotated as COG1348. However, we found that the search based on COG1348 classification recruited not only *nifH* gene sequences but also other genes not involved in nitrogen fixation. Since these genes were also annotated with the KEGG classification, we explored how many different KOs corresponded with COG1348. We found that among the 701 gene sequences annotated as COG1348, only 19 genes were assigned to K02588 (*nifH*, nitrogenase iron protein NifH). Among the remaining 682 gene sequences, 476 were assigned to K04037 (*chlL,* light-independent protochlorophyllide reductase subunit L) and 206 were annotated as K11333 (*bchX*, 3,8-divinyl chlorophyllide a/chlorophyllide a reductase subunit X). For instance, the gene *chlL*, annotated as COG1348, is involved in the synthesis of the Protochlorophyllide reductase, an enzyme widespread among photosynthetic microorganisms such as the abundant *Prochlorococcus* (Partensky *et al.*, 1999; Fujita and Bauer, 2000). Therefore, the assessment of nitrogen fixation based on eggNOG annotation can give an overestimated view of the nitrogen fixation process in the ocean.

In order to avoid this problem, we opted for the KEGG classification instead of using the eggNOG approach. Surprisingly, we recruited 28 *nifH* gene sequences from the OM-RGC annotated as K02588 (Table 1), which means 9 *nifH* genes more than those recruited with the eggNOG approach. Therefore, the eggNOG-based approach does not only overestimate the number of *nifH* genes involved in nitrogen fixation but also fails to identify some of the actual nitrogen fixation genes.

OM-RGC gene ID	Closest reference NifH protein [ACCN]	BLASTX (% identity)	Closest nifH nucleotide sequence [ACCN]	BLASTN (% identity)	No. Mismatches (nifH1;nifH2;nifH3;nifH4)
Cluster I - Alphaproteobacteria					
OM-RGC.v1.007377081	*Bradyrhizobium* sp. BTAi1 [WP_012045780.1]	100%	-	-	(0;0;0;0)
OM-RGC.v1.007388266	*Rhodovulum* sp. NI22 [WP_037205988.1]	100%	-	-	(0;0;0;0)
OM-RGC.v1.007501166	*Rhodobacter johrii* [WP_069332891.1]	100%	-	-	(0;0;0;0)
OM-RGC.v1.009841511	*Yangia pacifica* [WP_092419747.1]	99%	-	-	(0;0;0;NA)
OM-RGC.v1.038254821	*Pseudoruegeria marinistellae* [WP_068114929.1]	88%	*Yangia* sp. CCB-MM3 [CP014595.1]	88%	(**1**;NA;NA;**1**)
Cluster I - Gammaproteobacteria					
OM-RGC.v1.007436991	*Thiocapsa marina* [WP_007192139.1]	100%	-	-	(0;0;**1**;**1**)
OM-RGC.v1.007483613	*Pseudomonas stutzeri* [WP_003298004.1]	100%	-	-	(0;0;0;**1**)
OM-RGC.v1.007595848	*C.* Thiodiazotropha endolucinida [WP_069124654.1]	98%	-	-	(0;0;0;**1**)
OM-RGC.v1.007601814	*Marinobacterium litorale* [WP_027854782.1]	98%	-	-	(0;0;0;**1**)
OM-RGC.v1.007647800	*Amphritea atlantica* [WP_091356304.1]	98%	-	-	(0;0;0;**1**)
OM-RGC.v1.007667460	*Marinobacterium litorale* [WP_027854782.1]	97%	-	-	(0;0;0;**1**)
OM-RGC.v1.011403932	*C.* Thiodiazotropha endolucinida [WP_069124654.1]	99%	-	-	(NA;0;0;NA)
OM-RGC.v1.026833116	*Nitrincola* sp. A-D6 [WP_036523957.1]	100%	-	-	(0;NA;NA;**1**)
OM-RGC.v1.028674582	*Amphritea atlantica* [WP_091356304.1]	99%	-	-	(0;NA;NA;**1**)
OM-RGC.v1.037241215	*Amphritea atlantica* [WP_091356304.1]	100%	-	-	(0;NA;NA;**1**)
Cluster I - Epsilonproteobacteria					
OM-RGC.v1.006859767	*Arcobacter* sp. L [WP_014472750.1]	100%	-	-	(0;0;0;0)
Cluster I - Cyanobacteria					
OM-RGC.v1.007316555	*Trichodesmium erythraeum* [WP_011613474.1]	100%	-	-	(0;0;0;0)
OM-RGC.v1.007828191	*C.* Atelocyanobacterium thalassa [WP_012954002.1]	100%	-	-	(0;0;0;0)
Cluster III - Deltaproteobacteria					
OM-RGC.v1.007482987	*Pseudodesulfovibrio profundus* [WP_097013057.1]	100%	-	-	(0;0;0;**1**)
OM-RGC.v1.008529501	*Desulfovibrio dechloracetivorans* [WP_071544382.1]	89%	Uncultured bacterium [HM750631.1]	87%	(0;0;0;**1**)
OM-RGC.v1.008691244	*Desulfobulbus propionicus* [WP_015725710.1]	100%	-	-	(0;0;0;**1**)
OM-RGC.v1.008734443	*Desulfospira joergensenii* [WP_022667737.1]	96%	-	-	(0;0;0;**1**)
OM-RGC.v1.008759637	*Desulfovibrio* sp. U5L [WP_009108732.1]	100%	-	-	(0;0;0;**1**)
OM-RGC.v1.031135473	*Dehalococcoides mccartyi* [WP_010936850.1]	82%	*Desulfarculus baarsii* [CP002085.1]	77%	(NA;NA;NA;NA)
OM-RGC.v1.032496133	*Desulfovibrio vulgaris* [WP_011176593.1]	85%	Uncultured bacterium [DQ078045.1]	94%	(0;NA;NA;**1**)
Cluster III - Firmicutes					
OM-RGC.v1.010396209	*Anaerovirgula multivorans* [WP_089285040.1]	51%	Uncultured bacterium [KF847385.1]	77%	(NA;**4**;0;NA)
OM-RGC.v1.013419284	*Clostridium* sp. NCR [WP_035114035.1]	54%	Uncultured microorganism [HQ224439.1]	81%	(**2**;NA;NA;**1**)
OM-RGC.v1.031513582	*Lachnospiraceae bacterium* [WP_094177213.1]	59%	Uncultured microorganism [HQ223500.1]	76%	(**2**;NA;NA;**2**)

Table 1. **Phylogenetic distribution of *nifH* genes recruited from the OM-RGC.** *nifH* genes recruited from the OM-RGC were assigned to major *nifH* clusters based on the classification of Zehr *et al.* (2003). The closest *nifH* gene reference protein sequences found in NCBI (refseq_protein database) are shown. The closest *nifH* gene sequence found in NCBI (*nr/nt database*) is shown for those *nifH* genes sharing identity values lower than 95% with sequences in the *refseq_protein* database. For each *nifH* gene sequence recruited from the OM-RGC, the number of mismatches with the primers that most of studies use to assess the diversity of diazotrophs in marine samples, i.e., the primers nifH1, nifH2, nifH3 and nifH4 (Zehr and McReynolds, 1989; Zani *et al.*, 2000), are indicated in the last column. NA means that the recruited gene sequence does not overlap with the primer region.

We assigned a phylogenetic identity to the 28 *nifH* gene sequences recruited from the OM-RGC. These sequences were classified into canonical *nifH* Clusters (Zehr *et al.*, 2003), and we found that 18 *nifH* gene sequences belonged to Cluster I, and the remaining *nifH* sequences were classified into Cluster III (Table 1). Among the Cluster I *nifH* genes, only two of them were cyanobacterial sequences, whereas the rest belonged to the phylum Proteobacteria, specifically to class Alphaproteobacteria (n=5), Gammaproteobacteria (n=10) and Epsilonproteobacteria (n=1). The *nifH* sequences classified into Cluster III belonged to Deltaproteobacteria (n=7) and phylum Firmicutes (n=3). Interestingly, some of the recruited *nifH* gene sequences showed high divergence with their closest match in the NCBI database, both at the amino acid and nucleotide level (Table 1), suggesting that they represent novel diazotrophic diversity not found in previous studies.

Adequacy of the commonly used nifH primers to assess diazotrophic diversity

All the previous studies assessing the diversity of marine diazotrophic microorganisms (e.g. (Luo *et al.*, 2012; Bombar *et al.*, 2017)) have been based on the use of a set of four primers, i.e., nifH1, nifH2, nifH3 and nifH4 primers (Zehr and McReynolds, 1989; Zani *et al.*, 2000). These primers are used in two consecutive steps in nested PCR: one first PCR step using the external pair of primers nifH4 and nifH3, and a second PCR step using the primer pair nifH1/nifH2 (Zani *et al.*, 2000). Conversely, our approach reconstructs the *nifH* genes from the metagenomes and thus is free from the general biases associated to the use of primers (Acinas *et al.*, 2005), which also in the case of diazotrophic microorganisms has been shown to lead to an inaccurate view of their actual diversity (Gaby and Buckley, 2017). We thus evaluated whether these four primers would amplify the *nifH* gene variants obtained by our approach. Our analysis revealed that the nifH4 presented 1 or 2 mismatches with 18 *nifH* gene sequences (out of 28) (Table 1). This number might be even higher because in some cases the nifH4 primer-binding site region in the recruited genes was not covered. Additionally, three *nifH* gene sequences showed 1 or 2 mismatches with the nifH1 primer; one *nifH* gene sequence showed 4 mismatches with the nifH2 primer; and finally, another *nifH* gene sequence showed 1 mismatch with the nifH3 primer (Table 1). Except for the OM-RGC.v1.038254821 *nifH* sequence that showed one mismatch in the fourth position of the nifH4 primer (C instead of T), and the OM-RGC.v1.031513582 *nifH* sequence that had two mismatches with the nifH4 primer (G instead of T in 1st position and A instead of G/T in 3rd position), the remaining sequences showed all the same mismatch: A instead of T in the first position of the nifH4 primer. Since, as stated above, the nifH4 primer has been used in the initial PCR step of all *nifH* diversity studies, its use may have led to a wrong view of the key nitrogen-fixing players. The nifH1/nifH2 primer pair was originally designed for the detection of *Trichodesmium* (Zehr and McReynolds, 1989) and the nifH3/nifH4 primer pair to evaluate the diazotrophic assemblages in lakes (Zani *et al.*, 2000). Therefore, it can be expected, and now it is demonstrated, that the diversity recruited by these primers may be somehow biased towards diazotrophs that do not represent the key members of nitrogen fixation in marine systems. In fact, one of the reasons behind the lack of link between nitrogen fixation rates and diazotrophic diversity (Turk-Kubo *et al.*, 2014; Gradoville, Bombar, *et al.*, 2017) could be such biased representation of the diazotrophic diversity caused by the use of primers. This biased view, however, is particularly problematic when evaluating the heterotrophic diazotrophs, particularly the gammaproteobacterial (Cluster I) and deltaproteobacterial (Cluster III) diazotrophs (Table 1). However, for the key cyanobacterial diazotrophic members (*Trichodesmium* and UCYN-A) the nifH1-4 primers did not show any mismatch (Table 1). After this analysis of the recovered *nifH* diversity, we propose a modification of the primer nifH4 as follows:

5' – **T**TY TAY GGN AAR GGN GG -3' (nifH4) (Zani *et al.*, 2000)

5' – **W**TY TAY GGN AAR GGN GG -3' (nifH4_mod) (this study)

The modification of this primer should allow detecting some heterotrophic *nifH* variants whose existence has remained hidden until now.

Contribution of diazotrophs across ocean basins

We carried out an exploration of *nifH*-based abundance in order to determine the importance of diazotrophs within the studied bacterioplankton communities (*recA*-based abundance) at the global scale. To our knowledge this is the first global quantification of the relative abundance of different *nifH* variants in the ocean that does not rely on the use of primers, and thus may represent the most accurate quantification of the diazotroph contribution to communities conducted so far. We found diazotrophs distributed across all the studied oceanic regions and in most of the samples: their presence was detected in 46 surface samples (out of 63), 30 DCM samples (out of 42) and 29 mesopelagic samples (out of 30) (Fig. 1). However, their relative abundance was generally low (always below 0.6% of the communities), and in general their abundances were highest in the Atlantic and Pacific oceans and lowest in the Mediterranean Sea, the Red Sea and the Indian Ocean (Fig. 1). We observed significantly higher relative abundances of diazotrophs in mesopelagic (0.07% of the bacterioplankton community) than in surface (0.04%) or DCM waters (0.02%) (Fig. 2a). Pooling all samples together, the majority of *nifH* gene sequences were associated to gammaproteobacterial diazotrophs from Cluster I (63% of *nifH* sequences), followed by Cluster I cyanobacterial diazotrophs (16%) and deltaproteobacterial (Cluster III) *nifH* sequences (12%) (Fig. 2b). The relative contribution of the different clusters also varied with depth, and some of them showed a clear preference for a particular layer: For example, whereas cyanobacterial phylotypes showed their maximum contribution to diazotrophic communities in surface waters, Firmicutes did so in DCM and Alphaproteobacteria *nifH* phylotypes are preferentially located in mesopelagic waters (Fig. 2b). In all cases, however, the total pool of *nifH* sequences was dominated by gammaproteobacterial taxa, which ranged from 46% of the *nifH* sequences in surface to 81% in mesopelagic waters (Fig. 2b). These results are in agreement with the only global high-throughput *nifH* amplicon sequencing study, which showed a dominance of heterotrophic diazotrophs in all marine regions assessed (Farnelid *et al.*, 2011), and with the observation that gammaproteobacterial diazotrophs are often dominant (Turk-Kubo *et al.*, 2014).

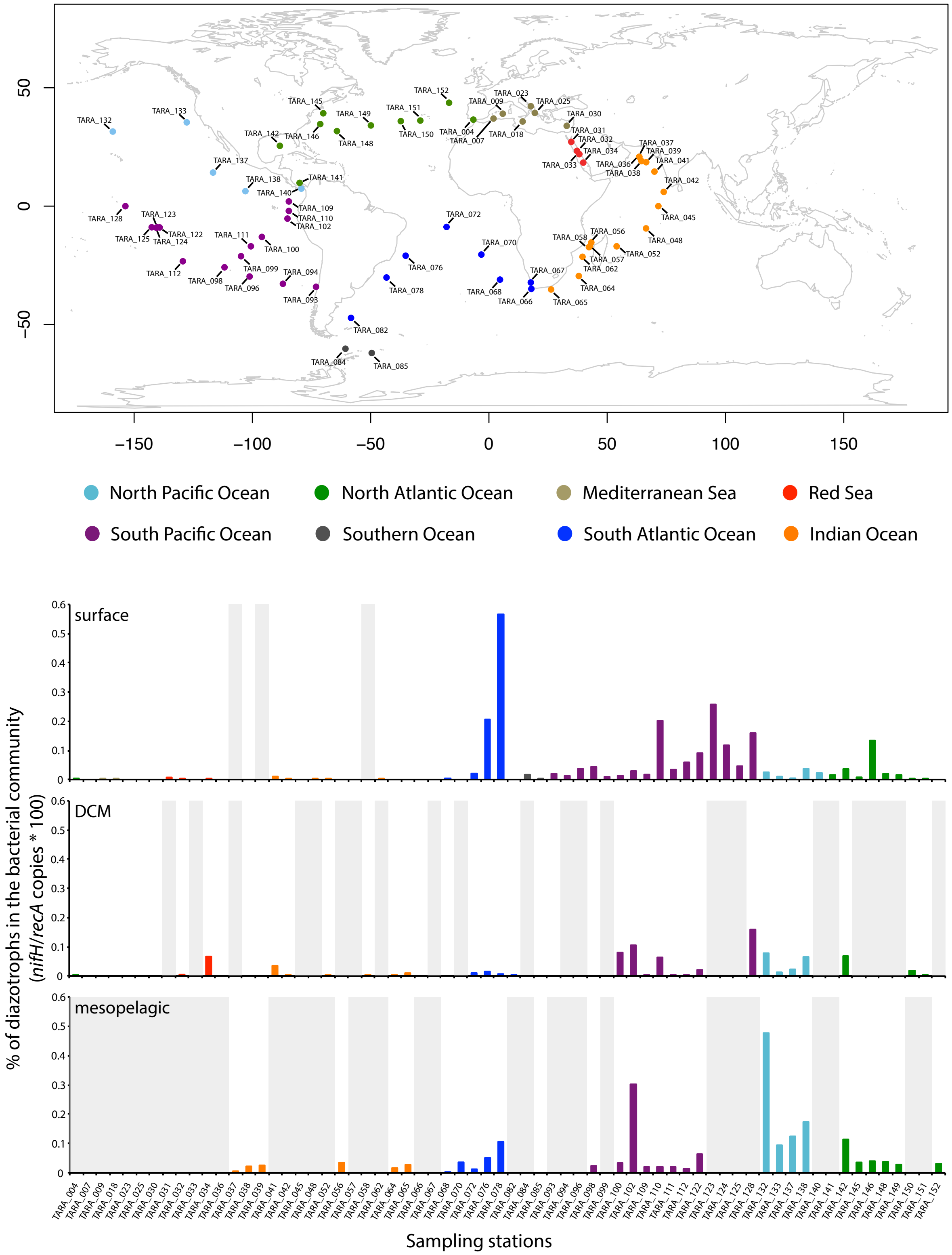


Figure 1. **Contribution of diazotrophs to total bacterioplankton community in the global ocean.** The upper panel shows the stations sampled during the *Tara* Ocean cruise. The lower panel shows the % of abundance of diazotrophs in the *Tara* Oceans samples, and is expressed as % of the total *recA* sequences per community. The absence of sample is indicated by the gray color.

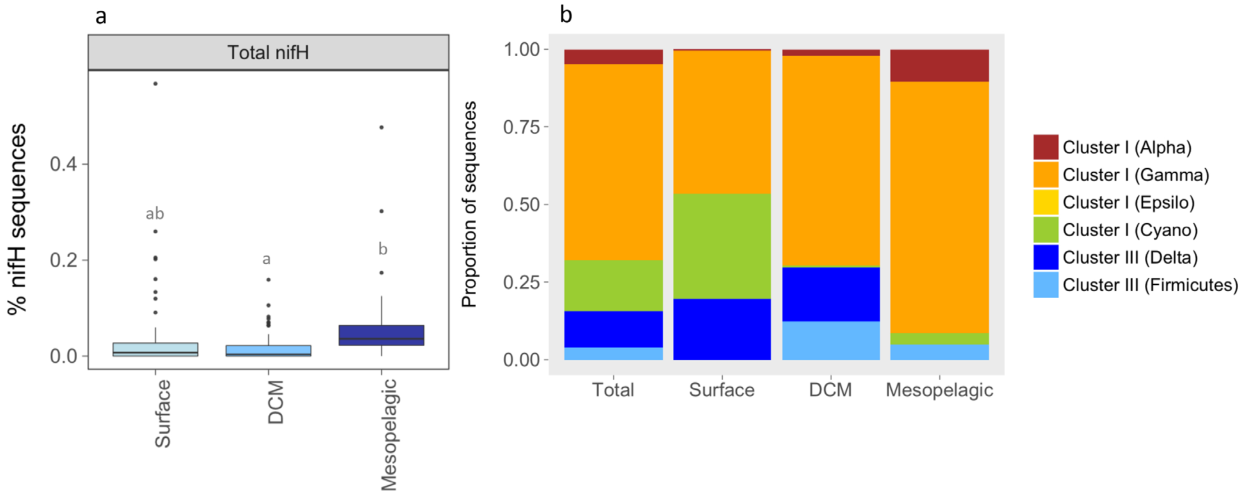


Figure 2. **a)** Relative contribution of total nifH sequences to prokaryotic communities between the different oceanic layers. Abundance values are expressed as % of the recA sequences per community. Different letters indicate significant differences (Tukey's post hoc test; $P < 0.05$) between depths. **b)** Relative contribution of the different nifH canonical clusters to the total nifH sequences considering the total dataset (Total) or each oceanic layer.

In terms of their contribution to the community, we observed significant differences between depths only in the case of Gamma- and Alphaproteobacteria, both of which showed higher relative abundances in mesopelagic waters than in surface and DCM waters (Fig. 3). Interestingly, the opposite pattern was previously found for the gammaproteobacterial phylotype Gamma A, showing that its ecological niche apparently overlaps with that of cyanobacterial diazotrophs, i.e., within the warm, oligotrophic, fully oxygenated surface waters of subtropical and tropical latitudes (Langlois *et al.*, 2015). However, this particular Gamma A clade was not detected in our dataset (see below). Finally, Deltaproteobacteria diazotrophs showed higher relative abundances in surface waters, but these differences were not statistically significant (Fig. 3).

Surprisingly, even though nitrogen is considered as a limiting nutrient for primary productivity in the open ocean (Karl *et al.*, 2002), we did not find a homogenous abundance contribution of diazotrophs to total bacterioplankton communities across ocean basins (Figure 1). In this sense, although biological nitrogen fixation is an important mechanism for new nitrogen supply (Capone *et al.*, 2005; Mouriño-Carballido *et al.*, 2011; Painter *et al.*, 2013; Martínez-Pérez *et al.*, 2016), turbulent diffusion across the nitracline has long been considered the dominant source of new nitrogen to the surface ocean and, consequently, it could explain the non-homogeneous contribution of diazotrophs across basins. For instance, in a recent study using data collected from the global Malaspina2010 expedition, the relative contribution of nitrate diffusive fluxes (due both to mechanical turbulence and salt fingers) to the new nitrogen supply was much higher (ca. 85%) than that of the nitrogen fixation (ca. 15%)(Fernández-Castro *et al.*, 2015). In that study, the biological nitrogen fixation was highest at the South Atlantic Gyral province (SATL), which is in agreement with our results on diazotrophic abundances, being the sampling station TARA_078 located at the SATL province the one showing the highest contribution of diazotrophs to the total bacterioplankton community (Figure 1). In general, however, we could not find any clear correlation between the total *nifH* abundances or those of the different clusters or phylotypes with any of the measured physico-chemical variables measured during the *Tara* Ocean cruise (environmental data in Sunagawa *et al.*, 2015).

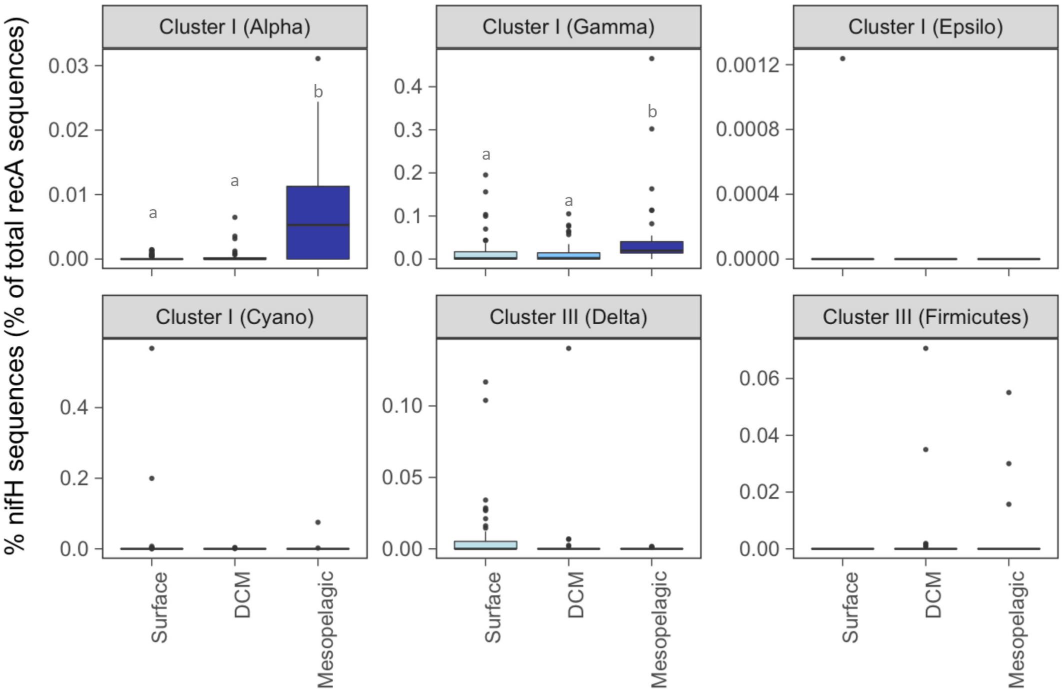


Figure 3. Relative contribution of the different nifH clusters to total communities across the different oceanic layers. Abundance values are expressed as % of the recA sequences per community. Different letters indicate significant differences (Tukey's post hoc test; $P < 0.05$) between depths. Note the different Y axes.

Contribution of diazotrophic phylotypes across ocean basins

In addition to exploring the abundance of diazotrophs at the nifH canonical cluster level, we also addressed the distribution of these diazotrophs at the sequence level (Fig. 4). Although we detected a total of 28 distinct *nifH* gene sequences, the number of *nifH* gene sequences per sample was much lower: 3 *nifH* variants on average, ranging between 0 (e.g. TARA_085 mesopelagic or TARA_056 surface) and 11 (TARA_122, mesopelagic). Interestingly, the highest relative abundance of diazotrophs in mesopelagic waters (Fig. 2) was also accompanied by a higher number of *nifH* gene sequences per sample in mesopelagic waters (~5 genetic variants per sample) than in the surface (~3) or DCM (~2) waters (Fig. 4). Moreover, we observed that some *nifH* gene sequences appeared only in one sample, like for example OM-RGC.v1.007436991 (Cluster I Gamma.), OM-RGC.v1.006859767 (Cluster I Epsilon.), OM-RGC.v1.008691244 (Cluster III Delta.) or OM-RGC.v1.031513582 (Cluster III Firmicutes). Interestingly, although different gammaproteobacterial *nifH* genetic variants accounted for a significant fraction of the *nifH* abundance (e.g., the OM-RGC.v1.007595848 (14.6%) or OM-RGC.v1.007601814 (14.4%) genes), the globally distributed group of marine gammaproteobacterial diazotroph Gamma A (Church *et al.*, 2005; Langlois *et al.*, 2015) was not detected in our dataset.

Among the different *nifH* variants detected, the most abundant was the cyanobacterial gene OM-RGC.v1.007828191 (16.2%). The OM-RGC.v1.007828191 gene was 100% identical to the *nifH* gene of the uncultured unicellular symbiotic cyanobacterium *Candidatus* Atelocyanobacterium thalassa (UCYN-A), in particular, to the UCYN-A1 sublineage (Thompson *et al.*, 2012) (Table 1). The UCYN-A clade is known to be one of the most widespread and important nitrogen fixers of the oligotrophic oceans (Zehr *et al.*, 2016). However, although UCYN-A1 was the most abundant *nifH* gene sequence in the *Tara* Oceans dataset, its presence was confined to a few samples (12 out of 135). Moreover, at least three different UCYN-A sublineages exist based on the diversity of the *nifH* gene (Thompson *et al.*, 2014), and we know that the nucleotide divergence between the *nifH*-based UCYN-A lineages is lower than 5%. In our study, we have explored the *nifH* genetic diversity using the OM-RGC (Sunagawa *et al.*, 2015) and, although the amount of genetic information is of great value, microdiversity signals such as the one shown for the UCYN-A *nifH* sublineages could be hidden or masked behind the 95% identity clustering threshold applied for the OM-RGC construction. This fact calls for the necessity of working directly with the metagenomic reads, besides with the reconstructed genes, in order to evaluate whether different *nifH* genetic variants can emerge from the available metagenomes.

As previously mentioned, in the Eastern Tropical South Pacific (ETSP) heterotrophic diazotrophs dominate the diazotrophic community throughout the water column (Fernandez *et al.*, 2011; Bonnet *et al.*, 2013; Turk-Kubo *et al.*, 2014; Loescher *et al.*, 2014; Gradoville *et al.*, 2017). However, a direct link between nitrogen fixation rates and heterotrophic diazotroph phylotypes in this particular area has not been demonstrated (Turk-Kubo *et al.*, 2014; Loescher *et al.*, 2014). In our study, we detected in the ETSP some *nifH* variants that presented 1 or more mismatches with the nifH primer pairs used to assess the diversity (Table 1). In particular, the gammaproteobacterial diazotroph corresponding with the OM-RGC.v1.007667460 gene showed a marked high abundance in this area (>0.1% relative abundance in TARA_102 mes. and TARA_110 surf.) (Fig. 1). So, we analysed whether the qPCR probes designed and used in Loescher *et al.,* 2014 to quantify diazotrophic abundance in this area would detect the OM-RGC.v1.007667460 sequence and, interestingly, they did not match with this specific phylotype. Therefore, a feasible hypothesis explaining this lack of connection between nitrogen fixation and diazotrophic diversity can be the biased view of the diazotrophic diversity obtained by the use of primers and probes.

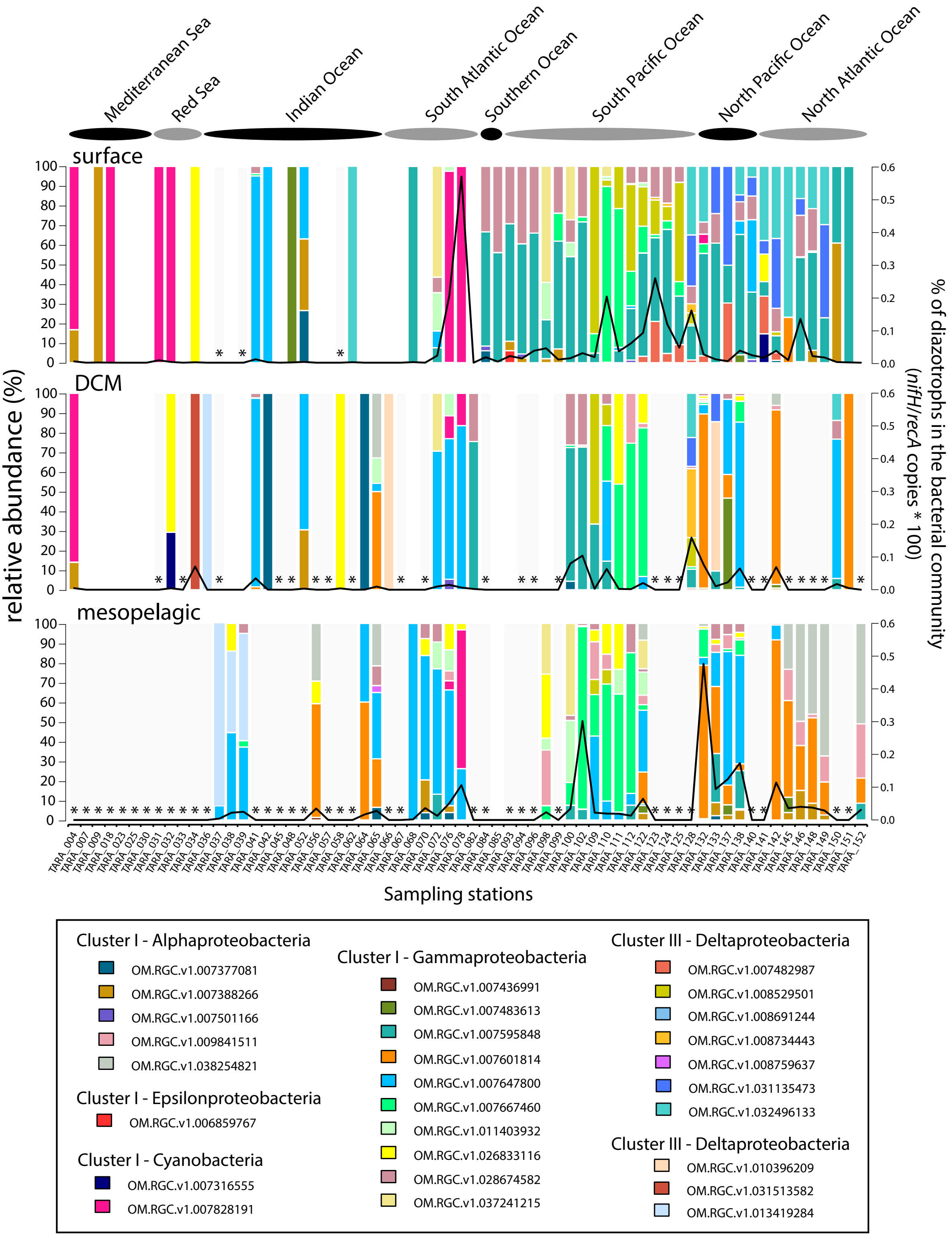


Figure 4. **Distribution of *nifH* phylotypes detected in the *Tara* Oceans expedition.** Relative contribution per sample of the 28 *nifH* genetic variants found in the OM-RGC (left axis). The black line indicates the contribution of diazotrophs to total bacterioplankton community as shown in Figure 1 (right axis). The gene id for each of the variants is indicated in the figure legend. The absence of sample is indicated with asterisks and with gray background color.

Identification of the 16S rRNA gene of diazotrophs

One difficulty in advancing towards a better understanding of the ecology of the marine diazotrophic players is, in most of the cases, the impossibility to get them in culture. In this sense, genome sequencing through single-cell sorting techniques and microscopic approaches has helped to understand the ecology of important diazotrophic players (Zehr *et al.*, 2008; Tripp *et al.*, 2010; Thompson *et al.*, 2012). However, a limiting step for applying visualization techniques such as the CARD-FISH (Catalyzed Reporter Deposition-Fluorescence in situ Hybridization) assay (Pernthaler *et al.*, 2004) lies on the identification of the 16S rRNA gene for the design of specific molecular probes. In this context, we tried to assign a 16S rRNA genetic identity to the *nifH* genes detected in the OM-RGC through correlation analysis of the *nifH* genes with the OTUs based on 16S $_{\mathrm{mi}}$TAGs of the *Tara* Oceans. We obtained positive and significant correlations with 16S OTUs in 5 cases (Fig. 5):

1. The OM-RGC.v1.007828191 *nifH* gene shared a 100% identity at the amino acid level with *C.* Atelocyanobacterium thalassa (UCYN-A) (Table 1) and was significantly correlated with the 16S rRNA gene of the UCYN-A1 sublineage (Fig. 5). The largest metagenomic contig of the *Tara* Ocean dataset containing the OM-RGC.v1.007828191 *nifH* gene was recruited from the TARA_078 surface sample and was 100% identical in its entire length (236.3 Kb) to UCYN-A1 sublineage (NC_013771.1).

2. The OM-RGC.v1.007482987 *nifH* gene shared a 100% identity at the amino acid level with *Pseudodesulfovibrio profundus* (Table 1) and was significantly correlated with the 16S rRNA gene of *Desulfovibrio profundus* DSM 11384 (AF418172.1) (Fig. 5). The closest genome in NCBI (*refseq_genomes*) was *Desulfovibrio profundus* 500-1 (NZ_LT907975.1) that showed, at nucleotide level, 99.87% and 99.04% identity with the 16S rRNA gene of *Desulfovibrio profundus* DSM 11384 and the OM-RGC.v1.007482987 *nifH* gene, respectively. The largest metagenomic contig obtained from the *Tara* Ocean dataset associated with this

nifH phylotype had a length of 23.6 Kb and was recruited from the 0.45-0.8 μm size fraction of the TARA_123 surface sample. In this case, the contig was 98% identical to *Desulfovibrio profundus* 500-1 but it only covered the 68% of the contig length. The remaining 32% of the contig (7.5 Kb) showed lower nucleotide identity (65-70%) with other *Desulfovibrio* spp. such as *Desulfovibrio brasiliensis* (NZ_BBCB01000062.1) or *Desulfovibrio frigidus* DSM 17176 (NZ_JONL01000001.1) suggesting that the OM-RGC.v1.007482987 *nifH* gene belonged to a lineage close but different to *Desulfovibrio profundus* 500-1.

3. The OM-RGC.v1.007601814 *nifH* gene was 98% identical at the amino acid level to *Marinobacterium litorale* (Table 1) and was significantly correlated with the 16S rRNA gene of an uncultured gammaproteobacterium (FJ497479.1) (Fig. 5). The closest genome to the 16S rRNA gene (FJ497479.1) was *Marinomonas* sp. MED121 (NZ_CH672429.1) and showed 79% identity. Likewise, the closest genome to the OM-RGC.v1.007601814 *nifH* gene was *Marinobacterium litorale* DSM 23545 (NZ_AUAZ01000022.1) with 86% of identity. The largest metagenomic contig obtained from the *Tara* Ocean dataset containing this *nifH* variant had a length of 102.8 Kb and was recruited from mesopelagic waters (TARA_132). This contig was 79% identical to the gammaproteobacteria *Oceanobacter kriegii* DSM 6294, although it only covered 46% of the contig length. Therefore, this diazotroph likely represents a new divergent genome.

4. The OM-RGC.v1.008734443 *nifH* gene showed 96% identity with *Desulfospira joergensenii* (Table 1) and significantly correlated with the 16S rRNA gene of an uncultured bacterium (FJ545634.1) (Fig. 5). *Desulfobacter postgatei* 2ac9 (NZ_CM001488.1) was the closest *nifH*-gene containing genome with 84% identity with the OM-RGC.v1.008734443 *nifH* gene. Interestingly, the 16S rRNA gene was related to members of the Candidate Phylum TM6. A recent study comparing the seven genomes available to date of this phylum revealed several features that may indicate that parasitism is widespread within this phylum, like for instance small genome size (1.0–1.5 Mb), lack of complete biosynthetic pathways, presence of

ATP/ADP translocases for parasitizing host ATP pools, or protein motifs to facilitate eukaryotic host interactions (Yeoh *et al.*, 2016). Moreover, none of these genomes showed nitrogen fixation genes. Unfortunately, the largest metagenomic contig obtained from the *Tara* Ocean dataset associated with this *nifH* phylotype was small (1 Kb) and, thus we could not explore, as in the previous cases, whether other genes were closely related to members of the Candidate Phylum TM6.

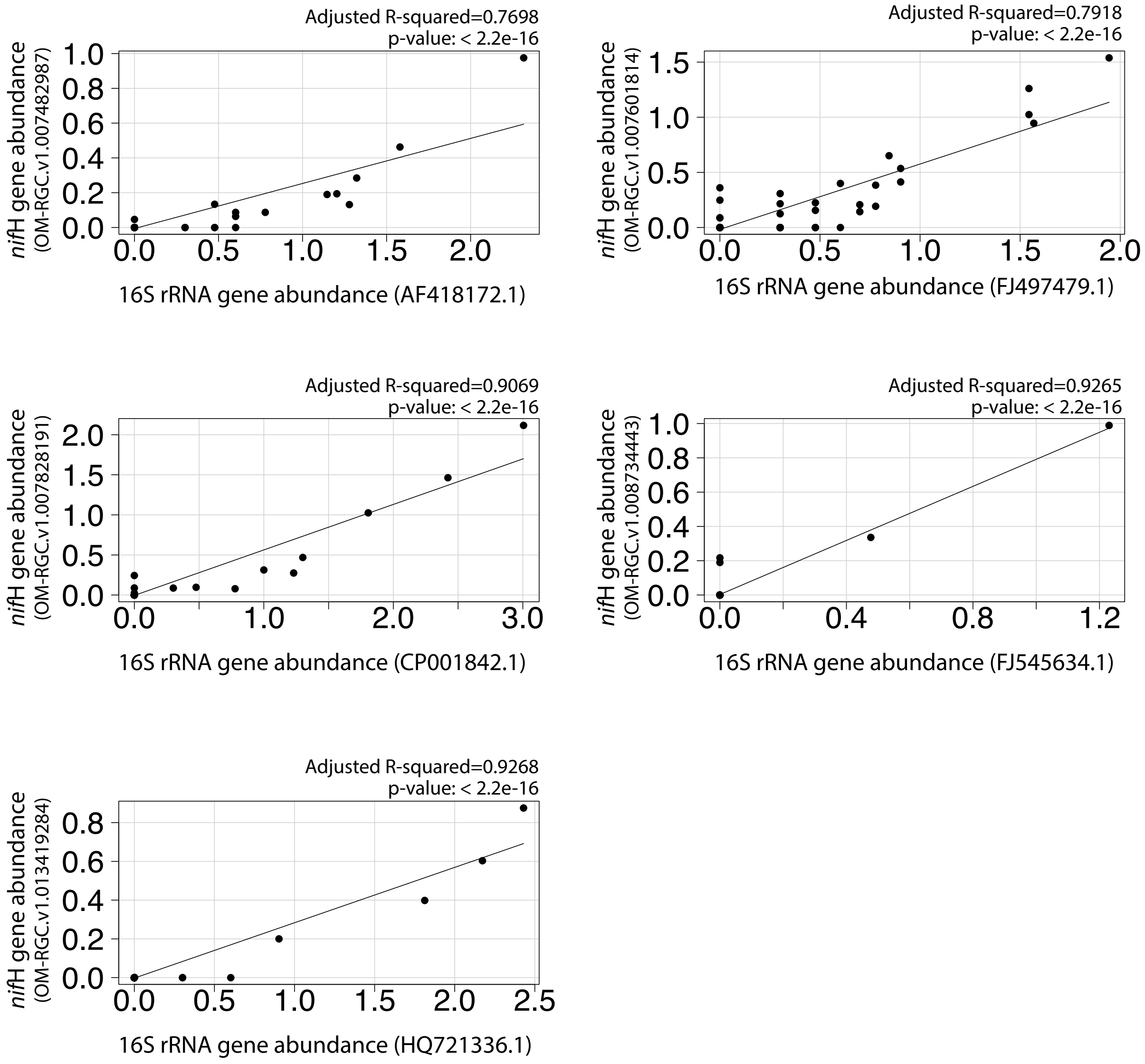


Figure 5. Correlation analysis between *nif*H variants and 16S rRNA genes in *Tara* Oceans. A total of 135 samples were used to perform a regression analysis between the abundance of *nif*H variants and that of 16S rRNA genes. Five *nif*H variants were significantly correlated with 16S rRNA genes (*P* value, regression line, and Adjusted R^2 values are shown). Abundances are represented in log-10 scale. Identification code in the OM-RGC of the *nif*H variants and and accession numbers for the 16S rRNA genes are shown between parenthesis.

5. Finally, the OM-RGC.v1.013419284 *nifH* gene was 54% identical at the amino acid level to *Clostridium* sp. NCR (Table 1) and was significantly correlated with the 16S rRNA gene of an uncultured bacterium (HQ721336.1) (Fig. 5). The associated 16S rRNA gene was related to members of the phylum Deferribacteres, which has not known diazotrophic members. Although the closest *nifH* gene belonged to the phylum Firmicutes, the phylogenetic distance was large enough not to be able to secure the taxonomical assignment to this group. Moreover, at the nucleotide level, the closest *nifH* was an uncultured bacterium that only covered 56% of the OM-RGC.v1.013419284 *nifH* gene length and shared 81% identity (Table 1). Therefore, a deeper exploration is needed to clarify whether the OM-RGC.v1.013419284 gene is a true *nifH* gene.

Concluding remarks

To our knowledge this is the first global study evaluating the significance of diazotrophs based on a non primer-biased approach. Numerous studies have tried to assess the contribution of cyanobacterial diazotrophs based on *nifH*-based abundances from qPCR assays (Luo *et al.*, 2012), but quantitative data on heterotrophic diazotrophs is, in contrast, scarce with the exception of the heterotrophic diazotroph phylotype Gamma A (Bombar *et al.*, 2017), and thus our study provides valuable insight on the relative contribution of heterotrophic diazotrophs across the global ocean. Moreover, all these studies are based on the use of primers, which may result in a biased view when applied in qPCR for estimating the abundance of diazotrophs (Gaby and Buckley, 2017), and based on our results, it may be also disregarding numerically important heterotrophic diazotrophs. We argue that our metagenomic approach is useful to improve the primers and probes used in PCR-based approaches. However, we acknowledge that the sequencing depth may limit the diversity that we cover given that we did not detect abundances higher than 0.6% of diazotrophs in the bacterial community. This means that our target microorganisms belong to the so-called ‘rare biosphere’ (Pedrós-Alió, 2012) and, consequently, a deep sequencing depth is mandatory to detect them in metagenomic studies. In any case, we were able to detect new *nifH* genetic variants in our dataset, mostly non-cyanobacterial ones, reflecting the potential importance of heterotrophic diazotrophs in the marine nitrogen cycle. Finally, we managed to link some of the *nifH* genetic variants to 16S rRNA phylotypes, which represents a step further in the taxonomic identification of these diazotrophs and, although we cannot guarantee an organismal link between both markers, this opens the door to techniques such as CARD-FISH for a visual exploration of the ecology of certain uncultivated diazotrophs. Finally, we argue that a deeper analysis of the available metagenomes through new bioinformatic approaches will allow the reconstruction of new diazotrophic microorganisms, which certainly will expand our comprehension of the marine nitrogen-fixing microorganisms diversity and ecology.

ACKNOWLEDGEMENTS

We thank P. Sánchez for providing access and bioinformatic resources through the Marine Bioinformatics Service (MARBITS) of the Institut de Ciències del Mar (ICM-CSIC) in Barcelona (Spain). We thank also C. Ruiz-González and M.R. Gradoville for critical reading and scientific discussions that have improved the quality of the present study. Funding was provided by BIOSENSOMICS to S.G.A through “Convocatoria 2015 de ayudas Fundación BBVA a investigadores y creadores culturales. F.M.C-C. held FPI fellowship from the Spanish Ministry of Economy, Industry and Competitiveness.

REFERENCES

Acinas SG, Sarma-Rupavtarm R, Klepac-ceraj V, Polz MF. (2005). PCR-induced sequence artifacts and bias: insights from comparison of two 16S rRNA clone libraries constructed from the same sample. *Appl Environ Microbiol* **71**:8966–8969.

Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, *et al.* (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci Data* **4**. doi:10.1038/sdata.2017.93.

Arumugam M, Harrington ED, Foerstner KU, Raes J, Bork P. (2010). SmashCommunity: A metagenomic annotation and analysis tool. *Bioinformatics* **26**:2977–2978.

Bombar D, Paerl RW, Riemann L. (2017). Marine Non-Cyanobacterial Diazotrophs: Moving beyond Molecular Detection. *Trends Microbiol* **24**:916–927.

Bonnet S, Dekaezemacker J, Turk-Kubo KA, Moutin T, Hamersley RM, Grosso O, *et al.* (2013). Aphotic N2 fixation in the eastern tropical South Pacific Ocean. *PLoS One* **8**. doi:10.1371/journal.pone.0081265.

Bryant JA, Clemente TM, Viviani DA, Fong AA, Thomas KA, Kemp P, *et al.* (2016). Diversity and Activity of Communities Inhabiting Plastic Debris in the North Pacific Gyre Jansson, JK (ed). *mSystems* **1**. doi:10.1128/mSystems.00024-16.

Capone DG, Burns JA, Montoya JP, Subramaniam A, Mahaffey C, Gunderson T, *et al.* (2005). Nitrogen fixation by Trichodesmium spp.: An important source of new nitrogen to the tropical and subtropical North Atlantic Ocean. *Global Biogeochem Cycles* **19**:1–17.

Capone G, Bronk DA, Mulholland MR, Carpenter EJ. (2008). Nitrogen in the Marine Environment. doi:10.1016/B978-0-12-372522-6.X0001-1.

Church MJ, Short CM, Jenkins BD, Karl DM, Zehr JP. (2005). Temporal patterns of nitrogenase gene (nifH) expression in the oligotrophic North Pacific Ocean. *Appl Environ Microbiol* **71**:5362–5370.

Dabundo R, Lehmann MF, Treibergs L, Tobias CR, Altabet MA, Moisander PH, *et al.* (2014). The Contamination of Commercial 15N2 Gas Stocks with 15N–Labeled Nitrate and Ammonium and Consequences for Nitrogen Fixation Measurements. *PLoS One* **9**:e110335.

Farnelid H, Andersson AF, Bertilsson S, Al-Soud WA, Hansen LH, Sørensen S, *et al.* (2011). Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* **6**:e19223.

Fernández-Castro B, Mouriño-Carballido B, Marañón E, Chouciño P, Gago J, Ramírez T, *et al.* (2015). Importance of salt fingering for new nitrogen supply in the oligotrophic ocean. *Nat Commun* **6**. doi:10.1038/ncomms9002.

Fernandez C, Farías L, Ulloa O. (2011). Nitrogen fixation in denitrified marine waters. *PLoS One* **6**. doi:10.1371/journal.pone.0020539.

Fujita Y, Bauer CE. (2000). 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.

Gaby JC, Buckley DH. (2017). The Use of Degenerate Primers in qPCR Analysis of Functional Genes Can Cause Dramatic Quantification Bias as Revealed by Investigation of nifH Primer Performance. *Microb Ecol* **74**:701–708.

Gradoville MR, Bombar D, Crump BC, Letelier RM, Zehr JP, White AE. (2017). Diversity and activity of nitrogen-fixing communities across ocean basins. *Limnol Oceanogr* **62**:1895–1909.

Gradoville MR, Crump BC, Letelier RM, Church MJ, White AE. (2017). Microbiome of Trichodesmium Colonies from the North Pacific Subtropical Gyre. *Front Microbiol* **8**:1122.

Grokopf T, Mohr W, Baustian T, Schunck H, Gill D, Kuypers MMM, *et al.* (2012). Doubling of marine dinitrogen-fixation rates based on direct measurements. *Nature* **488**:361–364.

He S, Malfatti SA, McFarland JW, Anderson FE, Pati A, Huntemann M, *et al.* (2015). Patterns in wetland microbial community composition and functional gene repertoire associated with methane emissions. *MBio* **6**:1–15.

Hemme CL, Tu Q, Shi Z, Qin Y, Gao W, Deng Y, *et al.* (2015). Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. *Front Microbiol* **6**:1205.

Johnston AWB, Li Y, Ogilvie L. (2005). Metagenomic marine nitrogen fixation - Feast or famine? *Trends Microbiol* **13**:416–420.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40**:D109–D114.

Karl D, Michaels A, Bergman B, Capone D, Carpenter E, Letelier R, *et al.* (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58(1)**:47–98.

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, *et al.* (2011). A holistic approach to marine eco-systems biology. *PLoS Biol* **9**:e1001177.

Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, *et al.* (2012). MOCAT: A Metagenomics Assembly and Gene Prediction Toolkit. *PLoS One* **7**:e47656.

Langlois R, Großkopf T, Mills M, Takeda S, LaRoche J. (2015). Widespread Distribution and Expression of Gamma A (UMB), an Uncultured, Diazotrophic, γ-Proteobacterial nifH Phylotype. *PLoS One* **10**:e0128912.

Loescher CR, Groskopf T, Desai FD, Gill D, Schunck H, Croot PL, *et al.* (2014). Facets of diazotrophy in the oxygen minimum zone waters off Peru. *ISME J* **8**:2180–2192.

Luo Y-W, Doney SC, Anderson LA, Benavides M, Berman-Frank I, Bode A, *et al.* (2012). Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth Syst Sci Data* **4**:47–73.

Magoč T, Salzberg SL. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**:2957–2963.

Mahaffey C, Michaels AF, Capone DG. (2005). The conundrum of marine N2 fixation. *Am J Sci* **305**:546–595.

Martínez-Pérez C, Mohr W, Löscher CR, Dekaezemacker J, Littmann S, Yilmaz P, *et al.* (2016). The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol* **1**. doi:10.1038/nmicrobiol.2016.163.

Meyer J. (2004). Miraculous catch of iron-sulfur protein sequences in the Sargasso Sea. *FEBS Lett* **570**:1–6.

Mohr W, Großkopf T, Wallace DWR, LaRoche J. (2010). Methodological underestimation of oceanic nitrogen fixation rates. *PLoS One* **5**:1–7.

Mouriño-Carballido B, Graña R, Fernàndez A, Bode A, Varela M, Domínguez JF, *et al.* (2011). Importance of N2 fixation vs. nitrate eddy diffusion along a latitudinal transect in the Atlantic Ocean. *Limnol Oceanogr* **56**:999–1007.

Ortiz M, Legatzki A, Neilson JW, Fryslie B, Nelson WM, Wing RA, *et al.* (2014). Making a living while starving in the dark: metagenomic insights into the energy dynamics of a carbonate cave. *ISME J* **8**:478–491.

Painter SC, Patey MD, Forryan A, Torres-Valdes S. (2013). Evaluating the balance between vertical diffusive nitrate supply and nitrogen fixation with reference to nitrate uptake in the eastern subtropical North Atlantic Ocean. *J Geophys Res Ocean* **118**:5732–5749.

Partensky F, Hess WR, Vaulot D. (1999). Prochlorococcus, a Marine Photosynthetic Prokaryote of Global Significance. *Microbiol Mol Biol Rev* **63**:106–127.

Pedrós-Alió C. (2012). The Rare Bacterial Biosphere. *Ann Rev Mar Sci* **4**:449–466.

Pernthaler A, Pernthaler J, Amann R. (2004). Section 3 update: Sensitive multi-color fluorescence in situ hybridization for the identification of environmental microorganisms. In:*Molecular Microbial Ecology Manual SE - 311*, Kowalchuk, GA, De Bruijn, FJ, Head, IM, Akkermans, AD, & Van Elsas, JD (eds), Springer Netherlands, pp. 2613–2627.

Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, *et al.* (2012). eggNOG v3.0: Orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res* **40**:D284–D289.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, *et al.* (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**:1261359.

Thompson A, Carter BJ, Turk-Kubo K, Malfatti F, Azam F, Zehr JP. (2014). Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol* n/a-n/a.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, *et al.* (2012). Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science (80-)* **337**:1546–1550.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**:90–94.

Turk-Kubo KA, Karamchandani M, Capone DG, Zehr JP. (2014). 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.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, *et al.* (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science (80-)* **304**:66–74.

Wang Z, Zhang X-X, Lu X, Liu B, Li Y, Long C, *et al.* (2014). Abundance and Diversity of Bacterial Nitrifiers and Denitrifiers and Their Functional Genes in Tannery Wastewater Treatment Plants Revealed by High-Throughput Sequencing. *PLoS One* **9**:e113603.

Yeoh YK, Sekiguchi Y, Parks DH, Hugenholtz P. (2016). Comparative Genomics of Candidate Phylum TM6 Suggests That Parasitism Is Widespread and Ancestral in This Lineage. *Mol Biol Evol* **33**:915–927.

Zani S, Mellon MT, Collier JL, Zehr JP. (2000). Expression of nifH genes in natural microbial assemblages in Lake George, New York, detected by reverse transcriptase PCR. *Appl Environ Microbiol* **66**:3119–3124.

Zehr JP. (2011). Nitrogen fixation by marine cyanobacteria. *Trends Microbiol* **19**:162–173.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, *et al.* (2008). Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**:1110–2.

Zehr JP, Jenkins BD, Short SM, Steward GF. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ Microbiol* **5**:539–554.

Zehr JP, Kudela RM. (2011). Nitrogen cycle of the open ocean: from genes to ecosystems. *Ann Rev Mar Sci* **3**:197–225.

Zehr JP, McReynolds LA. (1989). Use of degenerate oligonucleotides for amplification of the nifH gene from the marine cyanobacterium Trichodesmium thiebautii. *Appl Environ Microbiol* **55**:2522–2526.

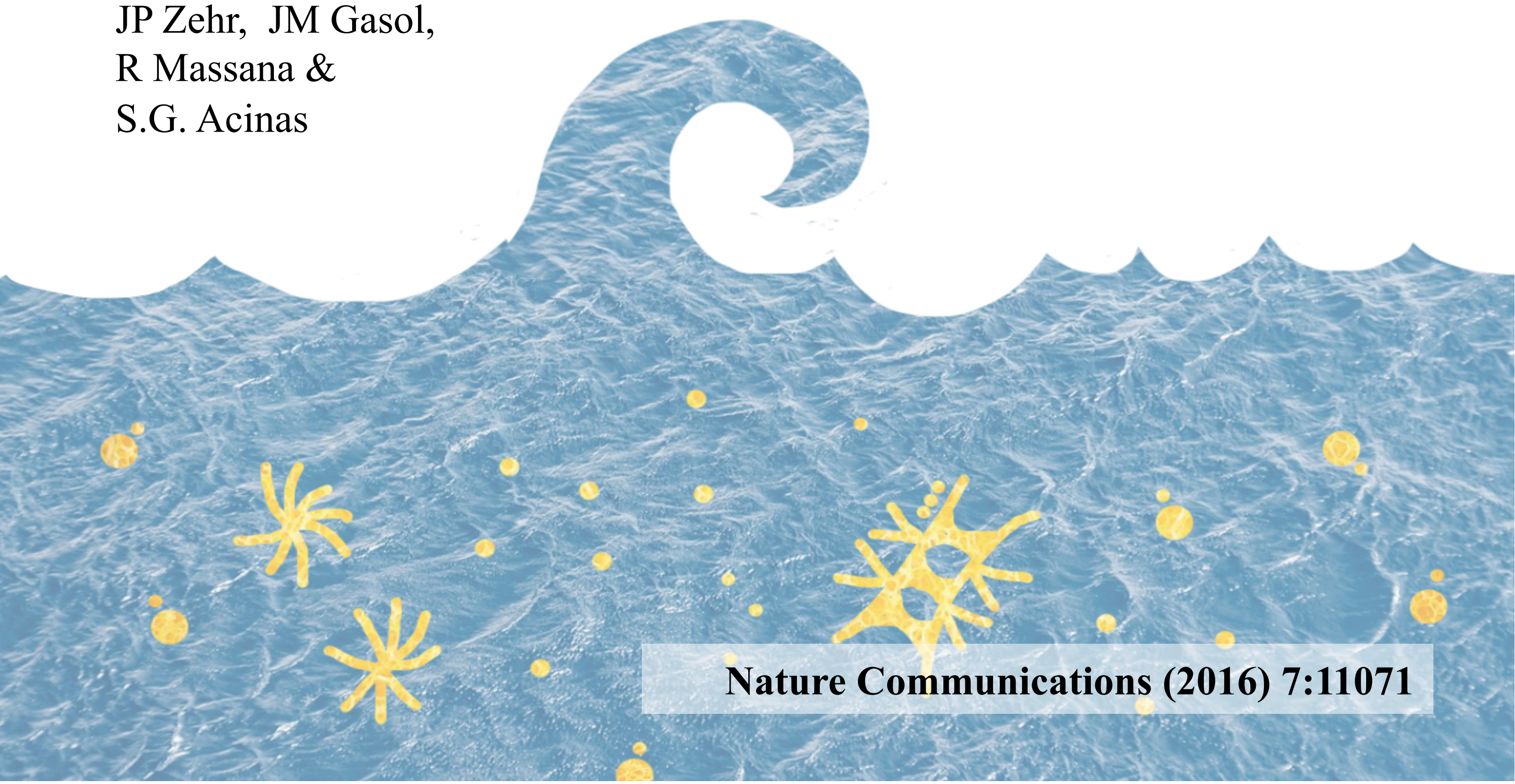
Zehr JP, Paerl H. (1998). Nitrogen Fixation in the Marine Environment: Genetic Potential and Nitrogenase Expression BT - Molecular Approaches to the Study of the Ocean. In: Cooksey, KE (ed), Springer Netherlands: Dordrecht, pp. 285–301.

Zehr JP, Shilova IN, Farnelid HM, Muñoz-Maríncarmen MDC, Turk-Kubo KA. (2016). Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. *Nat Microbiol* **2**. doi:10.1038/nmicrobiol.2016.214.

Chapter 2

Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton

Francisco M. Cornejo-Castillo, AM Cabello, G Salazar, P Sánchez-Baracaldo,
G Lima-Mendez, P HingamP, A. Alberti, S. Sunagawa, P Bork, C de Vargas,
J Raes, C Bowler, P Wincker,
JP Zehr, JM Gasol,
R Massana &
S.G. Acinas



Nature Communications (2016) 7:11071

Abstract

The unicellular cyanobacterium UCYN-A, one of the major contributors to nitrogen fixation in the open ocean, lives in symbiosis with single-celled phytoplankton. UCYN-A includes several closely-related lineages whose partner fidelity, genome-wide expression and time of evolutionary divergence remain to be resolved. Here we detect and distinguish UCYN-A1 and UCYN-A2 lineages in symbiosis with two distinct prymnesiophyte partners in the South Atlantic Ocean. Both symbiotic systems are lineage specific and differ in the number of UCYN-A cells involved. Our analyses infer a streamlined genome expression towards nitrogen fixation in both UCYN-A lineages. Comparative genomics reveal a strong purifying selection in UCYN-A1 and UCYN-A2 with a diversification process about 91 Mya, in the late Cretaceous, after the low nutrient regime period occurred during the Jurassic. These findings suggest that UCYN-A diversified in a co-evolutionary process wherein their prymnesiophyte partners acted as a barrier driving an allopatric speciation of extant UCYN-A lineages.

Introduction

Symbiotic relationships involving diazotrophic microorganisms, i.e. those capable of converting dissolved dinitrogen gas into ammonia, are of relevant interest in marine biogeochemistry because they represent major sources of fixed nitrogen, a limiting nutrient for primary production in the world's oceans (Karl *et al.*, 2002). As such, identifying these interactions is essential for understanding the role of symbiosis in biogeochemical cycles. Fortunately, the application of novel approaches such as high-throughput sequencing and single-cell genomics has greatly accelerated the pace of microbial symbiosis research (McFall-Ngai, 2008; Martinez-Garcia *et al.*, 2012). This is notable in the case of *Candidatus* Atelocyanobacterium thalassa (UCYN-A), a unicellular diazotrophic cyanobacterium, and its partner, a single-celled eukaryotic alga of the Class Prymnesiophyceae (Thompson *et al.*, 2012). Prymnesiophytes as well as UCYN-A are abundant and widely distributed members of the marine plankton and represent ecologically relevant players in carbon and nitrogen cycles (Montoya *et al.*, 2004; Jardillier *et al.*, 2010; Goebel *et al.*, 2010; Zehr and Kudela, 2011; Cabello *et al.*, 2015). The streamlined genome of UCYN-A and the striking lack of genes encoding the Photosystem II complex, the Calvin/Benson/Bassham cycle for carbon fixation, as well as other essential pathways such as the TCA cycle, hinted at a symbiotic lifestyle (Zehr *et al.*, 2008; Tripp *et al.*, 2010; Bombar *et al.*, 2014). UCYN-A is now known to establish a mutualistic relationship based on the exchange of fixed carbon and nitrogen with two different cell-sized prymnesiophyte partners, the unicellular alga *Braarudosphaera bigelowii* (7-10 µm) (Hagino *et al.*, 2013; Thompson *et al.*, 2014) and an uncultured closely-related prymnesiophyte (1-3 µm) (Thompson *et al.*, 2012; Krupke *et al.*, 2013).

Phylogenomic analyses have demonstrated the monophyly of UCYN-A within the marine cyanobacteria clade that includes *Crocosphaera* sp. and *Cyanothece* sp. clades (Bombar *et al.*, 2014). Phylogenetic analysis of the UCYN-A nitrogenase gene (*nifH*) sequences, a common marker used to address the diversity of N_2-fixing microorganisms, distinguished at least three distinct UCYN-A clades:

UCYN-A1, UCYN-A2 and UCYN-A3 (Thompson *et al.*, 2014). Comparative genomics revealed that UCYN-A1 and UCYN-A2 lineages share largely syntenic genomic structures suggesting that both lineages diverged after genome reduction from a common ancestor (Bombar *et al.*, 2014). Yet, their time of evolutionary divergence and evolutionary pressures remain unknown. It has been suggested that these two variants could be adapted to different niches, i.e. coastal waters (*B. bigelowii*) and open ocean (its closely-related prymnesiophyte) (Thompson *et al.*, 2014), but this ecological differentiation was recently ruled out (Cabello *et al.*, 2015). Although the two prymnesiophyte partners could follow different ecological strategies (Cabello *et al.*, 2015), the partner fidelity has never been tested in this symbiotic system and therefore, we cannot assume a similar ecological niche for their symbionts. Comparative gene expression studies could help to disentangle the ecological distinction of these two UCYN-A lineages but they are scarce and solely focused on the *nifH* gene expression without showing a clear differentiation in lineage-specific patterns (Thompson *et al.*, 2014).

By designing and applying new probes in double CAtalyzed Reporter Deposition Fluorescence *In Situ* Hybridization (CARD-FISH), we identified the specific symbiotic associations at the UCYN-A lineage level in samples from South Atlantic waters from *Tara* Oceans expedition, where we had previously verified significant abundances of the prymnesiophyte partners. The new probes allowed us to differentiate both symbiotic systems which resulted to vary in the number of UCYN-A cells involved. The coupled analyses of metagenomes and metatranscriptomes from surface and DCM depths that encompassed four different plankton size fractions distinguish different prymnesiophyte partners based on difference in cell sizes captured in different size fractions, complementing and extending the results obtained by CARD-FISH. Gene expression was explored in the two UCYN-A lineages in order to decipher whether distinct lineages, in association with distinct partners, exhibit different expression patterns. Finally, we investigated the evolutionary pressures acting on UCYN-A1 and UCYN-A2 lineages by comparative genomic analyses and performed phylogenomic analyses to estimate the age divergence of the two symbiotic lineages. Our findings support a symbiont-host co-evolutionary scenario in the marine environment originated from a single ancestral symbiotic event in the late Cretaceous from which, at least, two different UCYN-A lineages diversified to become lineage-specific nitrogen fixation factories in their prymnesiophyte partners. Together, these investigations improve our understanding of the relevance of co-evolutionary processes in marine ecosystems and the ecological significance of N_2-fixing symbiosis in the marine biogeochemical cycles.

Material and methods

Sample choice

From a total of 243 metagenomes from 68 globally distributed stations from *Tara* Oceans expedition (Karsenti *et al.*, 2011), the abundance of UCYN-A based on 16S $_{\mathrm{mi}}$TAGs (Sunagawa *et al.*, 2015; Logares *et al.*, 2013) and their corresponding prymnesiophyte partners evaluated by V9 18S iTAGS (Cabello *et al.*, 2015; de Vargas *et al.*, 2015), pointed out to a couple of stations, i.e. TARA_078 (30º 8' 12.12" S 43º 17' 23.64" W) and TARA_076 (20º 56' 7.44" S 35º 10' 49.08" W) in the South Atlantic Ocean in which this symbiotic system were significantly abundant (Cabello *et al.*, 2015) and, therefore these two stations were chosen to further explore the UCYN-A symbiotic system.

Sample collection

For the whole-cell CARD-FISH, 10 mL of surface seawater (pre-filtered with 20-µm pore-size mesh) was fixed with paraformaldehyde (1.5% final concentration) at 4ºC overnight and gently filtered through 0.2 µm pore-size polycarbonate filters (Millipore, GTTP, 25 mm diameter). For nucleic acid extractions and sequencing, surface seawater was collected and subsequently separated into four size fractions (0.2-3, 0.8-5, 5-20 and >0.8 µm pore-size filters) (Pesant *et al.*, 2015). After filtration, filters were kept for approximately 4 weeks at -20°C on the schooner and then at -80°C in the laboratory until processed for hybridization or sequencing.

Design of CARD-FISH probes

For the design of specific oligonucleotide probes targeting *B. bigelowii* and the closely-related prymnesiophyte partner, a total of 580 sequences, 18S rRNA gene sequences, belonging to the class Prymnesiophyceae were retrieved from the PR2 database (Guillou *et al.*, 2013), aligned using MAFFT (Katoh *et al.*, 2002) and the alignment was verified manually to remove chimeras and sequences with ambiguities (466 sequences were kept). A maximum likelihood phylogenetic tree was built using RAxML (Stamatakis, 2006) with 100 trees for both topology and bootstrap analyses and visualized with iTol (Letunic and Bork, 2007, 2011) (Supplementary Fig. 1). The newly designed probe UBRADO69 targeted *B. bigelowii*, while probe UPRYM69 targeted the closely-related prymnesiophyte partner (Supplementary Table 1). UBRADO69 and UPRYM69 probes differed in only one position, and required a competitor in order to avoid unspecific hybridizations. Therefore, the labeled probe UBRADO69 was used in combination with the unlabeled UPRYM69 oligonucleotide for the detection of *B. bigelowii*, and *vice versa* for the detection of the closely-related prymnesiophyte partner (Supplementary Table 1). Two helpers, Helper-A PRYM and Helper-B PRYM, were designed to improve the hybridization process for both probes (Supplementary Table 1). The UCYN-A732 probe designed against UCYN-A by targeting the 16S rRNA (Krupke *et al.*, 2013) has only one mismatch with the UCYN-A2 sequence and a competitor was designed to distinguish specifically UCYN-A1 and UCYN-A2 clades with high specificity (Supplementary Table 1). The specificity of the new probes was checked with the online tool ProbeCheck (http://www.cme.msu.edu/RDP/) and by searching in the GenBank database (http://www.ncbi.nlm.nih.gov/index.html) to detect potential matching sequences in non-target groups.

CARD-FISH assay and epifluorescence microscopy

A preliminary double hybridization assay using the universal haptophyte PRYM02 probe (Simon *et al.*, 2000) and UCYN-A732 was first applied to check whether the partner of UCYN-A in our sample belong to Class Prymnesiophyceae. In order to specifically target the different UCYN-A lineages and their prymnesiophyte hosts, a double CARD-FISH assay was performed for each partnership (according to the Multi-color CARD-FISH protocol (Pernthaler *et al.*, 2004). For the first hybridization step, the specific probe for one of the prymnesiophyte partners (UBRADO69 or UPRYM69) was used and, for the second step, the UCYN-A732 probe was used. To check the specificity of symbiont pairs, an additional double CARD-FISH was done with the UBRADO69 probe and the UCYN-A732 as described before with the addition of the UCYN-A732 competitor to the hybridization buffer (probe, helpers and competitor at [0.16 ng μl^{-1}]). Filters were embedded in low-gelling-point agarose 0.1% (w/v) to minimize cell loss, and cell walls were permeabilized with lysozyme (37ºC, 1 h) and acromopeptidase solutions (37ºC, 0.5 h). For the first CARD-FISH step (described in more detail in Cabello *et al.*, 2015) filters were hybridized overnight at 46ºC in 40% formamide (FA) hybridization buffer containing a mixture of the HRP (Horseradish peroxidase)-labeled probe, helpers and competitor oligonucleotides. Filters were then rinsed in washing buffer at 48°C and tyramide signal amplification (TSA) was performed for 40 min at RT in the dark in a buffer containing 4 μg ml^{-1} Alexa 488-labelled tyramide. Before the second hybridization, the HRP from the first probe was inactivated with 0.01M HCl for 10 minutes at RT in the dark (Pernthaler *et al.*, 2004). The second CARD-FISH used the probe UCYN-A732 and its corresponding helpers and was applied according to Krupke *et al.,* 2013. UCYN-A cells were hybridized for 3 h at 35°C in 50% FA hybridization buffer, rinsed in washing buffer for 15 min at 37°C and TSA was done as before but using 1 μg ml^{-1} Alexa 594 -labeled tyramide. Preparations were counterstained with 4', 6- diamidino-2-phenylindole (DAPI) at 5 μg ml^{-1}, mounted in antifading reagent (77% glycerol, 15% VECTASHIELD, and 8% 20x PBS) and kept frozen until microscopic analysis. A no-probe control showed there was not signal coming from endogenous peroxidases. Filters were observed by epifluorescence microscopy (Olympus BX61) at 1000x under UV (DAPI signal of the nucleous), blue light (green labeled host cells with Alexa 488) or green light (red labeled symbionts with Alexa 594) excitations. Micrographs were taken with an Olympus DP72 camera (Olympus America Inc.) attached to the microscope.

Hybridization conditions for the UPRYM69 and UBRADO69 probes were optimized testing different FA concentrations in the hybridization buffer and varying the hybridization temperature. The UPRYM69 probe (together with the competitor oligonucleotide) was tested in NE Atlantic surface samples, where UCYN-A1 host cells were ~86% of prymnesiophytes (~550 cells ml^{-1}). Initially we tried 20-30-40-50% FA in the buffer and the temperature of 35ºC for hybridization. At 20% FA host cells carrying UCYN-A (n=89) displayed a faint fluorescent signal (90%) or were not labeled (10%), whereas above 40% FA no hybridized cells were detected. Signal was improved by using helper oligonucleotides and host cells displayed a bright homogeneous signal at all FA concentrations, but we observed cross-hybridization (observed as fluorescent dots all over the cells) in larger prymnesiophyte-like cells not associated to UCYN-A even at 50% FA. Thus, we tested 40 and 50% FA in a hybridization temperature of 46ºC. The 40% FA showed optimal signal intensity, labeling small prymnesiophytes cells (about 2.5 μm) always carrying UCYN-A and no cross-hybridization was observed. We applied these conditions to hybridize the surface sample TARA_078. In this sample, in addition to the labeled small host cells observed in the NE Atlantic, we observed larger host cells not labeled by the UPRYM69 probe. To verify that these cells were the UCYN-A2 host we applied the UBRADO69 probe with the same conditions (as both probes differ in only 1 position) and we found the complementary result: the larger host was labeled but not the smaller one. With the optimized conditions (40% FA, 46ºC) the probes were labeling specifically the target host without cross-hybridization.

Nucleic acid extractions and sequencing

Surface and DCM seawater samples collected by *Tara* Oceans' station 76 and 78 in the South Atlantic Ocean (TARA_076, TARA_078) for metagenomic sequencing were size-fractionated. For surface samples, metagenomes from two and four fractions were analyzed in TARA_076 (0.2-3 μm and >0.8 μm) and TARA_078 (0.2-3 μm, 0.8-5 μm, 5-20 μm and >0.8 μm) respectively. For DCM samples, metagenomes from one fraction were analyzed in TARA_076 (>0.8 μm) and TARA_078 (>0.8 μm). Seawater samples for metatranscriptomic sequencing used also several size fractions. For surface samples, metatranscriptomes from two and three fractions were analyzed in TARA_076 (0.2-3 μm and >0.8 μm) and TARA_078 (0.2-3 μm, 5-20 μm and >0.8 μm) respectively. For DCM samples, metatranscriptomes from one fraction were analyzed in TARA_076 (>0.8 μm) and TARA_078 (>0.8 μm). DNA and RNA extraction protocols for the different size fractions and metagenome sequencing were already described in previous studies (Sunagawa *et al.*, 2015; Logares *et al.*, 2013; de Vargas *et al.*, 2015).

cDNA synthesis and sequencing

For 0.2-3 μm and >0.8 μm filters, bacterial rRNA depletion was carried out on 240-500 ng total RNA using Ribo-Zero Magnetic Kit for Bacteria (Epicentre, Madison, WI). The Ribo-Zero depletion protocol was modified to be adapted to low RNA input amounts (Alberti *et al.*, 2014). Depleted RNA was used to synthetize cDNA with SMARTer Stranded RNA-Seq Kit (Clontech, Mountain View, CA) (Alberti *et al.*, 2014). For 5-20 μm filter from TARA_078, cDNA was synthetized starting from 50 ng total RNA using SMARTer Ultra Low RNA Kit (Clontech) by oligodT priming following the manufacturer protocol. Full length double stranded cDNA was fragmented to a 150-600-bp size range using the E210 Covaris instrument (Covaris Inc., USA). Then, fragments were end-repaired and 3'-adenylated, and ligated to Illumina adapters by using NEBNext Sample Reagent Set (New England Biolabs, Ipswich, MA). Fragments were PCR-amplified using Illumina adapter specific primers and purified. All metatranscriptomic libraries were quantified by qPCR using the KAPA Library Quantification Kit for Illumina Libraries (KapaBiosystems, Wilmington, MA) and library profiles were assessed using the DNA High Sensitivity LabChip kit on an Agilent Bioanalyzer (Agilent Technologies, Santa Clara, CA). Libraries were sequenced on Illumina HiSeq2000 instrument (Illumina, San Diego,CA) using 100 base-length read chemistry in a paired-end mode. Sequencing depth for each sample is detailed in Table 1.

Nucleotide data deposition

Nucleotides data used in this study have been deposited in the European Nucleotide Archive (ENA) (www.ebi.ac.uk/ena) under the following accession numbers: ERR1001626-27, ERR1007415-18, ERR1013384-85, ERR599006, ERR599010, ERR599022, ERR599126, ERR599237, ERR599240, ERR599250, ERR599253, ERR599275 and ERR599311.

Fragment recruitment analysis from -omics datasets

BLAST+ v2.2.25 was used to recruit metagenomic and metatranscriptomic reads similar to the two UCYN-A genomes sequenced up to date (Tripp *et al.*, 2010; Bombar *et al.*, 2014) using default parameter values, except for the following: -perc_identity 50, -evalue 0.0001. Metagenomic/metatranscriptomic reads belonging to 23S, 16S and 5S rRNA genes or ITS regions as well as those aligned along less than 90% of its length were excluded (Table 1). The genome recovery was calculated as the percentage of nucleotide positions within the reference genomes aligned with metagenomic or metatranscriptomics reads higher than 95% identity, threshold used for representing members of the same population as the reference genome (Caro-quintero and Konstantinidis, 2011) (Table 1). To assess the gene expression at the genome level, we first used the gene positions to count the number of metatranscripts covering each gene. Then, we normalized these counts using two approaches (i) by using UCYN-A single copy house-keeping genes (*recA* and *gyrB* metatranscript counts), and (ii) by metagenomic read counts for each UCYN-A gene (in this case we also normalized by sequencing depth) (Supplementary Data 1 and 2).

Phylogenomic and relaxed molecular clock analyses

Sequence data for 57 cyanobacterial genomes were used to estimate the phylogenetic relationships of UCYN-A1 and UCYN-A2. We analyzed 135 protein sequences that have shown to be highly conserved, to have undergone a minimum number of gene duplications and also to represent a wide diversity of cellular functions (Blank and Sánchez-Baracaldo, 2010). Maximum likelihood analyses and bootstrap values were performed using RAxML 7.4.2 (Stamatakis, 2006). Bayesian relaxed molecular clock analyses as implemented in MCMCtree (Yang, 2007) and PhyloBayes 3.3b (Lartillot *et al.*, 2009) were performed to estimate divergence times of UCYN-A1 and UCYN-A2 (Supplementary Table 2). We applied the Uncorrelated Gamma Multipliers (UGM) model (Drummond *et al.*, 2006) as this model seems to fit better cyanobacteria nucleotide data sets based on Bayes Factors (Sánchez-Baracaldo *et al.*, 2014). Age divergences for UCYN-A1 and UCYN-A2 were estimated based on three genes: LSU (3002 characters), SSU (1546 characters) and rpoC1 (1887 characters). In PhyloBayes (Lartillot *et al.*, 2009), we implemented the CAT-GT replacement model of nucleotide evolution. For all non-calibrated nodes, we used a birth-death prior (Lepage *et al.*, 2007) on divergence times. A permissive gamma distributed root prior of 2,500 million years ago (Mya) was also implemented (SD = 200 Mya, which allowed the 95% credibility interval of the root node to range between 2,300 and 2,700 Mya). We treated all calibrations as soft allowing for 2.5 % on each side for an upper and lower bound. In MCMCTree, LSU, SSU and rpoC1 were treated as separate loci and branch lengths were estimated in BASEML (Yang, 2007). We used the HKY85 (Hasegawa *et al.*, 1985) model of nucleotide evolution based on Bayes factor analyses (Sánchez-Baracaldo *et al.*, 2014). We used 1 billion years per unit time for all analyses. The gamma prior G (α and β) used to describe how variable rates are across branches was specified as follows G (1, 7). The mean and standard deviation was specified as m $=\alpha/\beta$. The gamma priors for the substitution model parameters κ (transition/transversion rate ratio) and α (gamma shape parameter for variable rates among sites) were all specified by gamma distributions. Respective means and standard deviations were (6, 2) for κ and (1, 1) for α. For all analyses, we used fixed values for the birth-death process $\lambda = \mu = 1$ and $\rho = 0$. Analyses were performed at least twice to ensure convergence of the MCMC, although only one analysis is reported. For all age calibrations, both minimum and maximum bounds were soft and specified by uniform distributions between the maximum/minimum time constraints with 2.5% tail probabilities above/below these limits allowing for molecular data to correct for conflicting fossil information (Yang and Rannala, 2006). To check whether analyses had converged we used Tracer v1.5.0 (http://beast.bio.ed.ac.uk/Tracer). For the cyanobacterial root, 2,700 Mya (Brocks *et al.*, 2003) and 2,320 Mya (Bekker *et al.*, 2004) (the rise in atmospheric oxygen) were set as the maximum and minimum age respectively. Other fossils exhibiting unique morphological features were assigned to well-supported groups such as the Nostocales (Tomitani *et al.*, 2006) and the clade containing two *Pleurocapsa* genomes (PCC 7319 and PCC 7327) in the Pleurocapsales (Zhang, Y. and Golubic, 1987).

On-line supplementary information

Supplementary Data 1, Supplementary Data 2 and Supplementary Data 3 can be downloaded from the following web site: https://www.nature.com/articles/ncomms11071#supplementary-information

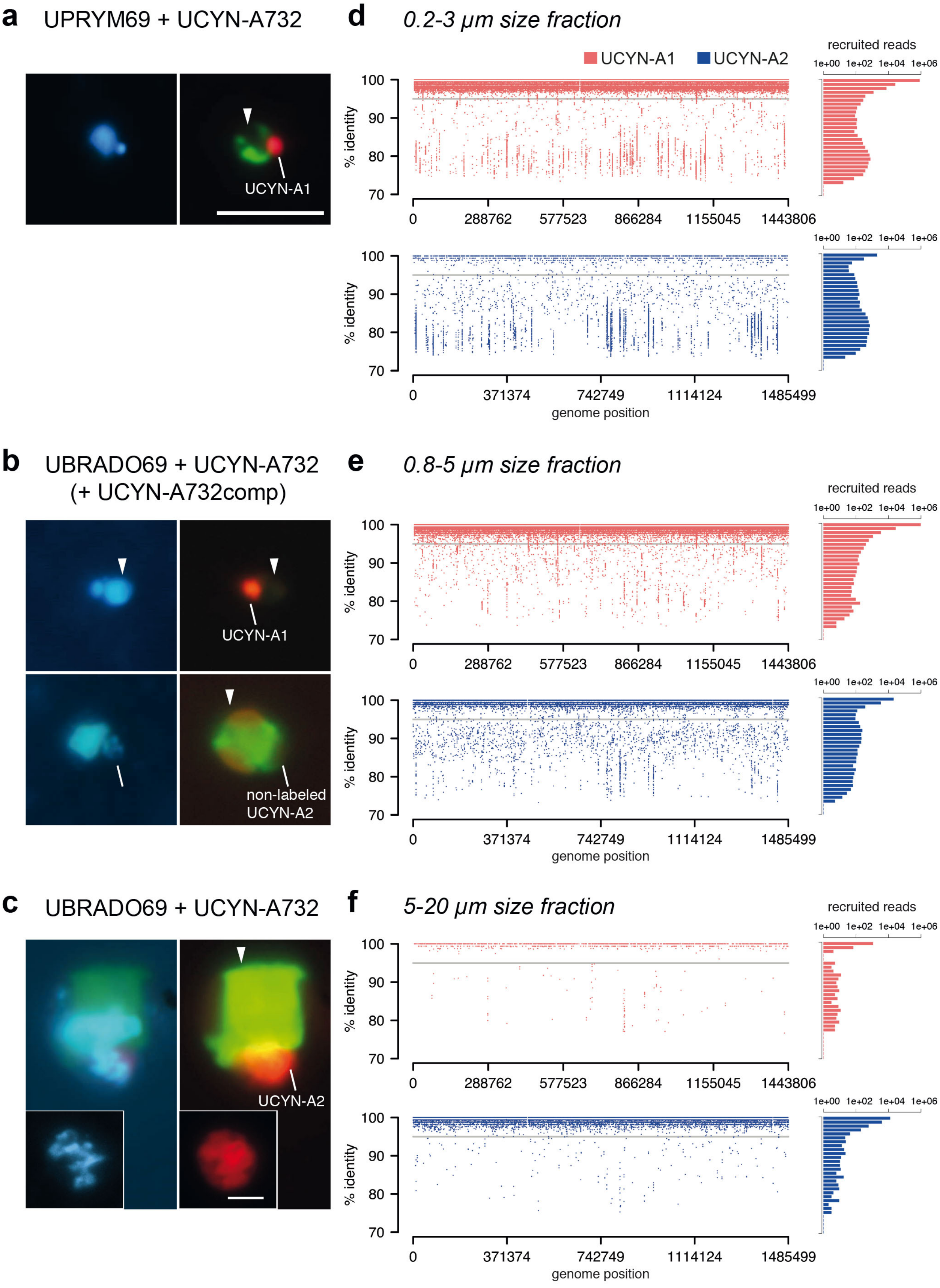
Results and discussion

Partner fidelity of two UCYN-A lineages

Microscopic in situ identification of different UCYN-A lineages as well as their prymnesiophyte partners by specific CARD-FISH staining is crucial to determine the specificity of their relationships. The CARD-FISH method has been successfully applied to identify unicellular diazotrophic cyanobacteria (Le Moal *et al.*, 2011) as well as specifically targeting the UCYN-A clade (Krupke *et al.*, 2013, 2014). However, to our knowledge there was not any reported probe to distinguished UCYN-A at the lineage level. We designed a competitor probe to be used with the UCYN-A732 probe (Krupke *et al.*, 2013) to distinguished UCYN-A1 and UCYN-A2 lineages (Fig. 1a-c, Supplementary Table 1). Similarly, we designed two probes to distinguish the two prymnesiophyte partners, *B. bigelowii* (UBRADO69 probe) and the closely-related prymnesiophyte (UPRYM69 probe)

(Fig. 1a-c, Supplementary Table 1). The UCYN-A732 probe, in the absence of its competitor, labeled UCYN-A cells inside either *B. bigelowii* or the closely-related prymnesiophyte partner (Fig. 1a,c). However, when the UBRADO69 probe was applied with the UCYN-A732 probe together with its competitor, UCYN-A cells were unlabeled or labeled when accompanying *B. bigelowii* or the closely-related prymnesiophyte partner respectively (Fig. 1b). It has been proposed that smaller UCYN-A cells are associated with smaller prymnesiophyte cells and vice versa, indicating different growth stages (Krupke *et al.*, 2014). However, those findings were interpreted from microscopic observations of the UCYN-A symbiosis detected with the general prymnesiophyte PRYM02 and UCYN-A732 (without its competitor) probes, i.e. without the ability to distinguish UCYN-A1 and UCYN-A2 cells. The results presented here show that both prymnesiophyte partners are phylogenetically closely-related but distinct species, and therefore, we suggest that the observed differences in cell sizes of prymnesiophyte partners reflect distinct species rather than different growth stages of the same species. These results demonstrate that UCYN-A lineages display partner fidelity with their prymnesiophyte partners, being *B. bigelowii* and the closely-related prymnesiophyte in specific association with UCYN-A2 and UCYN-A1 lineages respectively.

Figure 1. Partner specificity and variation of UCYN-A lineages with plankton size fraction. (**a-c**) Epifluorescence microscopy images with the double CARD-FISH assay showing the specificity of symbiont-host pairs and (**d-f**) fragment recruitment of UCYN-A lineages in size fractionated metagenomes from surface waters collected in station TARA_078. (**a-c**) Left panels correspond to the DAPI signal (blue-labeled DNA); right panels correspond to the combined signal of the prymnesiophyte specific probes (green-labeled host under blue light excitation) and the UCYN-A probe (red-labeled symbiont under green-light excitation). (**a**) UCYN-A1 with its prymnesiophyte partner; (**b**) the two UCYN-A symbiotic pairs, indicating the specific labeling of UCYN-A1 (upper) and *B. bigelowii* (lower) with their specific partners, the small prymnesiophyte closely-related to *B. bigelowii* and UCYN-A2 respectively; (**c**) *B. bigelowii* with UCYN-A2. The inset panel in (**c**) shows the detail of non-associated UCYN-A2 cells within a common symbiotic structure. Prymnesiophyte partners are indicated by arrow heads. Scale bar in (**a**) represents 5 μm and this scale is shared in (**a-c**) except in the inset of panel (**c**) where it indicates 2 μm. (**d-f**) On the left side, recruitment of metagenomic reads using UCYN-A1 and UCYN-A2 genomes as reference. Reads are plotted as red (UCYN-A1) or blue (UCYN-A2) dots depending on the closest hit genome, representing the covered genome positions (x axis) and the % of identity with the closest reference (y axis). A horizontal gray line set at 95% indicates the threshold for reads representing members of the same population as the reference genome. On the right side, histograms represent the number of recruited reads, in logarithmic scale, by UCYN-A1 (red) or UCYN-A2 (blue) genomes in intervals of 1% identity, from 100% to 70% identity.


a
UPRYM69 + UCYN-A732
UCYN-A1
b
UBRADO69 + UCYN-A732
(+ UCYN-A732comp)
UCYN-A1
non-labeled
UCYN-A2
c
UBRADO69 + UCYN-A732
UCYN-A2
d
0.2-3 µm size fraction
🟥 UCYN-A1
🟦 UCYN-A2
recruited reads
1e+00 1e+02 1e+04 1e+06
% identity
100
90
80
70
0
288762
577523
866284
1155045
1443806
0
371374
742749
1114124
1485499
genome position
e
0.8-5 µm size fraction
recruited reads
1e+00 1e+02 1e+04 1e+06
% identity
100
90
80
70
0
288762
577523
866284
1155045
1443806
0
371374
742749
1114124
1485499
genome position
f
5-20 µm size fraction
recruited reads
1e+00 1e+02 1e+04 1e+06
% identity
100
90
80
70
0
288762
577523
866284
1155045
1443806
0
371374
742749
1114124
1485499
genome position

The number of UCYN-A cells per partner is lineage-specific

Previous studies have shown that the prymnesiophyte partners can harbor one or two UCYN-A cells (Thompson *et al.*, 2012; Cabello *et al.*, 2015; Hagino *et al.*, 2013; Krupke *et al.*, 2013), pointing to a coupling between the prymnesiophyte cell division and the number of symbiotic cells, at least for UCYN-A1 (Cabello *et al.*, 2015). In our samples, only one UCYN-A1 cell per prymnesiophyte cell was detected (Fig. 1a,b). By contrast, *B. bigelowii* carried a symbiosome-like compartment with a variable but higher number of UCYN-A2 cells (~3-10 cells) (Fig. 1b,c). This structure was observed both attached to the host and in a free state, as an entity composed by several UCYN-A2 cells enclosed by a common envelope (Fig. 1c). In a previous study, the UCYN-A2 cells found in *B. bigelowii* were separated from the *B. bigelowii* cytoplasm by a single membrane, likely a perisymbiont membrane, and the envelope of the UCYN-A2 itself consisted of three layers, possibly an outer membrane, a peptidoglycan wall and a plasma membrane (Hagino *et al.*, 2013). Although UCYN-A1 and UCYN-A2 are very similar in terms of gene content, the genes involved in cell wall biogenesis and cell shape determination appear to be only present in UCYN-A2 suggesting clear structural differences associated with its host (Bombar *et al.*, 2014). Therefore, our observations hint at different symbiotic organizations: while the UCYN-A1 lineage has one or two separated cells per host, the UCYN-A2 lineage may harbour up to 10 cells per prymnesiophyte partner cell within a common symbiotic structure.

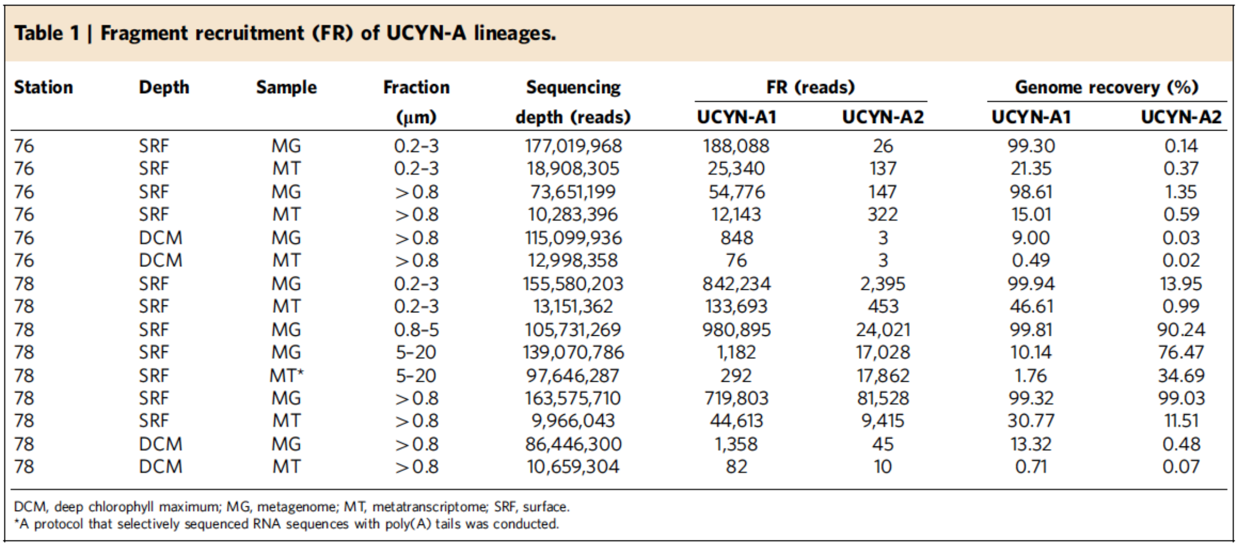


Table 1 | Fragment recruitment (FR) of UCYN-A lineages.

Station	Depth	Sample	Fraction	Sequencing	FR (reads)		Genome recovery (%)	
			(μm)	depth (reads)	UCYN-A1	UCYN-A2	UCYN-A1	UCYN-A2
76	SRF	MG	0.2-3	177,019,968	188,088	26	99.30	0.14
76	SRF	MT	0.2-3	18,908,305	25,340	137	21.35	0.37
76	SRF	MG	>0.8	73,651,199	54,776	147	98.61	1.35
76	SRF	MT	>0.8	10,283,396	12,143	322	15.01	0.59
76	DCM	MG	>0.8	115,099,936	848	3	9.00	0.03
76	DCM	MT	>0.8	12,998,358	76	3	0.49	0.02
78	SRF	MG	0.2-3	155,580,203	842,234	2,395	99.94	13.95
78	SRF	MT	0.2-3	13,151,362	133,693	453	46.61	0.99
78	SRF	MG	0.8-5	105,731,269	980,895	24,021	99.81	90.24
78	SRF	MG	5-20	139,070,786	1,182	17,028	10.14	76.47
78	SRF	MT*	5-20	97,646,287	292	17,862	1.76	34.69
78	SRF	MG	>0.8	163,575,710	719,803	81,528	99.32	99.03
78	SRF	MT	>0.8	9,966,043	44,613	9,415	30.77	11.51
78	DCM	MG	>0.8	86,446,300	1,358	45	13.32	0.48
78	DCM	MT	>0.8	10,659,304	82	10	0.71	0.07

DCM, deep chlorophyll maximum; MG, metagenome; MT, metatranscriptome; SRF, surface.
*A protocol that selectively sequenced RNA sequences with poly(A) tails was conducted.

UCYN-A lineages vary in different plankton size fractions

A total of eight marine metagenomes from stations TARA_078 and TARA_076 were analyzed to assess the distribution of UCYN-A lineages in several plankton size fractions (0.2-3, 0.8-5, 5-20 and >0.8 μm) of the microbial assemblages in surface and Deep Chlorophyll Maximum (DCM) waters (Table 1). We used the two UCYN-A genomes sequenced to-date as reference genomes (Tripp *et al.*, 2010; Bombar *et al.*, 2014) in fragment recruitment of these metagenomic samples (Table 1). Because of the UCYN-A partner fidelity displayed by double CARD-FISH (see above), metagenomic sequence reads from UCYN-A lineages should vary with size fraction as predicted by the different cell-sizes of the prymnesiophyte partners. The sequence reads from the UCYN-A1 lineage were primarily present in surface waters within the size-fraction range of the small prymnesiophyte partner (0.2-3, 0.8-5 and >0.8 μm) (Table 1). Almost 100% of the UCYN-A1 genome was recovered in each of the metagenomes from surface of these size fractions in the two stations. Likewise, UCYN-A1 sequence reads were poorly represented in the 5 to 20 μm size fraction (~10% of genome recovery) (Fig. 1d-f, Table 1). On the other hand, in TARA_078, the UCYN-A2 sequence read distribution in surface waters was consistent with the *B. bigelowii* cell size, i.e. UCYN-A2 reads were nearly absent in the 0.2 to 3 μm size fraction metagenomes but were more abundant in the 0.8-5, 5-20 and >0.8 μm fractions. In all these larger fractions, the UCYN-A2 reached high genome recovery values (90%, 76% and 99%, respectively) except for the >0.8 μm fraction in TARA_076 where UCYN-A2 was virtually absent (Fig. 1d-f, Table 1). In the >0.8 μm size fraction, UCYN-A1 was approximately 9 times more abundant than UCYN-A2 in TARA_078 (Table 1). Likewise, in same station the small prymnesiophyte partner was more abundant than *B. bigelowii* based on V9 18S rRNA tags (Cabello *et al.*, 2015). In the DCM samples, both UCYN-A lineages were poorly represented in the metagenome sequences, accounting for less than 14% and 1% of genome recovery for UCYN-A1 and UCYN-A2, respectively (Table 1). The same vertical distribution has been observed for their prymnesiophyte partners that were found preferentially in surface layers while the rest of the prymnesiophyte assemblage peaked at the DCM (Cabello *et al.*, 2015). Therefore, although the

UCYN-A1 lineage was in general more abundant than UCYN-A2, a transition from the UCYN-A1 to UCYN-A2 lineage was observed from smaller to larger size fractions, likely explained by the partner fidelity and the difference in cell size of their prymnesiophyte partners.

Another interesting finding was that most of the metagenomic (and metatranscriptomic) reads mapping to the UCYN-A1 or UCYN-A2 genomes had very high sequence identities (greater than 99% to their respective reference genome) (Fig. 1d-f), which suggests an extremely low microdiversity within populations that were sampled from geographically distant regions in the Pacific (ALOHA and SIO) and South Atlantic Oceans (this study). The size-fractionated sampling strategy combined with the metagenomic analyses reported in this study will be also important to uncover the genomic pool of new UCYN-A lineages, such us UCYN-A3, to identify the lineage-specific distribution of UCYN-A populations and to set the cell size range of their partners, a first step for their identification.

UCYN-A expression is streamlined to fuel nitrogen fixation

The analyses of seven size-fractionated metatranscriptomes from two stations (TARA_078 and TARA_076) and depths (surface and DCM) allowed for the first time a whole-genome transcription profiling of these widely distributed diazotrophic cyanobacteria (Table 1). In surface waters, UCYN-A1 transcripts were in general more abundant than those from UCYN-A2, except in the 5-20 μm size fraction (TARA_078) in which the latter were dominant (Table 1). The gene expression of 1131 and 1179 protein-coding genes in UCYN-A1 (Supplementary Data 1) and UCYN-A2 (Supplementary Data 2), respectively, were examined. In both lineages, the nitrogen fixation operon, including the *nifH* gene, was the most highly expressed gene-cluster accounting for a quarter of the total transcripts (Fig. 2a,b). In the >0.8 μm size fraction (TARA_078), despite UCYN-A1 being more abundant than UCYN-A2, the expressed *nifH* transcripts per cell were almost 2 times higher for UCYN-A2 (648.33) than for UCYN-A1 (396.60) (Supplementary Data 1 and 2).

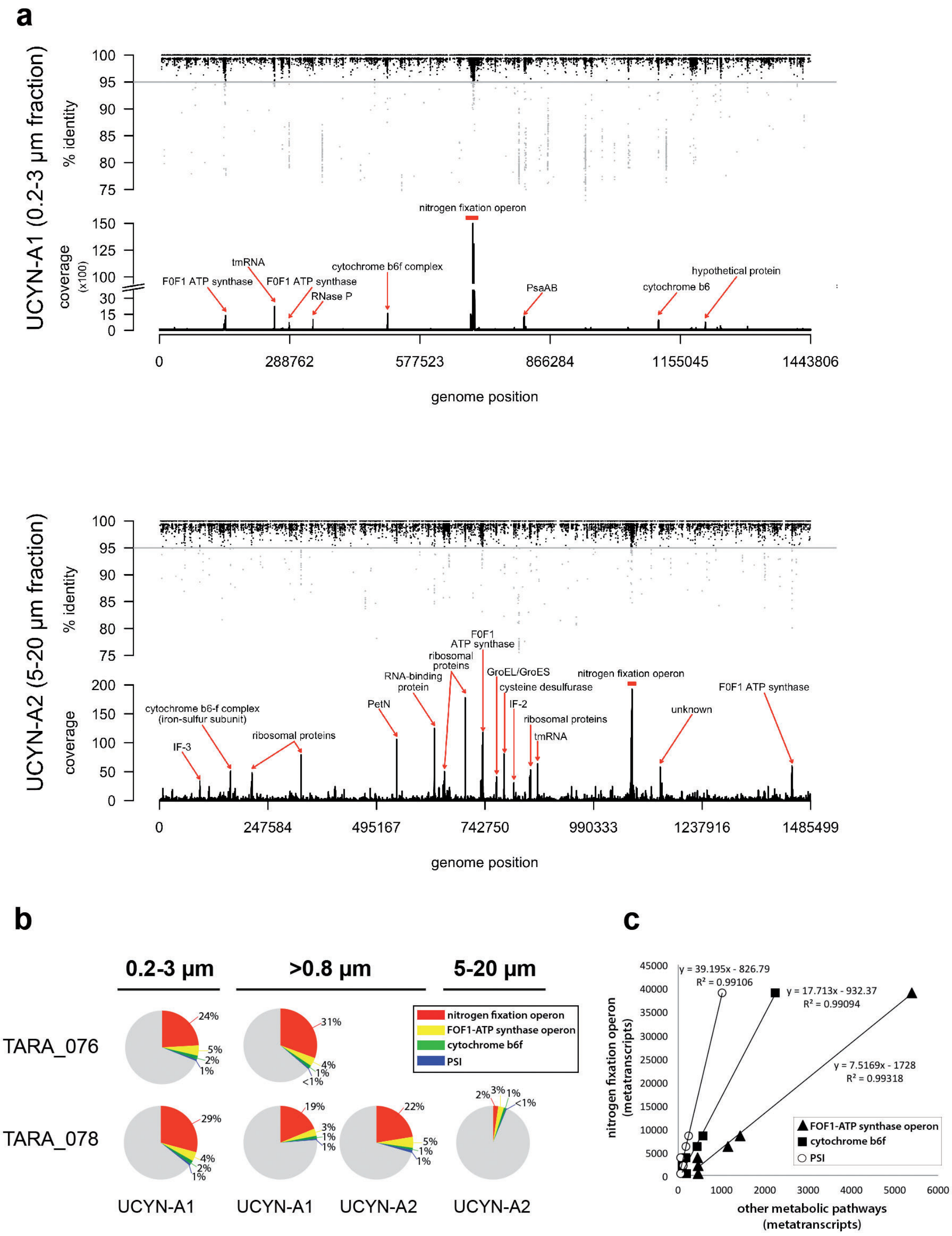


Figure 2. Genome expression in UCYN-A1 and UCYN-A2 lineages. (**a**) Metatranscriptome recruitment at the surface of the TARA_078 station of UCYN-A1 (0.2-3 μm) and UCYN-A2 (5-20 μm) transcripts. Transcripts are plotted as black dots representing the covered genome positions and the % of identity with the closest reference. A horizontal gray line set at 95% identity shows the threshold used to count the number of times, or coverage, that a gene was expressed. The most expressed genes in both lineages are highlighted. (**b**) Relative contribution of nitrogen fixation operon, FOF1-ATP synthase operon, cytochrome b_6f and PSI genes to the total UCYN-A transcripts contribution in surface samples; percentages are indicated. (**c**) Transcript counts of nitrogen fixation operon versus those of ATP synthase (triangle), cytochrome b_6f (square) and PSI (open circle) transcripts. All of these transcripts were significantly correlated ($P<10^{-5}$) and regression lines, regression equations and R^2 values are indicated in the figure.

It is well known that biological nitrogen fixation has a high energetic cost (16 moles of ATP to generate 2 moles of ammonia). Notably, the F0F1-ATP synthase operon and genes encoding for the cytochrome b_6f complex and Photosystem I complex (PSI) were highly transcribed and positively correlated ($P<10^{-5}$, N = 6, linear regression analysis) with the nitrogen fixation operon transcript abundances (Fig. 2c). These findings suggest that the generation of reducing power and the ATP synthesis could be coupled to fuel the nitrogen fixation process in UCYN-A. Likewise, UCYN-A2 might have higher nitrogen fixation rates per cell than UCYN-A1 based on the higher number of *nifH* transcripts per cell. It is reasonable to assume that the differences in *nifH* gene expression between UCYN-A lineages could simply reflect the differences in the cell size of their partners with differential nutrient requirements for growth. In addition, it has been indirectly demonstrated that the nitrogen fixation of UCYN-A supports the CO_2 fixation of its prymnesiophyte partner (Krupke *et al.*, 2015). Therefore, we hypothesize that the larger *B. bigelowii* host cell would meet its larger N nutrient requirements by partnering with a larger number of UCYN-A2 symbiotic cells.

Nitrogen-fixing microorganisms, and particularly cyanobacteria, should protect their nitrogenase from inactivation by oxygen. The absence of the ability to use photosystem II which evolves O_2 explains why UCYN- A appears to fix N_2 and express the nitrogenase genes during the day (Zehr, 2011). However, its association with an oxygen-evolving partner could make the nitrogenase enzyme in UCYN-A not completely safe from oxygen. We observed that the *sufB* gene (cysteine desulferase), involved in the assembly or repair of oxygen-labile iron-sulfur clusters under oxidative stress, was highly transcribed (Supplementary Data 1 and 2). It may be that UCYN-A requires high expression level of *sufB* genes to repair the nitrogenase enzyme from oxygenic inactivation, suggesting then a similar role than for the peroxidase genes found in their genomes (Tripp *et al.*, 2010; Bombar *et al.*, 2014). Our findings reveal that UCYN-A lineages dedicate a large transcriptional investment to fix nitrogen representing the first whole-genome expression profiling in environmental UCYN-A populations.

UCYN-A diverged during the late Cretaceous

Our findings on partner fidelity in UCYN-A point to the hypothesis of symbiont-host co-evolution (Thompson *et al.*, 2014). In order to analyze the selection pressure and evolution of the protein-coding genes, we calculated the number of synonymous or silent (Ks) and non-synonymous (Ka, inducing amino acid change) nucleotide substitutions (Li, 1993; Hurst, 2002) for 887 protein-coding genes shared by the UCYN-A1 and UCYN-A2 genomes (Supplementary Data 3). The Ka/Ks ratio may offer important clues about the selection pressure where ratios <1 indicate purifying selection and ratios >1 point to positive selection (McDonald and Kreitman, 1991). We found that 873 out of the 887 protein-coding genes were under purifying selection (*P*<0.05, Codon Based Z-test) (Supplementary Data 3). The 14 remaining genes also presented Ka/Ks < 1 but were not statistically well-supported (*P*>0.05). Purifying selection means that synonymous mutations are maintained, while non-synonymous mutations are continuously removed from the population. We did not detect signs of large-scale positive selection, i.e. no apparent strong adaptation to novel niches in UCYN-A lineages, suggesting that the evolutionary forces for niche adaptation would act on the prymnesiophyte partners rather than on UCYN-A. Our results are consistent with the fact that UCYN-A2 lacks the same major pathways and proteins that are absent in UCYN-A1 (Bombar *et al.*, 2014), indicating then that the symbionts were genetically adapted to their hosts before they were separated by speciation.

The age of divergence for UCYN-A1 and UCYN-A2 lineages was calculated by phylogenomic and Bayesian relaxed molecular clock analyses (Fig. 3, Supplementary Table 2). Our results indicate that UCYN-A1 and UCYN-A2 lineages diverged around 91 million years ago (Mya), i.e. during the late Cretaceous. In agreement, *B. bigelowii* has a fossil record extending back to the late Cretaceous (ca. 100 Mya) (Bown *et al.*, 2004), reported from neritic and pelagic sediments, e.g., in lower Paleogene sediments immediately above the K/Pg mass extinction level as well as in the Oligocene Diversity Minimum (Peleo-Alampay *et al.*, 1999; Bown *et al.*, 2004). In the Jurassic, between 190 and 100 Mya, nutrient availability in the ocean was lower than at any point during the last 550 Mya (Cárdenas and Harries, 2010). It is therefore likely that the symbiotic relationship between the common ancestor of UCYN-A1 and UCYN-A2 and a *Braarudosphaera*-related species was established by the late Cretaceous to cope with extremely low nutrient conditions and a generalized oligotrophy in marine surface waters, as it has been recognized for other symbiotic system such as the Acantharia–*Phaeocystis* symbiosis (Decelle *et al.*, 2012). UCYN-A then underwent purifying selection, progressively reducing its genome to the point that it became an obligate symbiont. An analogous discovery was the case of the two Rhopalodiaceae freshwater diatom species, *Rhopalodia gibba* and *Epithemia turgida* having acquired N_2-fixing endosymbionts (Kneip *et al.*, 2007; Nakayama *et al.*, 2011). Similar to the two UCYN-A partnerships described here, phylogenies of these two diatoms species and their intracellular symbionts were found to be congruent and, concordantly, a single symbiotic event has been proposed (Nakayama *et al.*, 2011). Probably, a similar scenario can be envisioned here for the two UCYN-A partnerships.

Taking into account that the number of symbiotic cells harbored by distinct prymnesiophyte partners is different and phylogenetically-dependent, i.e. the larger *B. bigelowii* can harbor a variable number (up to 10) of UCYN-A2 cells whilst the small prymnesiophyte partner harbored only one or two UCYN-A1 cells, it is reasonable to think that a larger nutrient acquisition could be linked to a larger number of symbionts. Indeed the whole genome expression patterns suggested a metabolic investment in UCYN-A1 and UCYN-A2 is mainly focused on the nitrogen fixation machinery. Our evolutionary analysis revealed that UCYN-A1 and UCYN-A2 were genetically adapted to their prymnesiophyte partners before UCYN-A speciation (purifying selection) but, on the contrary, the prymnesiophyte partners seem to follow different ecological strategies (Cabello *et al.*, 2015), suggesting a speciation process under positive selection. Our results suggest that the partner fidelity shown by UCYN-A lineages together with the speciation in the common ancestor of *B. bigelowii* and its closely-related prymnesiophyte may have forced an allopatric speciation of UCYN-A1 and UCYN-A2 populations in the late Cretaceous. Comparative genome analysis of the two prymnesiophyte partners would clarify whether these two algal species underwent positive selection through evolution by adaptation to novel niches. As revealed by *nifH* phylogenetic analysis it seems that novel UCYN-A lineages, such us UCYN-A3, and prymnesiophyte (or not prymnesiophyte) partners, will help to understand the evolutionary relationships of N_2-fixing cyanobacterial symbionts and the extent of their ecological relevance on marine biogeochemical cycles.

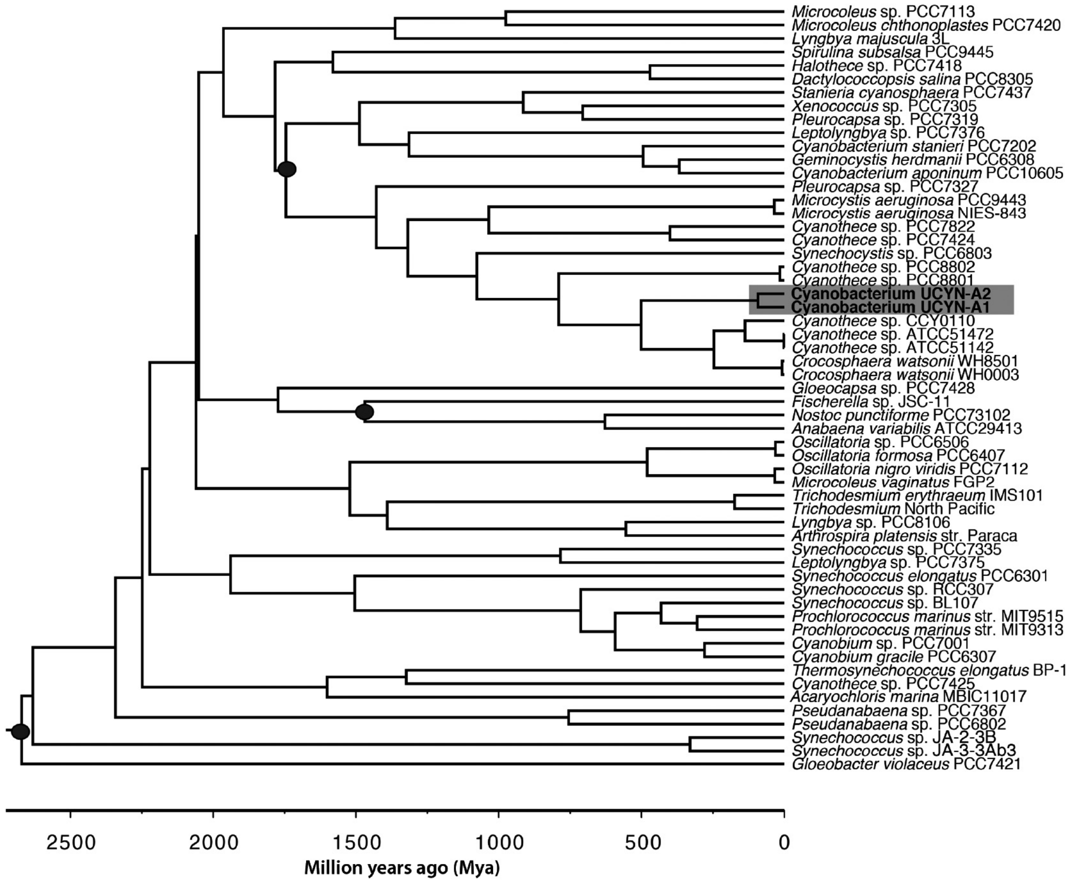

Microcoleus sp. PCC7113
Microcoleus chthonoplastes PCC7420
Lyngbya majuscula 3L
Spirulina subsalsa PCC9445
Halothece sp. PCC7418
Dactylococcopsis salina PCC8305
Stanieria cyanosphaera PCC7437
Xenococcus sp. PCC7305
Pleurocapsa sp. PCC7319
Leptolyngbya sp. PCC7376
Cyanobacterium stanieri PCC7202
Geminocystis herdmanii PCC6308
Cyanobacterium aponinum PCC10605
Pleurocapsa sp. PCC7327
Microcystis aeruginosa PCC9443
Microcystis aeruginosa NIES-843
Cyanothece sp. PCC7822
Cyanothece sp. PCC7424
Synechocystis sp. PCC6803
Cyanothece sp. PCC8802
Cyanothece sp. PCC8801
Cyanobacterium UCYN-A2
Cyanobacterium UCYN-A1
Cyanothece sp. CCY0110
Cyanothece sp. ATCC51472
Cyanothece sp. ATCC51142
Crocosphaera watsonii WH8501
Crocosphaera watsonii WH0003
Gloeocapsa sp. PCC7428
Fischerella sp. JSC-11
Nostoc punctiforme PCC73102
Anabaena variabilis ATCC29413
Oscillatoria sp. PCC6506
Oscillatoria formosa PCC6407
Oscillatoria nigro viridis PCC7112
Microcoleus vaginatus FGP2
Trichodesmium erythraeum IMS101
Trichodesmium North Pacific
Lyngbya sp. PCC8106
Arthrospira platensis str. Paraca
Synechococcus sp. PCC7335
Leptolyngbya sp. PCC7375
Synechococcus elongatus PCC6301
Synechococcus sp. RCC307
Synechococcus sp. BL107
Prochlorococcus marinus str. MIT9515
Prochlorococcus marinus str. MIT9313
Cyanobium sp. PCC7001
Cyanobium gracile PCC6307
Thermosynechococcus elongatus BP-1
Cyanothece sp. PCC7425
Acaryochloris marina MBIC11017
Pseudanabaena sp. PCC7367
Pseudanabaena sp. PCC6802
Synechococcus sp. JA-2-3B
Synechococcus sp. JA-3-3Ab3
Gloeobacter violaceus PCC7421
2500
2000
1500
1000
500
0
Million years ago (Mya)

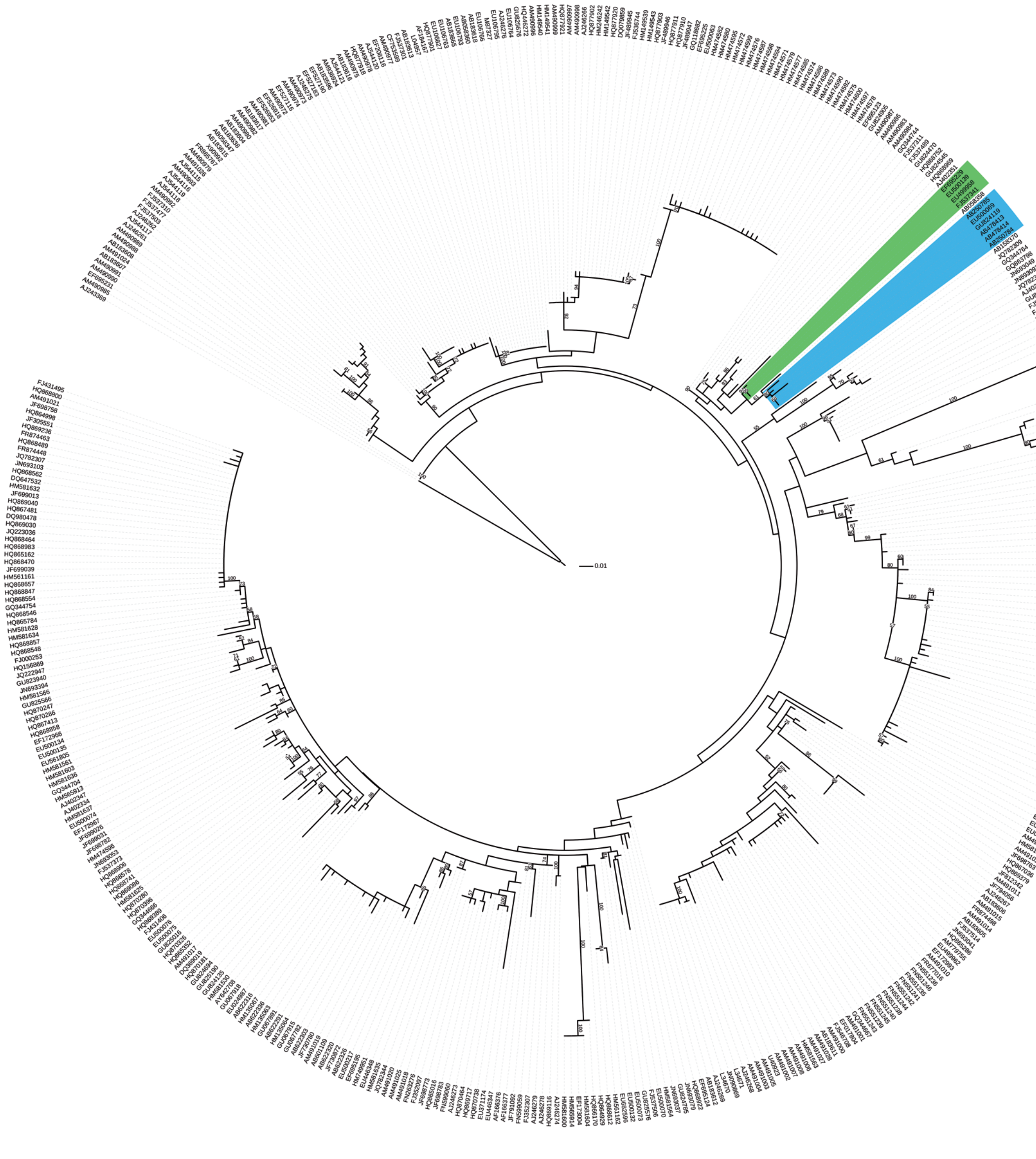
Figure 3. Time calibrated cyanobacteria tree. The phylogeny shown was estimated based on 135 proteins from 57 taxa. Three calibration points (black circles) were used for the tree presented and were treated as soft bounds. The root of the tree was set with a maximum age of 2,700 Mya and a minimum age of 2,320 Mya. Divergence time for the ancestor of cyanobacteria UCYN-A1 and UCYN-A2 (highlighted with a grey box) are given in with the corresponding values for the posterior 95% confidence intervals.

The present study offers new insights into the marine nitrogen-fixing UCYN-A symbiosis by disentangling the partner fidelity, host-symbiont organization and size distribution, gene expression and evolution of UCYN-A1 and UCYN-A2 lineages. These results demonstrate that specific UCYN-A symbiotic pairs co-exist without cross-symbiotic partnerships. The fact that its distribution occupies new plankton size fractions accordantly to the host size should be considered in global nitrogen fixation models. The number of UCYN-A1 and UCYN-A2 cells involved in this symbiosis differs and appears to be a conserved phylogenetic-trait. Remarkably, about a quarter of the UCYN-A transcripts were from nitrogen fixation genes, highlighting the importance of nitrogen fixation in this symbiosis. Our results present further evidences of a host and symbiont co-evolution scenario in the marine environment, probably derived from a single ancestral symbiotic event wherein at least two different lineages diversified in the late Cretaceous. Investigation of N_2-fixing cyanobacterial symbionts and their partners should provide clues for discovering new ecological compartments for nitrogen fixation that would increase our understanding of the nitrogen cycle in the ocean.

ACKNOWLEDGEMENTS

We are indebted to the scientists and crew on board of *Tara* Oceans expedition. We thank V. Balagué and I. Forn for technical support, P. Sánchez and R. Logares for providing access to bioinformatic resources and for bioinformatic advise and I. Ferrera for critical reading. This work was supported by the Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS, and TANIT (CONES 2010-0036) from the Agència de Gestió d´Ajusts Universitaris i Recerca to SGA. This research was further supported by European Molecular Biology Laboratory (EMBL), Investissements d'avenir grants Oceanomics (ANR-11-BTBR-OOO8) and France Génomique (ANR-10-INBS-09-08). FMCC and AMC held FPI (MICINN) fellowships and GS held Ph.D. JAE-Predoc (CSIC) fellowship. SGA was supported by FP7-OCEAN-2011 project "Micro B3". This article is contribution number 37 of *Tara* Oceans.

SUPPLEMENTARY INFORMATION



Supplementary Figure 1. Phylogenetic reconstruction of Class Prymnesiophyceae. Maximum likelihood phylogenetic tree of the Class Prymnesiophyceae based on the 18S rRNA gene. The tree includes 466 sequences (shown by their NCBI accession numbers) retrieved from the Protist Ribosomal Reference database (PR2). Bootstrap values above 50% are indicated. The UCYN-A1 host phylogroup targeted by probe UPRYM69 probe is highlighted in green while the UCYN-A2 host phylogroup targeted by probe UBRADO69 is highlighted in blue.

Supplementary Table 1 | Oligonucleotide probes used in CARD-FISH assays

Probe	Target organism	Sequence (5' to 3')	Reference
UCYN–A732	Unicellular cyanobacteria UCYN-A1	GTTACGGTCCAGTAGCAC	Krupke et al. (2013)
UCYN–A732 competitor	Unicellular cyanobacteria UCYN-A2	GTTGCGGTCCAGTAGCAC	This study
Helper A–732	Unicellular cyanobacteria UCYN-A	GCCTTCGCCACCGATGTTCTT	Krupke et al. (2013)
Helper B–732	Unicellular cyanobacteria UCYN-A	AGCTTTCGTCCCTGAGTGTCA	Krupke et al. (2013)
PRYM02	Prymnesiophyceae	GGAATACGAGTGCCCCTGAC	Simon et al. (2000)
UPRYM69*	UCYN-A1 host	CACATAGGAACATCCTCC	This study
UBRADO69*	*B. bigelowii*	CACATTGGAACATCCTCC	This study
Helper A-PRYM	Prymnesiophyceae	GAAAGGTGCTGAAGGAGT	This study
Helper B-PRYM	Prymnesiophyceae	AATCCCTAGTCGGCATGG	This study

*also used as competitor.

Supplementary Table 2 | Divergence time for the ancestor of cyanobacterium UCYN-A1 and UCYN-A2

	Phylobayes	MCMCtree
Divergence	Independent	Independent
UCYNA-1 and -2*	91 (46, 345)	141 (75, 234)

*Values in parenthesis correspond to the posterior 95% confidence intervals associated with median age estimates. Posterior age estimate in Million years.

REFERENCES

Alberti A, Belser C, Engelen S, Bertrand L, Orvain C, Brinas L, *et al.* (2014). Comparison of library preparation methods reveals their impact on interpretation of metatranscriptomic data. *BMC Genomics* **15**. doi:10.1186/1471-2164-15-912.

Bekker A, Holland HD, Wang P-L, Rumble D, Stein HJ, Hannah JL, *et al.* (2004). Dating the rise of atmospheric oxygen. *Nature* **427**:117–20.

Blank CE, Sánchez-Baracaldo P. (2010). Timing of morphological and ecological innovations in the cyanobacteria--a key to understanding the rise in atmospheric oxygen. *Geobiology* **8**:1–23.

Bombar D, Heller P, Sanchez-Baracaldo P, Carter BJ, Zehr JP. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *ISME J*. http://dx.doi.org/10.1038/ismej.2014.167.

Bown P, Lees J, Young J. (2004). Calcareous nannoplankton evolution and diversity through time. In:*Coccolithophores SE - 18*, Thierstein, H & Young, J (eds), Springer Berlin Heidelberg, pp. 481–508.

Brocks JJ, Buick R, Summons RE, Logan GA. (2003). A reconstruction of Archean biological diversity based on molecular fossils from the 2.78 to 2.45 billion-year-old Mount Bruce Supergroup, Hamersley Basin, Western Australia. *Geochim Cosmochim Acta* **67**:4321–4335.

Cabello AM, Cornejo-Castillo FM, Raho N, Blasco D, Vidal M, Audic S, *et al.* (2015). Global distribution and vertical patterns of a prymnesiophyte-cyanobacteria obligate symbiosis. *ISME J*. **10**:693-706.

Cárdenas AL, Harries PJ. (2010). Effect of nutrient availability on marine origination rates throughout the Phanerozoic eon. *Nat Geosci* **3**:430–434.

Caro-quintero A, Konstantinidis KT. (2011). Bacterial species may exist , metagenomics reveal. *Environ Microbiol*. doi:10.1111/j.1462-2920.2011.02668.x.

Decelle J, Probert I, Bittner L, Desdevises Y, Colin S, De Vargas C, *et al.* (2012). An original mode of symbiosis in open ocean plankton. *Proc Natl Acad Sci U S A* **109**:18000–18005.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biol* **4**:e88.

Goebel NL, Turk KA, Achilles KM, Paerl R, Hewson I, Morrison AE, *et al.* (2010). Abundance and distribution of major groups of diazotrophic cyanobacteria and their potential contribution to N_2 fixation in the tropical Atlantic Ocean. *Environ Microbiol* **12**:3272–89.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, *et al.* (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**:D597–D604.

Hagino K, Onuma R, Kawachi M, Horiguchi T. (2013). Discovery of an endosymbiotic nitrogen-fixing cyanobacterium UCYN-A in Braarudosphaera bigelowii (Prymnesiophyceae). *PLoS One* **8**:e81749.

Hasegawa M, Kishino H, Yano T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol* **22**:160–174.

Hurst LD. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18**:486–487.

Jardillier L, Zubkov M V, Pearman J, Scanlan DJ. (2010). Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME J* **4**:1180–92.

Karl D, Michaels A, Bergman B, Capone D, Carpenter E, Letelier R, *et al.* (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58(1)**:47–98.

Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, *et al.* (2011). A holistic approach to marine eco-systems biology. *PLoS Biol* **9**:e1001177.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**:3059–3066.

Kneip C, Lockhart P, Voss C, Maier U-G. (2007). Nitrogen fixation in eukaryotes--new models for symbiosis. *BMC Evol Biol* **7**:55.

Krupke A, Lavik G, Halm H, Fuchs BM, Amann RI, Kuypers MMM. (2014). Distribution of a consortium between unicellular algae and the N2 fixing cyanobacterium UCYN-A in the North Atlantic Ocean. *Environ Microbiol*. doi:10.1111/1462-2920.12431.

Krupke A, Mohr W, LaRoche J, Fuchs BM, Amann RI, Kuypers MMM. (2015). The effect of nutrients on carbon and nitrogen fixation by the UCYN-A-haptophyte symbiosis. *ISME J* **9**:1635–1647.

Krupke A, Musat N, Laroche J, Mohr W, Fuchs BM, Amann RI, *et al.* (2013). In situ identification and N_2 and C fixation rates of uncultivated cyanobacteria populations. *Syst Appl Microbiol* **36**:259–71.

Lartillot N, Lepage T, Blanquart S. (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**:2286–8.

Lepage T, Bryant D, Philippe H, Lartillot N. (2007). A general comparison of relaxed molecular clock models. *Mol Biol Evol* **24**:2669–80.

Letunic I, Bork P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**:127–8.

Letunic I, Bork P. (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**:W475-8.

Li WH. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* **36**:96–99.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, *et al.* (2013). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol*. doi:10.1111/1462-2920.12250.

Martinez-Garcia M, Brazel D, Poulton NJ, Swan BK, Gomez ML, Masland D, *et al.* (2012). Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J* **6**:703–707.

McDonald JH, Kreitman M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**:652–654.

McFall-Ngai M. (2008). Are biologists in 'future shock'? Symbiosis integrates biology across domains. *Nat Rev Microbiol* **6**:789–792.

Le Moal M, Collin H, Biegala IC. (2011). Intriguing diversity among diazotrophic picoplankton along a Mediterranean transect: a dominance of rhizobia. *Biogeosciences* **8**:827–840.

Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. (2004). High rates of N2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* **430**:1027–32.

Nakayama T, Ikegami Y, Nakayama T, Ishida K-I, Inagaki Y, Inouye I. (2011). Spheroid bodies in rhopalodiacean diatoms were derived from a single endosymbiotic cyanobacterium. *J Plant Res* **124**:93–7.

Peleo-Alampay AM, Mead GA, Wei W. (1999). Unusual Oligocene Braarudosphaera-rich layers of the South Atlantic and their palaeoceanograpghic implications. *J Nannoplankt Res* **21(1)**:17–26.

Pernthaler A, Pernthaler J, Amann R. (2004). Section 3 update: Sensitive multi-color fluorescence in situ hybridization for the identification of environmental microorganisms. In:*Molecular Microbial Ecology Manual SE - 311*, Kowalchuk, GA, De Bruijn, FJ, Head, IM, Akkermans, AD, & Van Elsas, JD (eds), Springer Netherlands, pp. 2613–2627.

Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, *et al.* (2015). Open science resources for the discovery and analysis of Tara Oceans Data. *bioRxiv*. http://biorxiv.org/content/early/2015/05/08/019117.abstract.

Sánchez-Baracaldo P, Ridgwell A, Raven JA. (2014). A neoproterozoic transition in the marine nitrogen cycle. *Curr Biol* **24**:652–7.

Simon N, Campbell L, Örnolfsdottir E, Groben R, Guillou L, Lange M, *et al.* (2000). Oligonucleotide probes for the identification of three algal groups by dot blot and fluorescent whole-cell hybridization. *J Eukaryot Microbiol* **47**:76–84.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinforma* **22**:2688–2690.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, *et al.* (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**:1261359.

Thompson A, Carter BJ, Turk-Kubo K, Malfatti F, Azam F, Zehr JP. (2014). Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol* n/a-n/a.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, *et al.* (2012). Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science (80-)* **337**:1546–1550.

Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. (2006). The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. *Proc Natl Acad Sci U S A* **103**:5442–7.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**:90–94.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, *et al.* (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**:1261605.

Yang Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**:1586–91.

Yang Z, Rannala B. (2006). Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol Biol Evol* **23**:212–26.

Zehr JP. (2011). Nitrogen fixation by marine cyanobacteria. *Trends Microbiol* **19**:162–73.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, *et al.* (2008). Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**:1110–2.

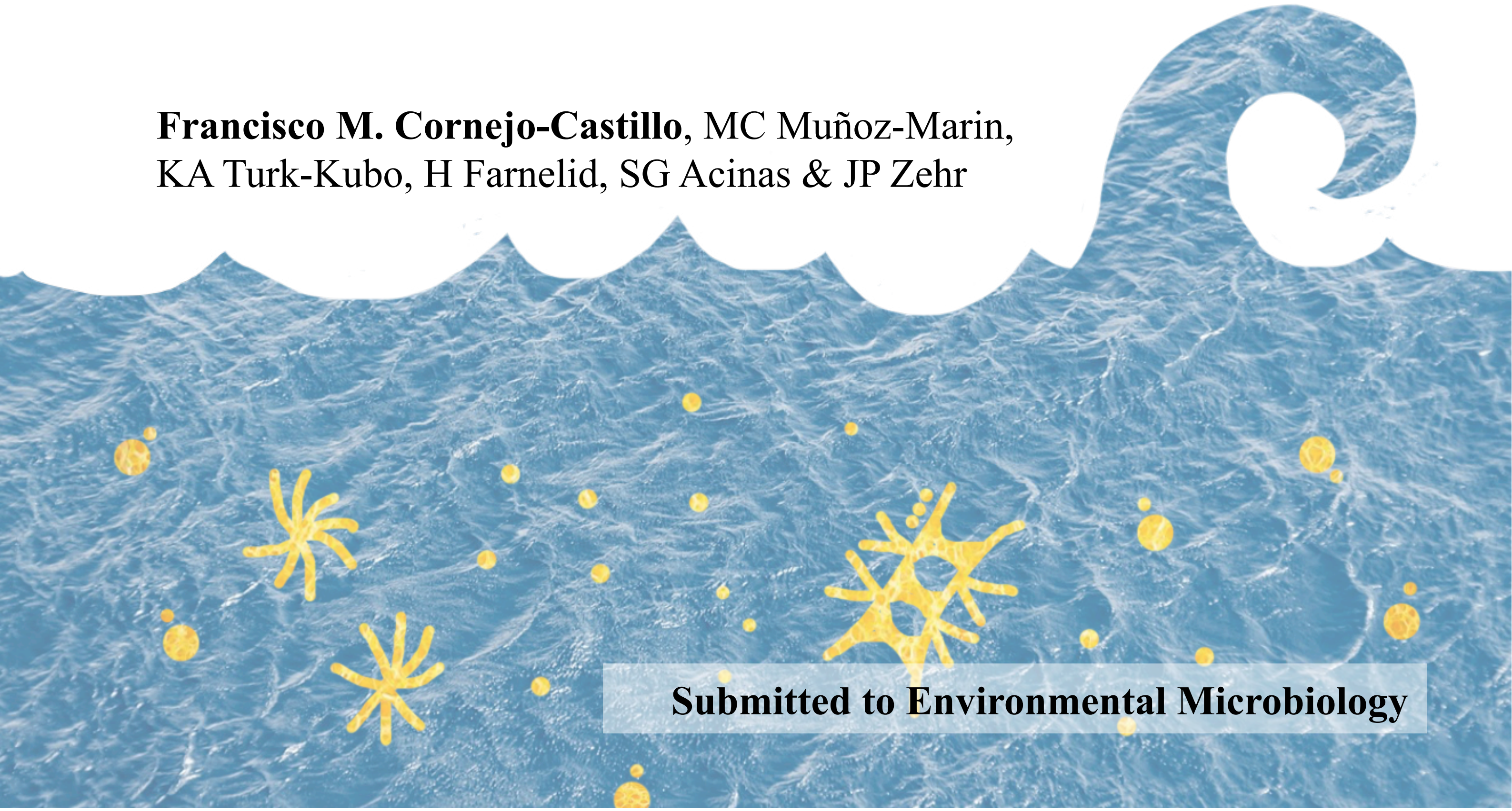
Zehr JP, Kudela RM. (2011). Nitrogen Cycle of the Open Ocean: From Genes to Ecosystems. *Ann Rev Mar Sci* **3**:197–225.

Zhang, Y. and Golubic S. (1987). Endolithic microfossils (cyanophyta) from early Proterozoic stromatolites, Hebei, China. *Acta Micropaleontol Sin* **4**:1–3.

Chapter 3

The emerging diversity of the nitrogen-fixing UCYN-A clade occupies different size fractions of the marine plankton

Francisco M. Cornejo-Castillo, MC Muñoz-Marin, KA Turk-Kubo, H Farnelid, SG Acinas & JP Zehr



Submitted to Environmental Microbiology

Abstract

The symbiotic unicellular cyanobacterium *Candidatus* Atelocyanobacterium thalassa (UCYN-A) is one of the most abundant and widespread nitrogen (N_2)-fixing cyanobacteria in the ocean. Although it remains uncultivated, multiple sublineages have been detected based on partial nitrogenase (*nifH*) gene sequences, including the four most commonly detected sublineages UCYN-A1, UCYN-A2, UCYN-A3 and UCYN-A4. Recently, advances in single cell visualization techniques using specific probes that target UCYN-A1 and UCYN-A2 and their respective hosts provided new insight into the morphology of these symbioses. Moreover, it has been suggested that UCYN-A2 is widely distributed throughout the oligotrophic oceans in addition to coastal waters. Very little is known about UCYN-A3 and the other sublineages beyond the *nifH* sequences from *nifH* gene diversity surveys. In this study, several different assays and methods were used that revealed discrepancies in identification of sublineages and led to new information on the diversity of the UCYN-A symbiosis. We report here that the UCYN-A association originally assumed to be UCYN-A2 at two open ocean sites is actually UCYN-A3. Our studies show that the size of the UCYN-A3 cells (both the cyanobacteria and host) and the number of cyanobacterial cells per host differs from that in the better characterized sublineages (UCYN-A1 and UCYN-A2). Moreover, the present study expands the known extent of UCYN-A genetic diversity by the reconstruction of about 13% of the UCYN-A3 genome from metagenomic data. Finally, our results unveil that the UCYN-A lineages are distributed along different size fractions of the plankton defined by the cell-size range of their prymnesiophyte hosts, uncovering new nitrogen fixation planktonic compartments.

Introduction

Biological nitrogen (N_2) fixation is a fundamental biogeochemical process in the ocean, whereby N_2 gas is reduced to ammonia, which supports primary production (Sohm *et al.*, 2011; Karl *et al.*, 2002). It has been long thought that the most important N_2-fixing microorganism ("diazotroph") in the open ocean was the free-living cyanobacterium *Trichodesmium* (Luo *et al.*, 2012). However, it is now clear that marine N_2-fixing cyanobacteria are more diverse and include a wide range of lifestyles, including symbionts such as the unicellular cyanobacterium *Candidatus* Atelocyanobacterium thalassa, commonly known as UCYN-A (Thompson and Zehr, 2013). UCYN-A lives in a mutualistic partnership with an uncultivated unicellular alga, a prymnesiophyte, closely related to *Braarudosphaera bigelowii* (Thompson *et al.*, 2012; Hagino *et al.*, 2013). This symbiosis is based on the exchange of carbon and nitrogen between partners (Thompson *et al.*, 2012; Krupke *et al.*, 2013), which explains how UCYN-A can thrive in oligotrophic environments despite lacking important biosynthetic pathways (Tripp *et al.*, 2010).

The nitrogenase (*nifH*) and 16S rRNA gene sequences of UCYN-A have been reported in a wide variety of oceanic environments which has suggested it has a major role in global N_2 fixation (Moisander *et al.*, 2010; Martínez-Pérez *et al.*, 2016; Farnelid *et al.*, 2016; Turk-Kubo *et al.*, 2017). Recent phylogenetic analyses based on partial UCYN-A *nifH* gene sequences have shown that there are at least four distinct sublineages, UCYN-A1, UCYN-A2, UCYN-A3 and UCYN-A4 (Thompson *et al.*, 2014; Farnelid *et al.*, 2016; Turk-Kubo *et al.*, 2017). However, only the UCYN-A1 and UCYN-A2 genomes have been sequenced (Tripp *et al.*, 2010; Bombar *et al.*, 2014). The UCYN-A2 sublineage is found specifically associated with *B. bigelowii* while UCYN-A1 is associated with a smaller, but closely-related prymnesiophyte (1-3 μm) (Thompson *et al.*, 2012, 2014; Cornejo-Castillo *et al.*, 2016). The hosts of the other UCYN-A sublineages are not yet known. Curiously, the UCYN-A2 symbiosis observed by Cabello *et al.* (2015) was only half the diameter of the originally described UCYN-A2 symbiosis (Thompson *et al.*, 2014), suggesting that UCYN-A2 in the open ocean was smaller than in coastal sites (4-5 µm compared to 7-10 µm). However, the FISH probe used at that time targeted all UCYN-A clades (Krupke *et al.*, 2013) making it impossible to distinguish between UCYN-A sublineages. Eventually, 16S rRNA gene sequences of UCYN-A1 and UCYN-A2 and 18S rRNA gene sequences of their respective hosts made it possible to design CARD-FISH probes that differentiated UCYN-A1 and UCYN-A2 sublineages (Cornejo-Castillo *et al.*, 2016). These were used to show that the UCYN-A2 symbiosis was actually composed of several (3-10) UCYN-A2 cells per host cell, in contrast to UCYN-A1, which had only one cell per host (Krupke *et al.*, 2013; Cabello *et al.*, 2015; Cornejo-Castillo *et al.*, 2016; Martínez-Pérez *et al.*, 2016). Metagenomic analysis of size-fractionated samples from the *Tara* Oceans expedition showed that the UCYN-A1 genome was recovered in the small size-fraction (0.2-3 µm), whereas the UCYN-A2 genome was found in the 0.8-5 and 5-20 µm size-fraction, in agreement with all previous observations of size of the symbiosis by CARD-FISH and qPCR of flow cytometry sorted cells (Thompson *et al.*, 2014; Cabello *et al.*, 2015; Cornejo-Castillo *et al.*, 2016).

Recent studies have shown that UCYN-A1 and UCYN-A2 symbiosis both have wide global distributions and that they often coexist in space and time (Cabello *et al.*, 2015; Cornejo-Castillo *et al.*, 2016). A third phylogenetically distinct group, UCYN-A3, appears also to have a global distribution, and is commonly detected in oligotrophic waters, including at Station ALOHA in the North Pacific Subtropical Gyre (NPSG), and co-occurs with UCYN-A1 (Turk-Kubo *et al.*, 2017). Likewise, the UCYN-A4 sublineage has been observed to co-occur with UCYN-A2 in coastal waters (Turk-Kubo *et al.*, 2017).

In this work, we studied UCYN-A populations in two different oceanic regimes: Station ALOHA in the NPSG and Southern California Coastal Current waters near the Scripps Institution of Oceanography (SIO) Pier in La Jolla, CA, USA. We identified the UCYN-A sublineages that were present in these samples using a PCR assay that specifically targets the *nifH* gene from UCYN-A and quantified

UCYN-A cell abundances using quantitative PCR assays for each previously described cyanobacteria/prymnesiophyte pair. Cell size ranges and morphological features of UCYN-A sublineages were determined using double CARD-FISH. We also analyzed size-fractionated metagenome sequence libraries collected during the *Tara* Oceans expedition in the South Atlantic to detect and reconstruct the genome(s) of divergent UCYN-A populations other than those of UCYN-A1 and UCYN-A2, and to determine the evolutionary relationships of the UCYN-A sublineages to other unicellular N_2-fixing cyanobacteria.

Materials and Methods

Sampling procedures

Samples were collected from the Scripps Institution of Oceanography (SIO) Ellen Browning Scripps Memorial Pier (32º 52’ 1.7” N 117º 15’ 27.3” W) in La Jolla, CA between 28th July and 1st August 2014 and on the 14th July 2015. Surface samples (0 m) were obtained using a bucket at the end of the Pier. Samples were also collected from CTD casts at Station ALOHA (22º 45’ N 158º W, at 45 meters depth) during the C-MORE Cruise C-20 (http://hahana.soest.hawaii.edu/hot/cruises.html) between 6-10th April, 2015. At each sampling site, water was collected in 2 L polypropylene bottles for CARD-FISH and 10 L polypropylene bottles for DNA, and were covered with black plastic until fixed or filtered.

For the CARD-FISH assay, 190 mL of seawater was fixed in the dark for 1 hour at 4ºC with 10 mL 37% formaldehyde (1.87% v/v final concentration) for two replicates. For each sample, 100 mL was filtered at a maximum vacuum pressure of 100 mm Hg onto 0.6 µm pore-size polycarbonate membrane filter, 25 mm diameter (EMD Millipore) with a support filter of 0.8 µm pore-size, 25 mm polycarbonate cellulose acetate membrane filter (Sterlitech Corporation) and kept frozen -80º C until processed.

Duplicates of DNA samples from SIO were collected by filtering 500 mL of seawater through 47 mm, 0.22 µm pore-size, Supor filters (Pall Corporation) using a peristaltic pump. At Station ALOHA, 4 L samples were filtered onto 0.22 µm pore-size Sterivex cartridges (Millipore) using low pressure with a peristaltic pump. Filters were placed in sterile 2 ml bead-beating tubes with sterile glass beads and stored at -80 ºC until extraction.

DNA extraction

DNA extractions were carried out with a modification of the Qiagen DNeasy Plant Kit (Moisander *et al.*, 2008). Briefly, 400 µL AP1 buffer was added to the bead-beating tubes, followed by three sequential freeze-thaw cycles using liquid nitrogen and a 65ºC water bath. The tubes were agitated for 2 min with a FastPrep-24 bead beater (MP Biomedicals), and incubated for 1h at 55ºC with 20 mg ml^{-1} proteinase K (Qiagen). Samples were treated for 10 min at 65ºC with 4 µL RNase A (100 mg/mL) and then the filters were removed using sterile needles. The tubes were centrifuged for 5 min at 14,000 rpm at 4ºC, and the supernatant was further purified using the QIAcube automated extraction platform according to the manufacturer's protocol (Qiagen). Samples were eluted using 100 µL AE buffer and stored at 20ºC.

Quantitative PCR (qPCR) assay

Taqman® qPCR assays were used to measure the abundances of UCYN-A1 and UCYN-A2 and their respective hosts (Church *et al.*, 2005; Thompson *et al.*, 2014) (Supplementary Table 1). Each assay used TaqMan® Gene Expression MasterMix (Invitrogen) at 1X concentration, 0.4 µM forward and reverse primers, 0.2 µM Taqman® probe and 2 µL of the DNA extract, for a final volume of 25 µL.

The four assays were initially incubated for 10 min at 95°C to relax target DNA, and data was collected at the end of each of 45 cycles of 15 s at 95°C and 60 s at 60°C for all assays except for the UCYN-A2 *nifH* gene assay that used an annealing temperature of 64°C.

For each assay standards were included as positive controls and used to quantify the gene copy numbers. Standards were generated using linear plasmids containing cloned inserts of PCR amplified genes from environmental samples containing either the UCYN-A1 or the UCYN-A2 *nifH* gene or the respective prymnesiophyte 18S rRNA genes (Host-A1/Host-A2) (Thompson *et al.*, 2014). Standards were added to the environmental DNA samples to test for PCR inhibition. To investigate the relation between the cyanobacteria and host abundances, the ratio of UCYN-A per host was calculated assuming one copy of *nifH* in the UCYN-A genome and one copy of the 18S rRNA gene in the *B. bigelowii* genome. However, it must be noted that qPCR is not reliable for absolute quantification of single gene copies and that the cell-to-cell ratio estimates need to be interpreted with caution since we do not know the number of copies of the 18S rRNA gene in the host genome.

Double CARD-FISH assay

The double CARD-FISH assay was carried out following the protocol designed by Cabello *et al.* (2015) and Cornejo-Castillo *et al.* (2016). The sequences of the probes used are compiled in Supplementary Table 2. Following hybridization, the filters were rinsed in a washing buffer (9 mM NaCl, 5 mM EDTA, 0.01% SDS, 20 mM Tris-HCl pH 8) at 37°C, and the TSA reaction performed using the TSA Plus Cyanine 3 System (Perkin Elmer, Inc) for 10 min at room temperature in the dark following the manufacturer's instructions. Filters were stained with 5 μg ml^{-1} DAPI (4′, 6- diamidino-2-phenylindole), mounted in antifading reagent (77% glycerol, 15% VECTASHIELD and 8% 20 Å~ PBS) and the micrographs were obtained using Leica SP5 Confocal Microscope at the University of California, Santa Cruz Life Sciences Microscopy Center. Filters were observed under ultraviolet (DAPI), blue (host stained with Alexa 488 in green) and green (UCYN-A stained with Cy3 in red) excitation wavelengths.

Microscopic observations and cell counting (100 associations per sample) were performed with a Carl Zeiss Axioplan-2 Imaging Fluorescent Microscope

(Zeiss). Cell dimensions were estimated using AxioVision 4.8.1 and Image J software (Schindelin *et al.*, 2012).

Sequence processing

UCYN-A *nifH* gene fragments were amplified using a nested PCR assay designed to amplify known UCYN-A diversity. The first round of amplification was carried out using a widely used universal *nifH* primer set, nifH3/nifH4 (Zani *et al.*, 2000) and the second round of amplification used newly developed universal UCYN-A *nifH* primers described in Turk-Kubo *et al*. (2017). The UCYN-A specific primers were both modified with 5’ common sequence linkers, to facilitate library preparation using a dual PCR strategy (Green *et al.*, 2015). UCYN-A *nifH* PCR amplicons were multiplexed with other samples for a targeted depth of coverage of ca. 40,000 sequences, and sequenced using the Illumina MiSeq platform (2 x 250 bp paired ends).

Raw reads were merged and quality-filtered (phred score of 20) using the PEAR aligner (Zhang *et al.*, 2014), chimeras were removed using UCHIME (Edgar *et al.*, 2011), and sequences were clustered at 99% sequence similarity using USEARCH v6.1 (Edgar, 2010) through QIIME (Caporaso *et al.*, 2010) . Cluster representatives with greater than 500 sequences (which accounted for 92% of all recovered sequences) were imported into ARB (Ludwig *et al.*, 2004), where they were translated into amino acids and sequences with stop codons were removed. Representative sequences that passed all quality filter steps were aligned to existing UCYN-A alignments in a curated *nifH* database (Heller *et al.*, 2014), and exported, along with representative sequences from each sublineage, for tree construction in MEGA6 (Tamura *et al.*, 2013). Maximum likelihood trees were calculated using the Tamura-Nei branch length correction and node support was determined with 1000 bootstrap replicates. Distribution data for the representative sequences was simplified from USEARCH v6.1 output files using a custom python script, and visualized using the interactive tree of life (iTOL) web tool (Letunic and Bork, 2007).

Fragment recruitment and UCYN-A3 genome reconstruction

The analysis of the abundance of UCYN-A based on 16S $_{\mathrm{mi}}$TAGs along 243 metagenomes from 68 globally distributed stations from *Tara* Oceans expedition revealed only two stations, TARA_078 (30º 8’ 12.12” S 43º 17’ 23.64” W) and TARA_076 (20º 56’ 7.44” S 35º 10’ 49.08” W) in the South Atlantic Ocean, where UCYN-A was significantly abundant (Cornejo-Castillo *et al.*, 2016). Therefore, a total of 8 metagenomes spanning four size-fractions (0.2–3, 0.8–5, 5–20 and >0.8 µm) from these two sampling stations, TARA_078 and TARA_076) were analyzed for UCYN-A gene fragments. Both seawater collection and DNA extraction protocols for the different size-fractions and metagenome sequencing are described in Cornejo-Castillo *et al.* (2016).

BLAST+ v2.2.25 was used to recruit metagenomic reads closely related to UCYN-A1 and UCYN-A2 genomes using default parameters with some modifications: -perc_identity 70, -evalue 0.00001. A reference database was constructed that contained the two UCYN-A genomes sequenced to date, UCYN-A1 and UCYN-A2. Metagenomic reads aligned to the ribosomal operon were excluded from the analysis. Likewise, reads aligned over less than 90% of its length were excluded to avoid random alignments.

A *de novo* metagenome co-assembly process based on fragment recruitment results was carried out to reconstruct a fraction of the UCYN-A3 genome (Supplementary Figure 1). Instead of using contigs from individual metagenomes, we used all reads closely related to UCYN-A1 and UCYN-A2 extracted from all 8 metagenomes to build the UCYN-A3 contigs. The criteria to select this subset of metagenomic reads was based on the identity shared between metagenomic reads and the reference genomes to define genomic species (Caro-quintero and Konstantinidis, 2011). Thus, reads with an identity between 80-95% to the reference genomes were assumed to belong to a divergent UCYN-A sublineage. Subsequently, these selected reads were assembled to build contigs using MEGAHIT v1.0.4-beta (Li *et al.*,

2015) with the following parameters: --presets meta-sensitive -m 0.97 -t 24. The metagenomic samples used for the novel genome reconstruction correspond to the 0.8-5 and >0.8 µm size-fractions. No reads were found in the other size fractions. Every single reconstructed gene was compared to GenBank using BLASTN against the nt database to verify its taxonomic assignment to the UCYN-A clade. High-Performance computing analyses were run at the Marine Bioinformatics Service (MARBITS) of the Institut de Ciències del Mar (ICM-CSIC) in Barcelona (Spain).

Phylogenomic analysis of UCYN-A sublineages

Sequence data for 165 protein-coding genes were used to estimate the evolutionary relationships of the new UCYN-A sublineage. These genes were extracted from the following cyanobacterial genome sequences: *Cyanothece* sp. PCC 7822 (NC_014501.1), *Cyanothece* sp. PCC 7424 (NC_011729.1), *Cyanothece* sp. PCC 8801 (NC_011726.1), *Cyanothece* sp. PCC 8802 (NC_013161.1), endosymbiont of *Epithemia turgida* EtSB (NZ_AP012549.1), *Cyanothece* sp. ATCC 51142 (NC_010546.1), *Pleurocapsa* sp. PCC 7327 (NC_019689.1), *Candidatus* Atelocyanobacterium thalassa SIO64986 (UCYN-A2; JPSP00000000.1) and *Candidatus* Atelocyanobacterium thalassa ALOHA (UCYN-A1; NC_013771.1). For the new UCYN-A sublineage, these 165 protein-coding genes were extracted from the newly assembled contigs. All genes were independently aligned using the translation align MUSCLE algorithm implemented in the Geneious software (Geneious Pro 4.8.5). Once aligned, the 165 genes were concatenated and, as result, the combined sequence length was 88,107 bp. Finally, a maximum likelihood phylogenetic tree was built using RAxML (Stamatakis, 2006) with 100 trees for both topology and bootstrap analyses.

On-line supplementary information

Supplementary Table 4 and Supplementary Table 5 can be downloaded from the following link: https://www.dropbox.com/sh/6b401yn1nhaom2i/AACaXmeY1nJlXSCyR-FBD3HYa?dl=0

Results

1. Diversity of UCYN-A

1.1. *nifH* gene sequences

UCYN-A diversity was assessed by amplifying *nifH* gene sequences from SIO Pier and Station ALOHA (Figure 1A). In samples taken in two consecutive summers at the SIO Pier, the UCYN-A community was defined primarily by three different *nifH* phylotypes: OTU00, OTU01, and OTU03 (Figure 1B). OTU00 clusters within the UCYN-A1 sublineage, with relative abundances ranging between 28.8-51.3%. UCYN-A2 sequences were also recovered and the sequence type with the highest relative abundances was OTU01 (ranging between 43.7-66.1%), which is 100% identical to the UCYN-A2 *nifH* sequence reported by Thompson *et al.* (2014). A third sequence type was recovered, OTU03, which clusters with UCYN-A4, a new sublineage described by Farnelid *et al.* (2016). The UCYN-A4 sequence was present during both years, but relative abundances in 2014 (0.5%) were lower than in 2015 (2.9-6.1%).

At Station ALOHA, the UCYN-A community was comprised primarily of UCYN-A1 (OTU00) and UCYN-A3 (OTU02) (Figure 1B). UCYN-A1 was the dominant phylotype, and accounted for 87% of the UCYN-A sequences recovered (Figure 1B). Additionally, the UCYN-A3 sublineage was also recovered at lower relative abundances and accounted for 9.4% of the *nifH* sequences. UCYN-A2 sequences were also recovered but at much lower relative abundances (ca. 0.7%).

1.2 Visualization of UCYN-A associations

To visualize cells of the different UCYN-A sublineages, and to better define the size ranges of each of the sublineages in coastal and open-ocean environments, we applied a double CARD-FISH assay to samples collected at the SIO Pier and at Station ALOHA.

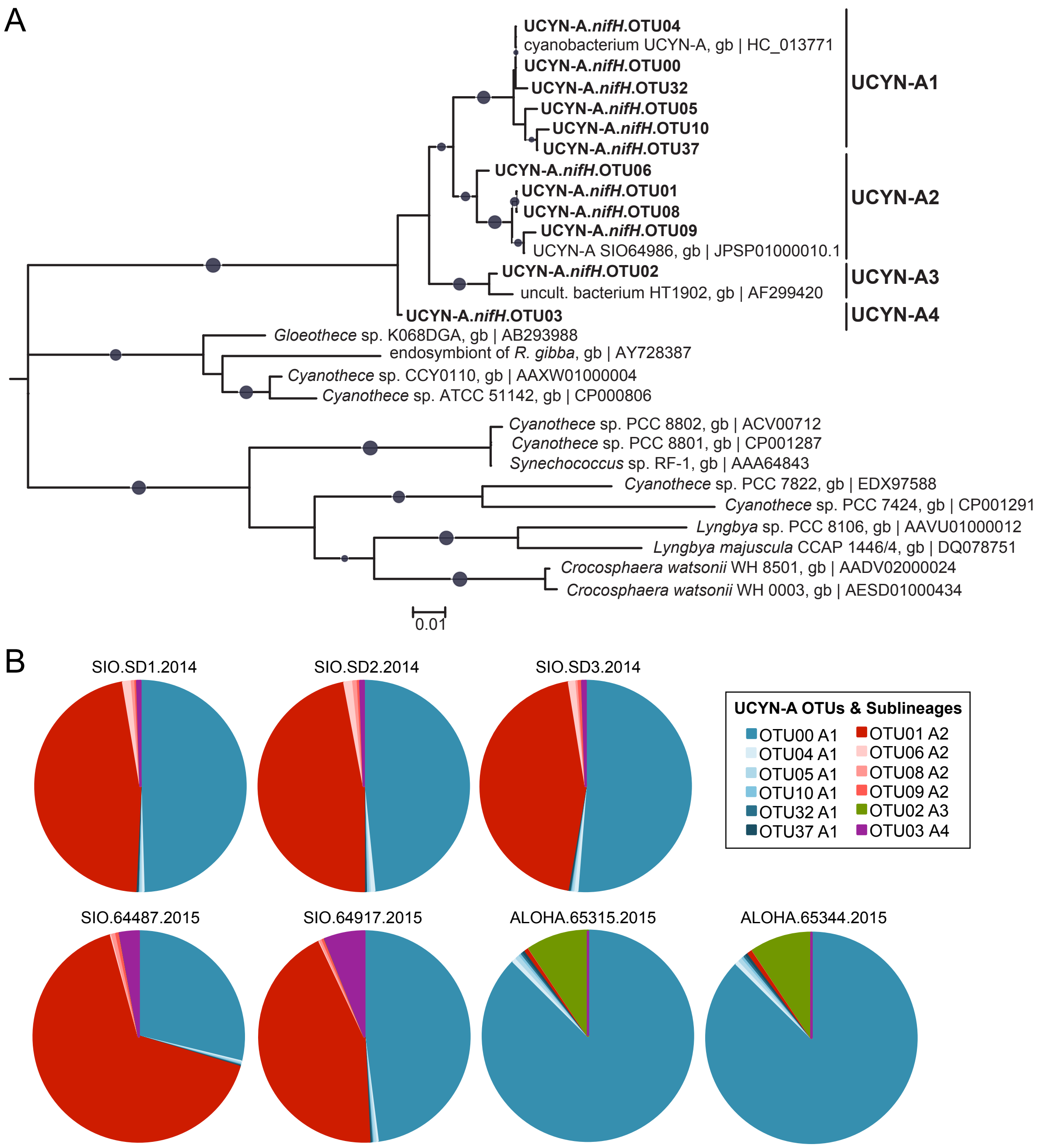


Figure 1. **Diversity and relative abundances of UCYN-A sublineages at the SIO Pier and Station ALOHA**. A) Maximum likelihood phylogenetic tree constructed using partial *nifH* genes from the 12 UCYN-A OTUs with the highest relative abundances across the dataset and closely related UCYN-A and unicellular cyanobacterial sequences as reference sequences. UCYN-A OTUs recovered in this study are in bold, and were defined by clustering at 99% nucleotide identity. Nodes with bootstrap support greater than 50 and based on 1000 replicate trees are identified with a black circle; the size of the circle correlates to the bootstrap value, with larger circles on nodes with stronger bootstrap support. Sublineages are labeled to the right of the tree based on Thompson *et al.*, (2014) and Farnelid *et al*. (2016). B) Relative abundances of these 12 UCYN-A OTU in each sample. SIO – Scripps Pier; ALOHA – Station ALOHA.

At SIO Pier, the UCYN-A1 host averaged 2.3 ± 0.3 µm (n=100 cells) (Supplementary Figure 2), and consistently associated with a single UCYN-A1 cell that ranged between 1.0 ± 0.2 µm in diameter (n=100 cells) (Figure 2A, B). With an average cell diameter of 7.3 ± 1.0 µm (n=100 cells), the UCYN-A2 host at SIO had between 5-10 UCYN-A2 cells per host, with an average size of 3.3 ± 0.5 µm in diameter (n=100 cells) (Figure 2C, D) (Supplementary Figure 1).

The sizes of UCYN-A1 and its prymnesiophyte host did not differ significantly (p value > 0.05) between the two sampling locations (SIO Pier and Station ALOHA) (Figure 2A, 3A and Supplementary Figure 2). However, at Station ALOHA, using the UCYN-A2 and Host-A2 probes, the targeted association appeared to be much smaller, that is 3.6 µm **± 0.7** µm (n=100 cells) in diameter compared to 7.3 µm at SIO (p value < 0.05) (Figure 2C and 3C).

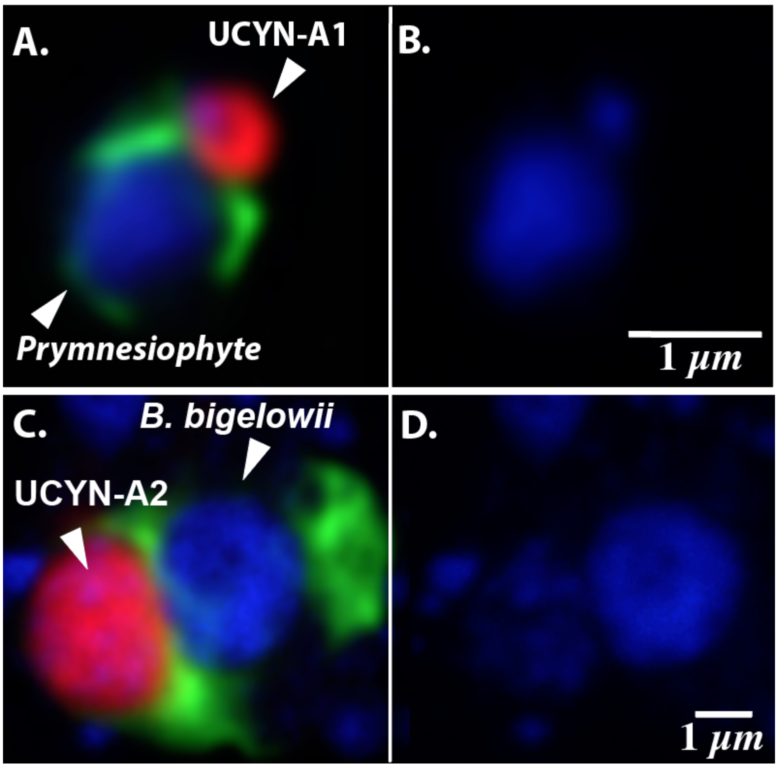


Figure 2. Micrographs of UCYN-A associations at SIO Pier. Epifluorescence microscopy images using the double-CARD-FISH assay showing the specificity of symbiont–host pairs **(A,C)**: On top **(A,B)**, UCYN-A1 with its prymnesiophyte partner labeled with the UCYN-A1-732 and UPRYM69 probes with competitors. On the bottom **(C,D)**, UCYN-A2 association labeled with the UCYN-A2-732 and UBRADO69 probes with competitors. Right panels **(B,D),** correspond to the 4´-6-diamidino-2-phenylindole signal (DAPI; blue).

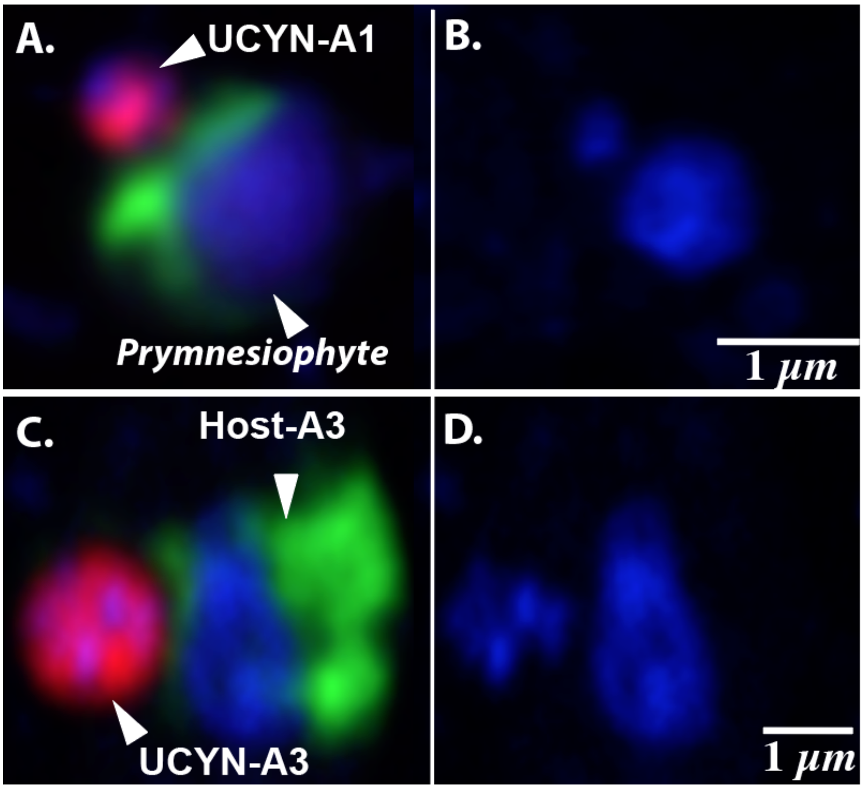


Figure 3. Micrographs of UCYN-A associations at Station ALOHA. Epifluorescence microscopy images with the double-CARD-FISH assay showing the specificity of symbiont–host pairs **(A,C)**: On top **(A,B)**, UCYN-A1 with its prymnesiophyte host labeled with the UCYN-A1-732 and UPRYM69 probes with competitors. On the bottom **(C,D)**, new UCYN-A association labeled with the UCYN-A2-732 and UBRADO69 probes with competitors. Right panels **(B-D),** correspond to the 4´-6-diamidino-2-phenylindole signal (DAPI; blue).

Based on the sequencing results described above from the corresponding DNA samples, where UCYN-A2 was virtually absent, but UCYN-A3 was the second most abundant UCYN-A sublineage, we assume the association that hybridized to the UCYN-A2 and Host-A2 probes, was UCYN-A3, not UCYN-A2.

To further target and characterize new UCYN-A associations (other than UCYN-A1 and UCYN-A2), we nonspecifically targeted all UCYN-A cells with the probe UCYN-A1 732 without competitors (universal UCYN-A probe) at both stations (Figure 4). The dual labeling with the UPRYM69 probe and competitor allowed us to distinguish the UCYN-A1 host from other prymnesiophyte-like cells that were associated with UCYN-A. At Station ALOHA, we did not detect new

UCYN-A associations, but at SIO Pier another UCYN-A association was observed (Figure 4). The size of this UCYN-A cell was similar to the size of UCYN-A1, 0.9 ± 0.2 µm (n=5) and was not significantly different in diameter (p value > 0.05) (Supplementary Figure 2). However, the host cell of this UCYN-A type was not detected with either the UCYN-A1 or UCYN-A2 host probes. The abundance of this new UCYN-A association was estimated to be approximately 210 cells L^{-1}, based on CARD-FISH epifluorescence cell counts.

1.3 UCYN-A and host gene copy ratios

We quantified the gene copy ratios of UCYN-A and hosts in coastal (SIO Pier) and open ocean (Station ALOHA) regions using previously established qPCR assays (Church *et al.*, 2005; Thompson *et al.*, 2014) targeting the *nifH* gene of UCYN-A and the 18S rRNA gene of the two known hosts (Figure 5). We will refer to the UCYN-A2 qPCR assay designed by Thompson *et al.* (2014) as UCYN-A2/UCYN-A3, since the UCYN-A2 qPCR primers and probe do not contain sufficient mismatches with the UCYN-A3 sublineage to prevent cross-hybridization and amplification (Farnelid *et al.*, 2016).

In samples from the SIO Pier, the UCYN-A2/UCYN-A3 and the UCYN-A2 host (*B. bigelowii*) gene copy ratios averaged $1.2*10^5 \pm 8*10^3$ *nifH* copies L^{-1} and $4.7*10^5 \pm 3.7*10^4$ 18S rRNA gene copies L^{-1}, respectively, with the host gene copy ratios four times greater than the symbiont (Figure 5). UCYN-A1 averaged $1.9*10^5 \pm 1.6 *10^4$ *nifH* copies L^{-1}, but the prymnesiophyte host originally described by Thompson *et al*. (2012) was not detected using the UCYN-A1 host assay (Supplementary Table 1). Based on the UCYN-A *nifH* libraries described above, the UCYN-A2/A3 qPCR assay was quantifying UCYN-A2 in the SIO samples.

In contrast, at Station ALOHA, both UCYN-A1 and UCYN-A1 hosts were detected at gene copy ratios averaging $3 *10^5 \pm 2.5*10^4$ *nifH* copies L^{-1} and $3.3*10^5 \pm 1.4*10^4$ 18S rRNA gene copies L^{-1}, respectively (Figure 5). Likewise, the gene copies of UCYN-A2/UCYN-A3 *nifH* gene were on average $6*10^4 \pm 4.6*10^3$ *nifH* copies L^{-1} and the 18S rRNA gene of the UCYN-A2 host averaged $2*10^4 \pm 3.1*10^3$ rRNA gene copies L^{-1}. These numbers suggest a ratio of about 3 UCYN-A cells per host (Figure 5), although the 18S rRNA gene copy number for the host is not known. Based on the sequencing results described above, the UCYN-A2/A3 qPCR assay was quantifying UCYN-A3 in the Station ALOHA samples.

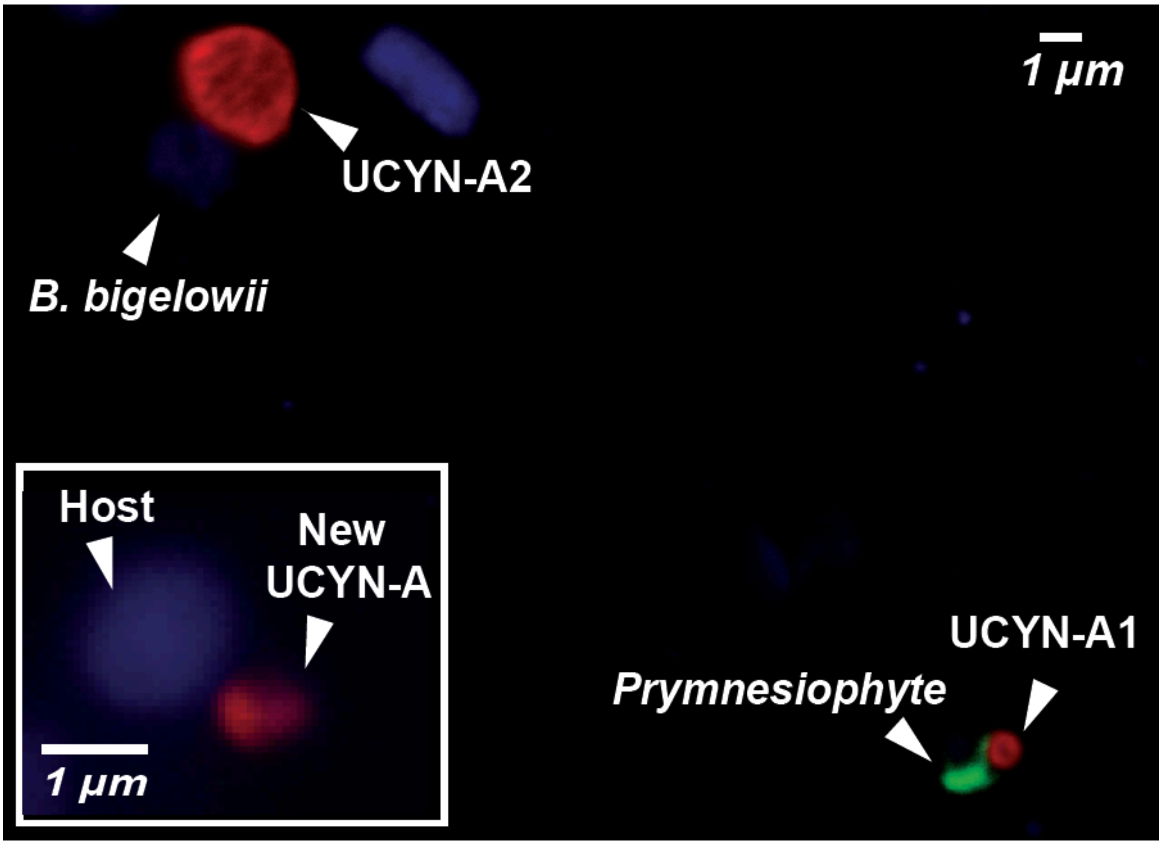


Figure 4. New UCYN-A association from the SIO Pier. Epifluorescence microscopy images with the double-CARD-FISH assay showing Host-A1 labeled with UPRYM69 probe with the competitor and all UCYN-A with the UCYN-A1-732 without competitor. White box shows an inset image of the new UCYN-A association found in the same slide with the same CARD-FISH probes. Similar new association was found using Host-A2 labeled with UBRADO69 probe with the competitor and all UCYN-A with the UCYN-A1-732 without competitor (non shown).

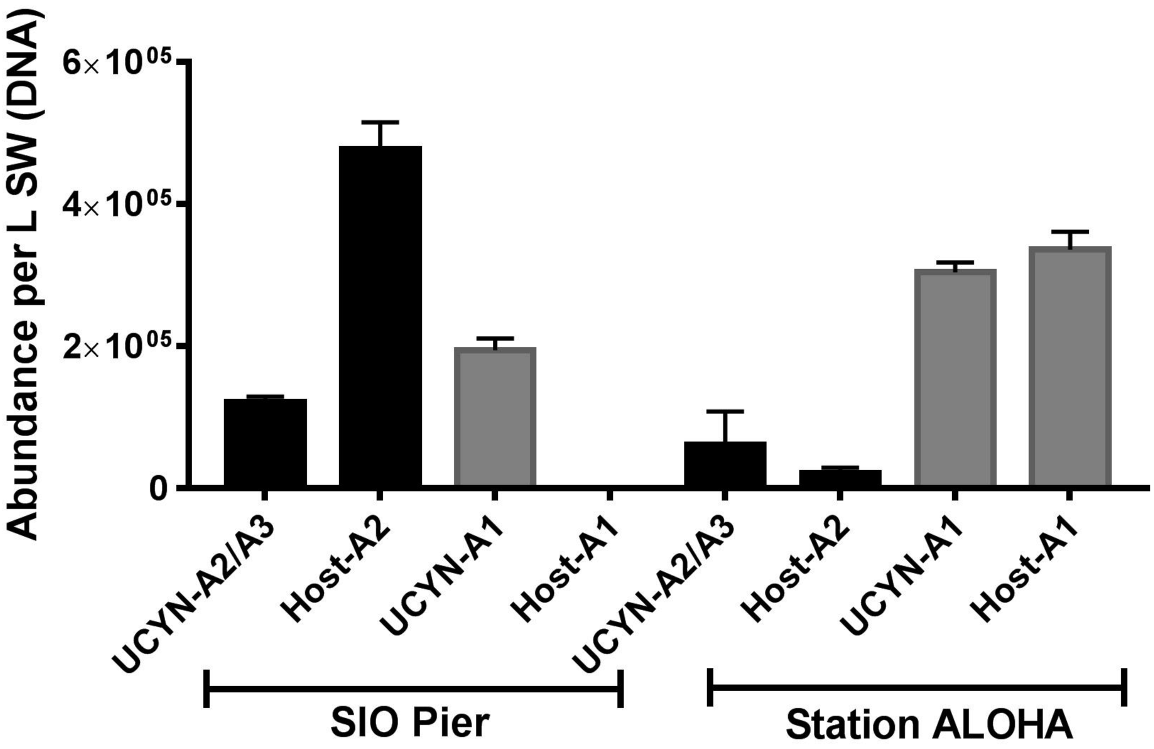


Figure 5. Gene copy averages of UCYN-A1 and UCYN-A2/A3 *nifH* and Prymnesiophyte (UCYN-A1 host) and *B. bigelowii* (UCYN-A2 host) 18S rRNA genes determined by qPCR at SIO Pier and Station ALOHA over three days.

2. UCYN-A3 characterization from metagenomes

2.1 Metagenomic detection of a new UCYN-A population.

During the *Tara* Oceans expedition, the distribution of the UCYN-A1 and UCYN-A2 sublineages was analyzed by metagenome fragment recruitment at stations TARA_078 and TARA_076 in the South Atlantic Ocean in several plankton size-fractions (0.2-3, 0.8-5, 5-20 and >0.8 μm) (Cornejo-Castillo *et al.*, 2016). In order to detect new divergent UCYN-A populations, these metagenomes were re-analyzed in this study. Metagenomic reads belonging to UCYN-A1, UCYN-A2 as well as closely related sequences (see Material and Methods) were explored by fragment recruitment (Supplementary Table 3).

Reads assigned to the UCYN-A1 sublineage (>95% identity with UCYN-A1 genome) were primarily present in the 0.2-3, 0.8-5 and >0.8 μm size fractions at stations TARA_078 and TARA_076 (Supplementary Table 3), encompassing the cell size ranges of the small prymnesiophyte partner (<3 μm) obtained in all previous studies (Thompson *et al.*, 2012, 2014; Cornejo-Castillo *et al.*, 2016; Martínez-Pérez *et al.*, 2016). Likewise, reads assigned to the UCYN-A2 sublineage (>95% identity with UCYN-A2 genome) were in the size-fraction range of *B. bigelowii* at station TARA_078 (0.8-5, 5-20 and >0.8 μm) in agreement with previous studies (Thompson *et al.*, 2014; Cabello *et al.*, 2015; Cornejo-Castillo *et al.*, 2016). Interestingly, at station TARA_076, UCYN-A2 was absent but recruitment of reads with sequence identities less than 95% to the UCYN-A2 genome appeared in metagenomes from both the 0.8-5 and >0.8 μm size-fractions (Figure 6, Supplementary Table 3). The recruitment pattern appeared within the size-fraction corresponding with the cell size-range of the UCYN-A3 host (3.6 μm ± **0.7** μm; n=100 cells) detected at Station ALOHA in this study (Supplementary Figure 2). These findings suggest that this new divergent UCYN-A genome sequence population is UCYN-A3.

2.2 Metagenomic reconstruction of an environmental UCYN-A3 genome

To gain insight into the gene content of the new divergent UCYN-A3 population detected in fragment recruitment, metagenomic reads recruited from the 0.8-5 and >0.8 μm size fractions of the two surface TARA samples (TARA_076 and TARA_078) were co-assembled (Supplementary Figure 1). A total of 180,557 bp of the UCYN-A3 genome, summarized in 247 contigs containing 293 genes (including the *nifH* gene), were assembled (Supplementary Table 4). Considering that the genome sizes of the two UCYN-A genomes sequenced to-date is approximately 1.4 Mb, it can be assumed that approximately 13% of the UCYN-A3 genome was assembled. In order to verify that these genes belonged to the UCYN-A clade, every single gene was compared to GenBank using BLASTN against the nt database. The best hit for almost all of the reconstructed genes was UCYN-A2. A few genes were closer to UCYN-A1, however, since the UCYN-A2 genome is not completely closed, these genes could be closer to the homologous genes missing in UCYN-A2. When compared with the UCYN-A2 genome, some of the genes that were missing in the UCYN-A1 genome were present in the UCYN-A3 genome. Additionally, in terms of gene synteny, all contigs containing more than one gene had the same gene ordination as in the UCYN-A2 genome (Supplementary Table 4).

2.3 Phylogenomic analysis and evolution of the UCYN-A3 sublineage

A phylogenetic tree was constructed to place the new UCYN-A3 sublineage in its evolutionary context. Maximum likelihood analysis of a total of 165 protein-coding genes (Supplementary Table 4; genes marked with asterisk) shared by closely-related N_2-fixing cyanobacteria confirmed that UCYN-A3, together with UCYN-A1 and UCYN-A2, form a well-supported monophyletic group (Figure 7). Interestingly, UCYN-A3 formed a sub-group with UCYN-A2 (Figure 7).

We explored the possible causes of the evolutionary diversification of UCYN-A3 by studying the selection pressure acting on the protein-coding genes shared by both UCYN-A2 and UCYN-A3 sublineages. We calculated the number of synonymous or silent (Ks) and non-synonymous (Ka, inducing amino acid change) nucleotide substitutions in these genes. The Ka/Ks ratio indicates whether purifying (<1) or positive (>1) selection has happened between phylogenetically closely-related organisms (McDonald and Kreitman, 1991). We assessed the Ka/Ks ratio for 291 protein-coding genes shared by the UCYN-A2 and UCYN-A3 sublineage (Supplementary Table 5). We found that only 1 out of 291 genes was under positive selection, in particular, a gene coding for the subunit I of the cytochrome C oxidase. The vast majority of the genes, 261 out of 291, were subjected to purifying selection; and the rest of the analyzed genes did not show significant results (P>0.05, Codon Based Z-test) (Supplementary Table 5).

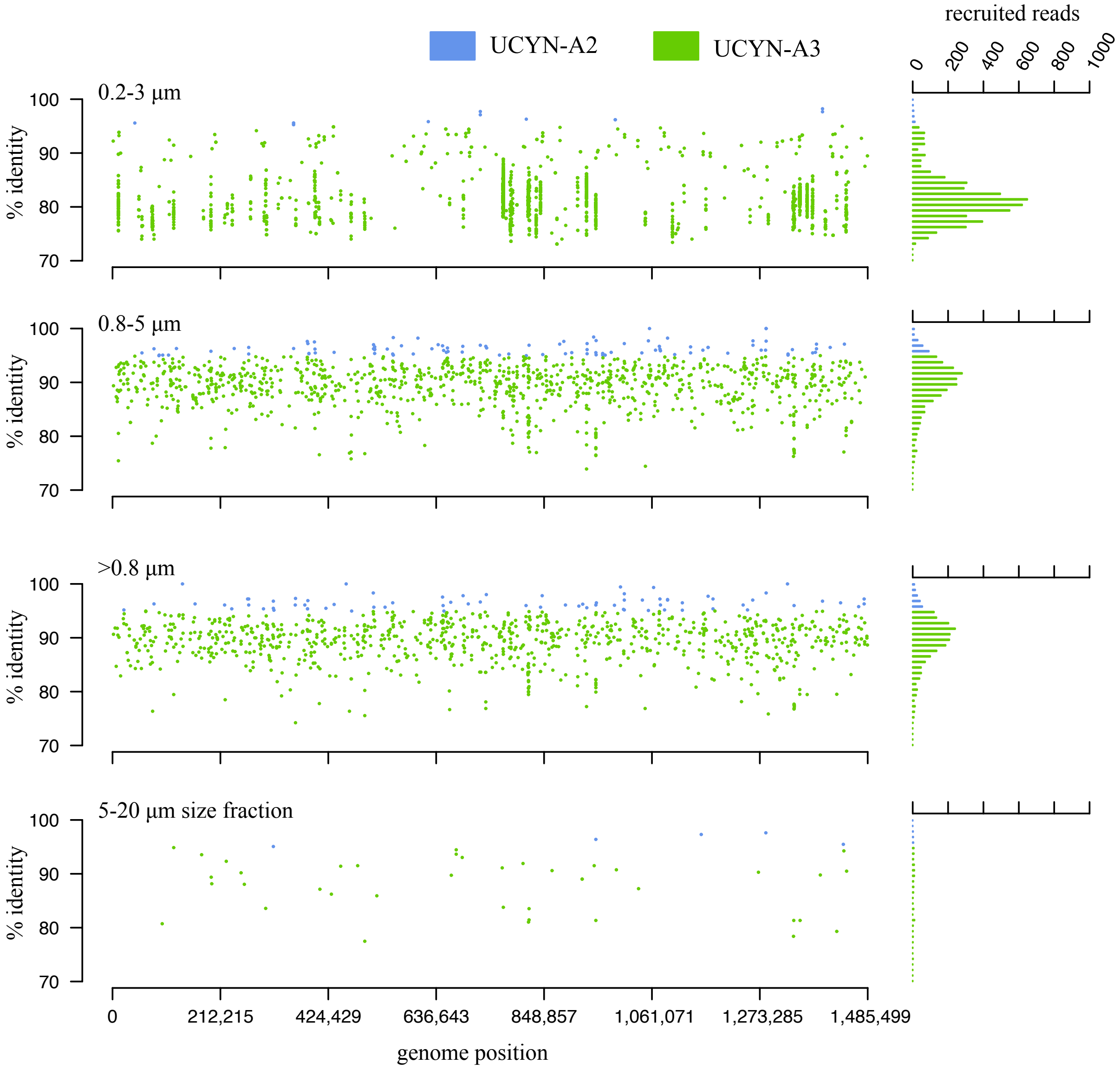


Figure 6: **Metagenome fragment recruitment of UCYN-A lineages in size-fractionated metagenomes from surface waters collected at station TARA_076.** Recruitment plot of metagenomic reads more similar to UCYN-A2 genome sequences, using UCYN-A1 and UCYN-A2 genomes as reference genomes. Reads are plotted as colored dots, representing the covered genome positions (x axis) and the % of identity with the UCYN-A2 reference genome (y axis). Reads with identity higher or lower to 95% are considered to represent UCYN-A2 (blue dots) or UCYN-A3 (green dots) populations respectively. On the right side, histograms represent the number of recruited reads in intervals of 1% identity, from 100 to 70% identity.

Discussion

UCYN-A3, a new UCYN-A clade

To characterize the UCYN-A population structure in two different marine environments we used a combination of established methods including qPCR assays targeting the *nifH* genes of UCYN-A1 and UCYN-A2/UCYNA3 sublineages and the 18S rRNA gene of the UCYN-A1 and UCYN-A2 hosts, visualization using double CARD-FISH, as well as Illumina sequencing of UCYN-A *nifH* gene fragments. The presence or absence of each symbiont and its partner at both stations based on the utilization of the different techniques is summarized in Supplementary Table 6. The multi-approach strategy revealed some discrepancies between techniques in the identification of each sublineage that led us to new insights into the diversity of the UCYN-A symbiosis.

UCYN-A1 *nifH* gene sequences were detected at both stations at relatively high abundances, and CARD-FISH analysis indicated consistent morphology in both environments. We measured a ratio of *nifH*:18S rRNA genes close to 1:1 (~0.9) at Station ALOHA, which is consistent with the 1:1 symbiont-host cell ratio observed previously (Thompson *et al.*, 2014).

Based on qPCR and CARD-FISH assays, it originally appeared that UCYN-A2 was present at both stations. However, we have several lines of evidence supporting that the association detected at Station ALOHA was in fact UCYN-A3. First, UCYN-A3 *nifH* amplicon sequences were present at relative abundances ~ 10% at Station ALOHA, while UCYN-A2 sequences were virtually absent, which is consistent with a recent report of UCYN-A3 in the NPSG (including Station ALOHA) by Turk-Kubo *et al.*, (2017). Second, the UCYN-A2 qPCR assay cannot be used to distinguish between UCYN-A2 and UCYN-A3, as it does not contain sufficient mismatches to prevent cross-hybridization with UCYN-A3 (and UCYN-A4) sublineages (Farnelid *et al.*, 2016). Finally, the discrepancy in cell sizes between the two associations observed at the SIO Pier and at Station ALOHA targeted with CARD-FISH probes originally assumed to be specific for UCYN-A2/*B. bigelowii*, indicates that the UCYN-A association originally assumed to be UCYN-A2 at Station ALOHA is actually UCYN-A3. Thus it appears that the CARD-FISH probes designed to target UCYN-A2/*B. bigelowii* also target UCYN-A3 and its unknown host.

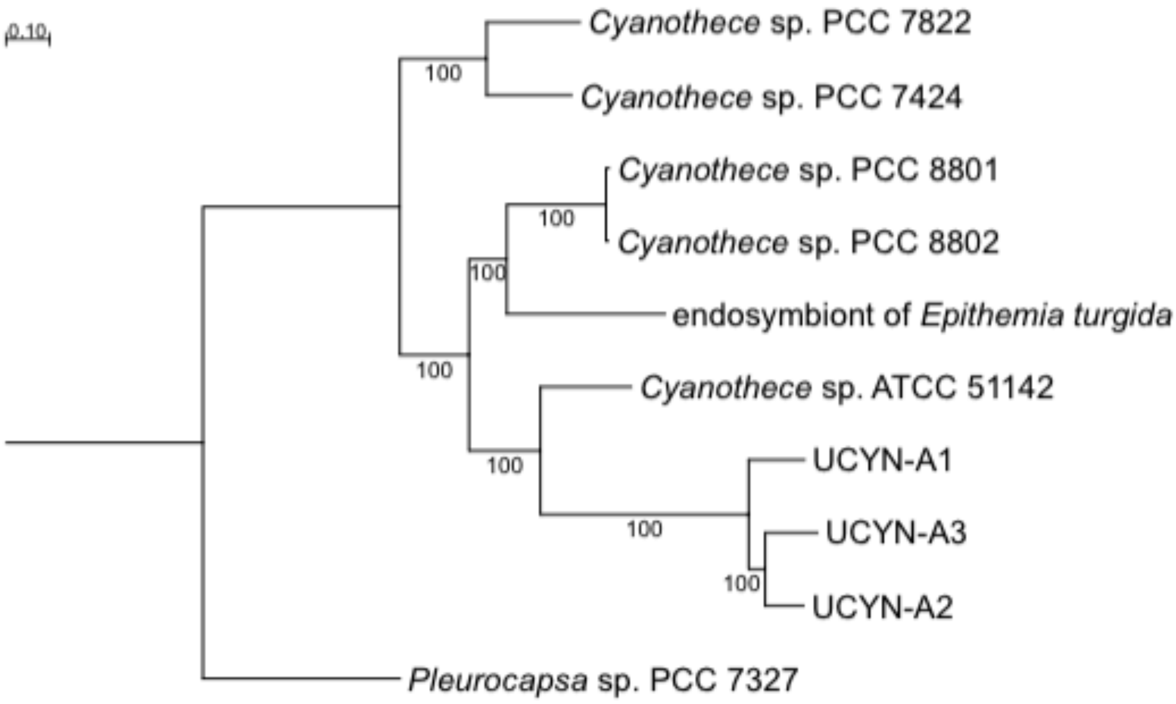


Figure 7: **Maximum likelihood tree from analysis of a total of 165 protein-coding genes shared by UCYN-A sublineages and nine closely related cyanobacteria.** *Pleurocapsa* sp. PCC 7326 was used as outgroup. The data set was bootstrapped 100 times, and bootstrap values are shown.

Previous studies have also reported a smaller UCYN-A2 host (Cabello *et al.*, 2015; Martínez-Pérez *et al.*, 2016) suggesting that in the open ocean the UCYN-A2 host might be smaller (4-5 µm) than at coastal sites (7-10 µm). These studies, where the UCYN-A2 host was smaller than reported at SIO Pier, were most likely detecting the UCYN-A3 host instead. In order to determine this possibility, we recruited metagenomes from the TARA_076 station of the *Tara* Oceans expedition where UCYN-A2 was not detected. Instead of the UCYN-A2 genome, a divergent UCYN-A genome, UCYN-A3, was detected in the size-fractions consistent with the size range observed for the UCYN-A3 association described above.

Prior to this study, UCYN-A3 had been defined as a sublineage based solely on the phylogeny of UCYN-A *nifH* sequences. One of the reconstructed genes in TARA_076 was the *nifH* gene, which was key to identify this new divergent population as UCYN-A3. The present study goes further by defining UCYN-A3 not only as a *nifH* sequence variant but also as a new UCYN-A genomic species, since the partial reconstruction of its genome revealed a new UCYN-A divergent genome. The phylogenetic relationships among UCYN-A sublineages showed that UCYN-A3 is closer to UCYN-A2 than to UCYN-A1 and that UCYN-A3 share a more recent common ancestor with UCYN-A2 than the ancestor shared with UCYN-A1. Therefore, since the divergence of the UCYN-A1 and UCYN-A2 sublineages was estimated to occur approximately 91 Myr ago (Cornejo-Castillo *et al.*, 2016), the UCYN-A2 and UCYN-A3 sublineages would have diverged more recently (Figure 7). The selection pressure showed a large-scale purifying or stabilizing pressure acting on the UCYN-A2 and UCYN-A3 sublineages. In fact, this may suggest that the last common ancestor of UCYN-A2 and UCYN-A3 sublineages were adapted to the same habitat (or to the same host) before they diverged into different sublineages. A possible scenario would be that the last common ancestor of UCYN-A2 and UCYN-A3 lived in symbiosis with a common prymnesiophyte host that diversified into different species and, as a consequence, the host speciation caused the isolation of the UCYN-A population that was the origin the UCYN-A2 and UCYN-A3 sublineages.

Partner fidelity

The diversity of the UCYN-A *nifH* gene reported here is similar to other studies, with the sublineages UCYN-A1 and UCYN-A3 co-occurring in the open ocean and the sublineage UCYN-A2 co-occurring with UCYN-A4 in coastal waters (Farnelid *et al.*, 2016; Turk-Kubo *et al.*, 2017) (Figure 1). The co-occurrence of UCYN-A sublineages suggests that distinct ecotypes have overlapping niches. However, in this study UCYN-A1 was also found at the SIO Pier in two different years, where it had not been observed previously (Thompson *et al.*, 2014). The different distributions of UCYN-A diversity in this coastal water might be explained by seasonal dynamics. In fact, there is evidence of temporal shifts of *Synechococcus* clades at the SIO Pier (Tai and Palenik, 2009) and the same potential environmental factors that result in shifts of *Synechococcus* clades could also similarly affect the distribution of UCYN-A sublineages.

UCYN-A strains could potentially vary in their associations with different hosts. It is currently known that UCYN-A1 and UCYN-A2 sublineages are specifically associated with distinct hosts (Cornejo-Castillo *et al.*, 2016). However, with the increasing number of UCYN-A sublineages it is unknown whether new relationships involving new hosts and/or more 'promiscuous' UCYN-A strains may exist (Figure 8). Despite being able to visualize the host-A1 (prymnesiophyte) using the same CARD-FISH probe at both stations, we could not identify the 18S rRNA gene of the BIOSOPE T60.34 sequence (GenBank accession no. FJ537341) at the SIO Pier using qPCR. This suggests that although the prymnesiophyte host cells from both stations look morphologically similar, they do not have identical 18S rRNA gene sequences. However, we assume that the UCYN-A1 host detected at the SIO Pier is closely related to the originally detected UCYN-A1 prymnesiophyte host since the Host-A1 probe for CARD-FISH hybridized with the host (Figure 2A, 2B).

In contrast to the UCYN-A2/*B. bigelowii* ratio of about 3 reported by Thompson *et al.*, (2014), in our samples, we observed the opposite pattern, that is, the host-A2 (*B. bigelowii*) was almost 4 times more ratio than the cyanobacterium UCYN-A2. A recent publication suggested a possible correlation between rDNA copy number and organism size (de Vargas *et al.*, 2015). Based on this correlation, *B. bigelowii,* with a size between 7-10 µm, could have approximately 3 copies rRNA genes for cell, which could explain our qPCR ratios. In contrast, the ratio reported by Thompson *et al.*, (2014) cannot be explained if the host has 3 copies per cell (de Vargas *et al.*, 2015).

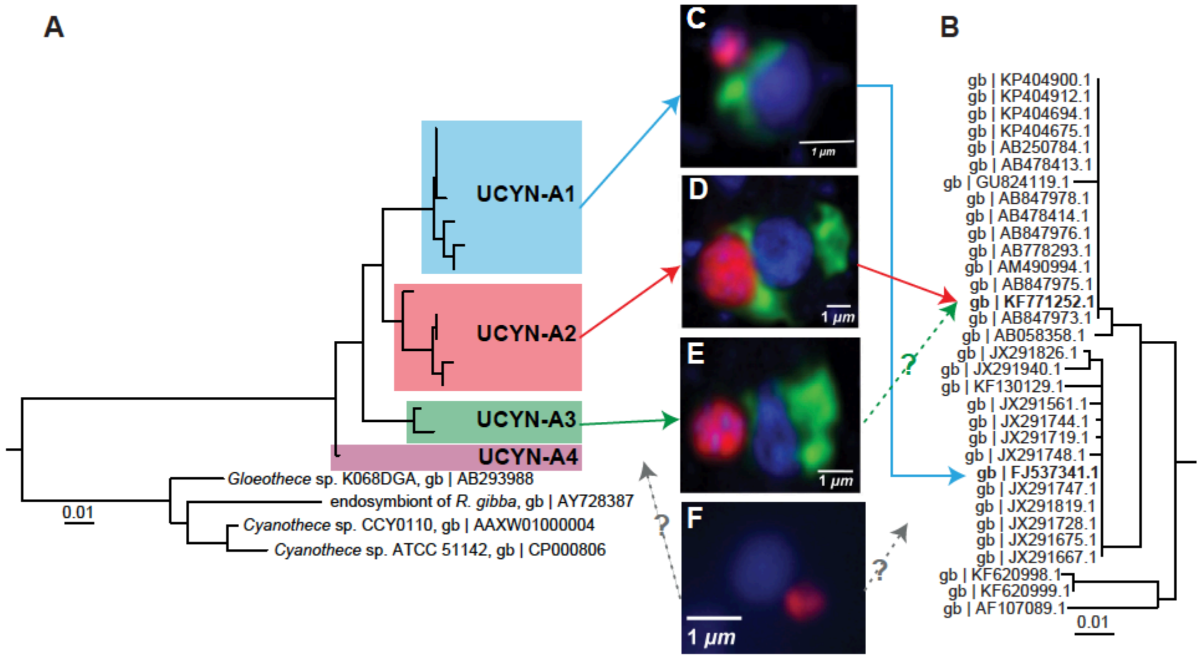


Figure 8. **Synthesis of diversity, morphological features and host specificity for the known UCYN-A sublineages**. A) Maximum likelihood nucleotide tree based on partial *nifH* genes from representative sequences from each UCYN-A sublineage. B) Maximum likelihood tree based on 18S rRNA gene fragments from sequences closely related to UCYN-A1 (gb|FJ537341.1) and UCYN-A2 (gb|KF771252.1) hosts in the Genbank nr database aligned using Silva. The tree was constructed in MEGA. Epifluorescence microscopy images of UCYN-A1 (C), UCYN-A2 (D), UCYN-A3 (E) and an unknown UCYN-A sublineage (F), described in detail in Fig. 2, 3 and 4. Solid arrows indicate known links between sublineages and morphologies and host associations; dotted arrows indicate that the links between sublineages and morphologies and/or host associations are not known.

To investigate this discrepancy, we evaluated the specificity and cross-hybridization of the UCYN-A2 host qPCR assay using all sequences available in Genbank. Using primers (forward and reverse) and the hybridization probe (internal oligo), the 18S rRNA genes from a variety of different sequences were amplified *in silico*. The sequences, which had 100% nucleotide identity with *B. bigelowii* across the primer probe sites, included sequences from other haptophytes such as

Haptolina brevifilum (AM490995.2), *Chrysochromulina parkeae* (AM490994.1), *Chrysochromulina brevifilum* MBIC10518 (AB058358.1) and one unidentified haptophyte clone OTU5 (KF878252.1). These results showed that the primers used in this work for the UCYN-A2 host and previously by Thompson *et al.* (2014) were not specific for *B. bigelowii*, and that these primers could have amplified different eukaryotes that were not UCYN-A hosts.

The non-specificity of the UCYN-A2 host qPCR assay explains the detection of the "UCYN-A2 host" at Station ALOHA ($2.6*10^4$ rRNA gene copies L^{-1}), despite the very low relative ratio of UCYN-A2 sequences based on *nifH* sequencing (Figure 1). However, these results along with the CARD-FISH analyses do suggest that UCYN-A3 could be associated with a host genetically similar to the host of UCYN-A2, with a cell size intermediate between the sizes of the UCYN-A1 and UCYN-A2 hosts (3.61 ± 0.67 µm (Figure 3C, 3D)).

Conclusions and future directions

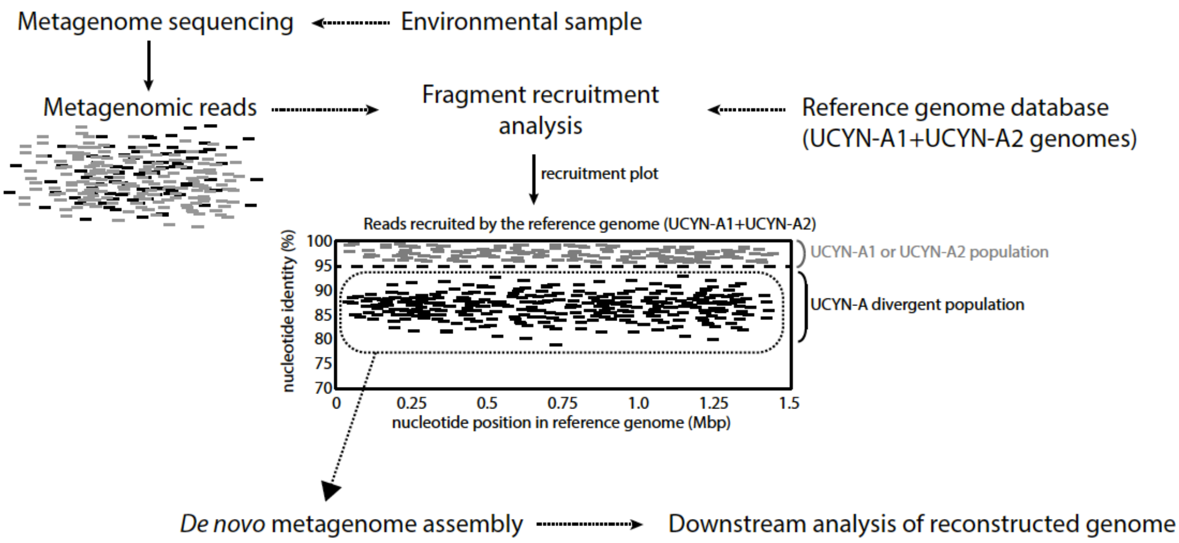
Our results obtained based on CARD-FISH counts, together with the UCYN-A *nifH* gene sequences and qPCR ratios provide new insights into the global distribution of UCYN-A sublineages. The wide occurrence of multiple UCYN-A sublineages and prymensiophyte hosts that vary in size has implications for N_2 fixation rates as they could have been hidden in higher compartments and, consequently, UCYN-A may contribute to N_2 fixation in a much larger oceanic area than previously thought.

Moreover, this is the first study reporting microscopic images of the UCYN-A3 sublineage, and a partial genome using a novel approach combining fragment recruitment and genome assembly techniques. With the availability of these UCYNA-3 gene sequences, the recruitment of the whole genome and metatranscriptomic should be possible in the near future. Thus the discovery and characterization of new UCYN-A sublineages and their hosts will help to determine the significance of these biogeochemically relevant microorganisms.

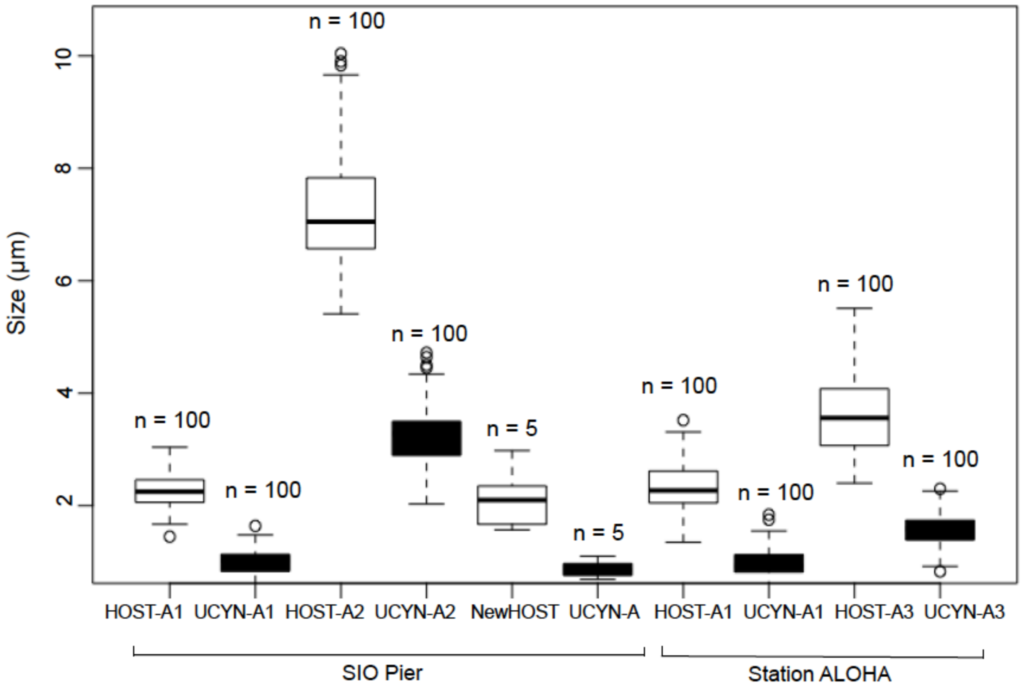
Finally, as we learn more about the diversity of UCYN-A, it is clear that we must compare and validate different assays and methods. Furthermore, there is a great need for the development of molecular probes/primers that can specifically target distinct UCYN-A sublineages which is critical for elucidating the ecology and the evolution of the UCYN-A symbiosis.

SUPPLEMENTARY MATERIAL

Supplementary Figure 1. Overview of the strategy followed to reconstruct the UCYN-A3 genome.


Metagenome sequencing
Environmental sample
Metagenomic reads
Fragment recruitment analysis
Reference genome database (UCYN-A1+UCYN-A2 genomes)
recruitment plot
Reads recruited by the reference genome (UCYN-A1+UCYN-A2)
nucleotide identity (%)
100
95
90
85
80
75
70
0
0.25
0.5
0.75
1.0
1.25
1.5
nucleotide position in reference genome (Mbp)
UCYN-A1 or UCYN-A2 population
UCYN-A divergent population
De novo metagenome assembly
Downstream analysis of reconstructed genome

Supplementary Figure 2. Size of the UCYN-A lineage at the SIO Pier and Station ALOHA Abbreviation: n, total number of cells observed using an Imaging Fluorescent Microscope (Zeiss, Berlin, Germany).



Supplementary Table 1: List of primers and probes utilized in quantification of gene abundance by quantitative PCR.

Gene	Primer	Sequence (5' to 3')	Reference
nifH UCYN-A1	Forward	AGCTATAACAACGTTTTATGCGTTGA	Church *et al*. (2005)
	Reverse	ACCACGACCAGCACATCCA	Church *et al*. (2005)
	Probe	TCTGGTGGTCCTGAGCCTGGA	Church *et al*. (2005)
18S rRNA Host-A1	Forward	AGGTTTGCCGGTCTGCCGAT	Designed by Thompson (not published)
	Reverse	ATCCGTCTCCGACACCCGCTC	Designed by Thompson (not published)
	Probe	CTGGTAGAACTGTCCTTCC	Designed by Thompson (not published)
nifH UCYN-A2	Forward	GGTTACAACAACGTTTTATGTGTTGA	Thompson *et al*. (2014)
	Reverse	ACCACGACCAGCACATCCA	Church *et al*. (2005)
	Probe	TCTGGTGGTCCTGAGCCCGGA	Thompson *et al*. (2014)
18S rRNA Host-A2	Forward	GGTTTTGCCGGTCTGCCGTT	Thompson *et al*. (2014)
	Reverse	ATCCGTCTCCGACACCCACTC	Thompson *et al*. (2014)
	Probe	CTGGTGCGAGCGTCCTTCCT	Thompson *et al*. (2014)

Supplementary Table 2: List of probes utilized in the visualization of UCYN-A associations by double CARD-FISH.

Probe Name	**Target**	**Sequence (5' to 3')**	**Reference**
UPRYM69	Host-A1	CACAT**A**GGAACATCCTCC	Cornejo-Castillo *et al.*. (2016)
UPRYM69 competitor	Host-A2 used as Host-A1 competitor	CACAT**T**GGAACATCCTCC	Cornejo-Castillo *et al.*. (2016)
UBRADO69	Host-A2	CACAT**T**GGAACATCCTCC	Cornejo-Castillo *et al.*. (2016)
UBRADO69 competitor	Host_A1 used as Host-A2 competitor	CACAT**A**GGAACATCCTCC	Cornejo-Castillo *et al.*. (2016)
Helper A-PRYM	*Haptophyta*	GAAAGGTGCTGAAGGAGT	Cornejo-Castillo *et al.*. (2016)
Helper B-PRYM	*Haptophyta*	AATCCCTAGTCGGCATGG	Cornejo-Castillo *et al.*. (2016)
UCYN–A1 732	UCYN-A1	GTT**A**CGGTCCAGTAGCAC	Krupke *et al.*. (2013)
UCYN-A1 competitor	UCYN-A2 used as A1 competitor	GTT**G**CGGTCCAGTAGCAC	Cornejo-Castillo *et al.*. (2016)
UCYN–A2 732	UCYN-A2	GTT**G**CGGTCCAGTAGCAC	Cornejo-Castillo *et al.*. (2016)
UCYN-A2 competitor	UCYN-A1 used as A2 competitor	GTT**A**CGGTCCAGTAGCAC	Krupke *et al.*. (2013)
Helper A–732	UCYN-A	GCCTTCGCCACCGATGTTCTT	Krupke *et al.*. (2013)
Helper B–732	UCYN-A	AGCTTTCGTCCCTGAGTGTCA	Krupke *et al.*. (2013)

Supplementary Table 3: Fragment recruitment of UCYN-A lineages.

Station	Metagenome fraction (µm)	Sequencing depth (reads)	UCYN-A1				UCYN-A2			
			Number of recruited reads		Genome coverage (%)		Number of recruited reads		Genome coverage (%)	
			≥ 95% identity	< 95% identity	≥ 95% identity	< 95% identity	≥ 95% identity	< 95% identity	≥ 95% identity	< 95% identity
TARA_076	0.2-3	177,019,968	188,095	3,383	99.28	3.62	26	4,482	0.14	3.78
TARA_076	0.8-5	68,314,651	131,215	871	98.43	7.07	196	2,113	1.65	16.60
TARA_076	>0.8	73,651,199	54,776	716	98.61	5.89	147	1,829	1.35	14.83
TARA_076	5-20	91,873,867	361	44	4.02	0.45	9	80	0.10	0.77
TARA_078	0.2-3	155,580,203	842,234	5,885	99.31	6.64	2,395	6,360	13.43	8.60
TARA_078	0.8-5	105,731,269	980,895	2,347	99.32	10.02	24,021	2,379	89.47	15.04
TARA_078	>0.8	163,575,710	719,803	2,971	99.32	12.49	81,528	5,123	99.03	25.90
TARA_078	5-20	139,070,786	1,182	121	9.99	0.68	17,028	209	75.69	1.60

Supplementary Table 6. Presence or absence of each symbiont and its partner at SIO Pier and Station ALOHA based on the utilization of the different techniques. Gray boxes show the interpretation of the different results.

ASSAYS	SIO Pier	Station ALOHA
UCYN-A *nifH* gene sequencing(1)	UCYN-A1, UCYN-A2, UCYN-A4	UCYNA1, UCYN-3
UCYN-A1 (qPCR)(2)	+	+
Host-A1 (qPCR)(3)	-	+
UCYN-A1 (CARD-FISH)(4)	+	+
Host-A1 (CARD-FISH)(4)	+	+
INTERPRETATION	**UCYN-A1 present at SIO is not associated with an identical host to that identified by Thompson *et al.*., (2012). However, the Host-A1 at SIO is closely related to the original Host-A1 since it hybridizes with the Host-A1 CARD-FISH probe.**	**UCYN-A1 is associated with the host identified previously by Thompson *et al.*., (2012).**
UCYN-A2/A3 (qPCR)(2)	+	+
Host-A2 (qPCR)(3)	+	+
UCYN-A2 (CARD-FISH)(4)	+	+
Host-A2 (CARD-FISH)(4)	+	+
INTERPRETATION	**UCYN-A2 is associated with the host identified previously by Thompson *et al.*., (2014). But the Host-A2 qPCR assay was not specific for *B. bigelowii* since it detected 4 times more Host-A2 gene copies (18S rRNA) compared to UCYN-A2 gene copies (*nifH*).**	**The sequencing and the discrepancy in sizes revealed that what was thought to be UCYN-A2 in Station ALOHA is most likely UCYN-A3. Also, UCYN-A3 could be in association with the Host-A2 or a closely related host since it hybridized with the Host-A2 CARD-FISH probe and it amplified in the Host-A2 qPCR assay.**
UCYN-A (universal) (CARD-FISH)(5)	+	+

(1) UCYN-A specific *nifH* gene amplification and sequencing using Illumina MiSeq Turk-Kubo *et al.*. (2017). (2) *nifH* gene qPCR assays designed for UCYN-A1 and UCYN-A2 sublineages by Thompson *et al.*. (2014) and Church *et al.*. (2005). (3) 18S rRNA gene qPCR assay targeting the identified hosts Host-A1 and Host-A2 designed by Thompson *et al.*. (2014). (4) CARD-FISH probes specific for known UCYN-A sublineages and hosts using the competitors designed by Cornejo-Castillo *et al.*. (2016). (5) UCYN-A1-735 probe without competitor probe. For more information see Materials and Methods.

REFERENCES

Bombar D, Heller P, Sanchez-Baracaldo P, Carter BJ, Zehr JP. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *ISME J*. http://dx.doi.org/10.1038/ismej.2014.167.

Cabello AM, Cornejo-Castillo FM, Raho N, Blasco D, Vidal M, Audic S, *et al.* (2015). Global distribution and vertical patterns of a prymnesiophyte-cyanobacteria obligate symbiosis. *ISME J*. **10**:693-706.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**:335.

Caro-quintero A, Konstantinidis KT. (2011). Bacterial species may exist, metagenomics reveal. *Environ Microbiol*. doi:10.1111/j.1462-2920.2011.02668.x.

Church MJ, Jenkins BD, Karl DM, Zehr JP. (2005). Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat Microb Ecol* **38**:3–14.

Cornejo-Castillo FM, Cabello AM, Salazar G, Sánchez-Baracaldo P, Lima-Mendez G, Hingamp P, *et al.* (2016). Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat Commun* **7**. doi:10.1038/ncomms11071.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**:2460–2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**:2194–2200.

Farnelid H, Turk-Kubo K, Del Carmen Muñoz-Marín M, Zehr JP. (2016). New insights into the ecology of the globally significant uncultured nitrogen-fixing symbiont UCYN-A. *Aquat Microb Ecol* **77**:128–138.

Green SJ, Venkatramanan R, Naqib A. (2015). Deconstructing the Polymerase Chain Reaction: Understanding and Correcting Bias Associated with Primer Degeneracies and Primer-Template Mismatches. *PLoS One* **10**:e0128122.

Hagino K, Onuma R, Kawachi M, Horiguchi T. (2013). Discovery of an endosymbiotic nitrogen-fixing cyanobacterium UCYN-A in Braarudosphaera bigelowii (Prymnesiophyceae). *PLoS One* **8**:e81749.

Heller P, Tripp HJ, Turk-Kubo K, Zehr JP. (2014). ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank . *Bioinformatics* **30**:2883–2890.

Karl D, Michaels A, Bergman B, Capone D, Carpenter E, Letelier R, *et al.* (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58(1)**:47–98.

Krupke A, Musat N, Laroche J, Mohr W, Fuchs BM, Amann RI, *et al.* (2013). In situ identification and N_2 and C fixation rates of uncultivated cyanobacteria populations. *Syst Appl Microbiol* **36**:259–71.

Letunic I, Bork P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**:127–8.

Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**:1674–1676.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, *et al.* (2004). ARB: a software environment for sequence data. *Nucleic Acids Res* **32**:1363–1371.

Luo Y-W, Doney SC, Anderson LA, Benavides M, Berman-Frank I, Bode A, *et al.* (2012). Database of diazotrophs in global ocean: abundance, biomass and nitrogen fixation rates. *Earth Syst Sci Data* **4**:47–73.

Martínez-Pérez C, Mohr W, Löscher CR, Dekaezemacker J, Littmann S, Yilmaz P, *et al.* (2016). The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol* **1**. doi:10.1038/nmicrobiol.2016.163.

McDonald JH, Kreitman M. (1991). Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**:652–654.

Moisander PH, Beinart RA, Hewson I, White AE, Johnson KS, Carlson CA, *et al.* (2010). Unicellular cyanobacterial distributions broaden the oceanic N2 fixation domain. *Science* **327**:1512–4.

Moisander PH, Beinart RA, Voss M, Zehr JP. (2008). Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *Isme J* **2**:954.

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, *et al.* (2012). Fiji: an open-source platform for biological-image analysis. *Nat Methods* **9**:676.

Sohm JA, Webb EA, Capone DG. (2011). Emerging patterns of marine nitrogen fixation. *Nat Rev Microbiol* **9**:499–508.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinforma* **22**:2688–2690.

Tai V, Palenik B. (2009). Temporal variation of Synechococcus clades at a coastal Pacific Ocean monitoring site. *Isme J* **3**:903.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Mol Biol Evol* **30**:2725–2729.

Thompson A, Carter BJ, Turk-Kubo K, Malfatti F, Azam F, Zehr JP. (2014). Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol* n/a-n/a.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, *et al.* (2012). Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science (80-)* **337**:1546–1550.

Thompson AW, Zehr JP. (2013). Cellular interactions: Lessons from the nitrogen-fixing cyanobacteria. *J Phycol* **49**:1024–1035.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**:90–94.

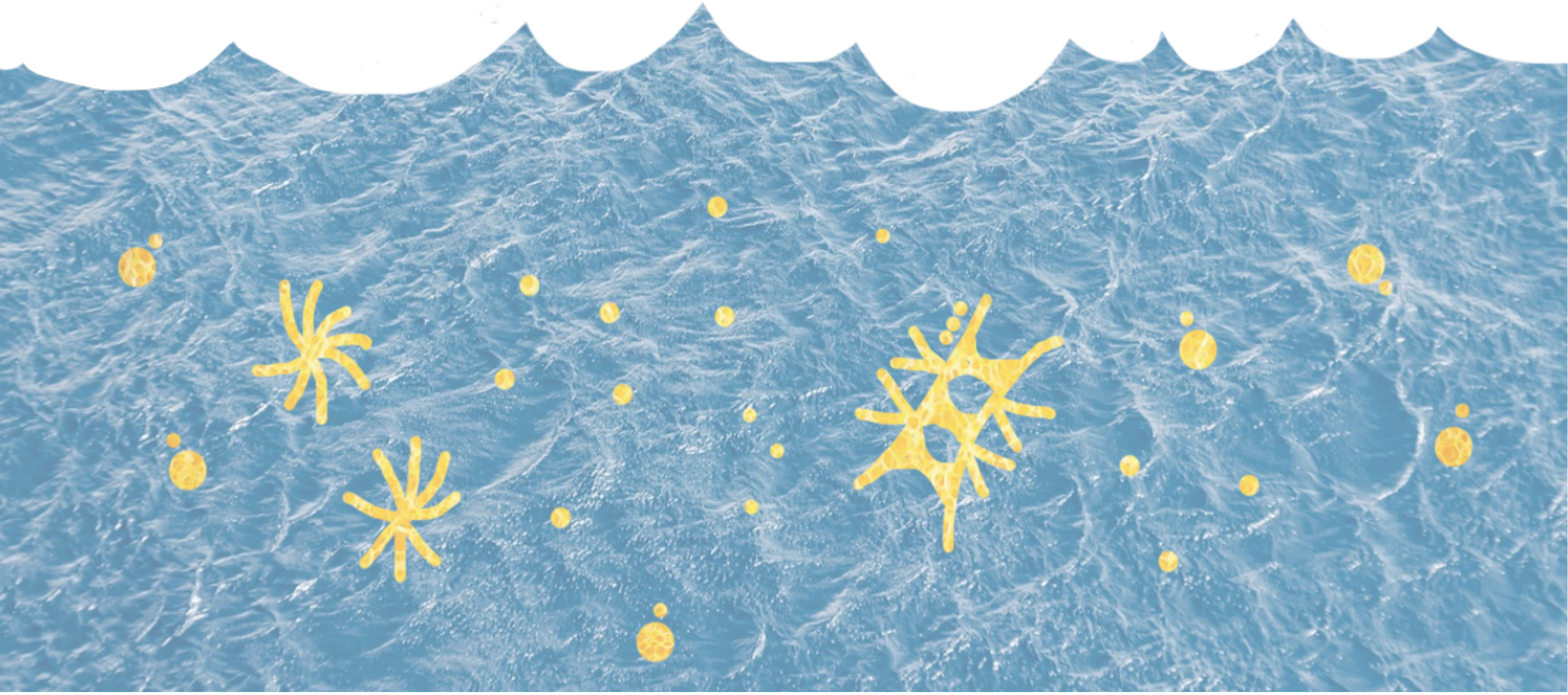
Turk-Kubo KA, Farnelid HM, Shilova IN, Henke B, Zehr JP. (2017). Distinct ecological niches of marine symbiotic N2-fixing cyanobacterium Candidatus Atelocyanobacterium thalassa sublineages. *J Phycol* **53**:451–461.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, *et al.* (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**:1261605.

Zani S, Mellon MT, Collier JL, Zehr JP. (2000). Expression of nifH genes in natural microbial assemblages in Lake George, New York, detected by reverse transcriptase PCR. *Appl Environ Microbiol* **66**:3119–3124.

Zhang J, Kobert K, Flouri T, Stamatakis A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* **30**:614–620.

General Discussion



General Discussion

Overview of the thesis and main scientific contributions

The general aim of this thesis was to gain insight into the diversity, ecology and evolution of the marine nitrogen-fixing microorganisms in the open ocean. The circumnavigation expedition *Tara* Oceans allowed us to address this goal through the unprecedented amount of genomic and biologic data generated at the global scale, which are already available in public repositories. In **Chapter 1** we conducted a global exploration of the *nifH* gene extracted from metagenomic data derived from 68 globally distributed stations. This approach is different from most surveys of *nifH* diversity conducted so far because it overcomes the biases caused by the use of primers in PCR or qPCR approaches (i.e., loss of diversity and biased quantification of the relative abundances of different diazotrophs). This study thus provides the first 'primer-free' global map of the distribution of open ocean diazotrophic communities across ocean basins and throughout the water column, from surface to mesopelagic waters. The metagenomic approach used allowed to estimate the contribution of these diazotrophic microorganisms within prokaryotic communities, something that has never been done before at this global scale, unveiling that they often occur at very low abundances, and suggesting that they belong to the so-called rare biosphere. Interestingly we found that in general, the abundance of diazotrophs was significantly higher in the mesopelagic than in photic surface or DCM waters, and that some of the detected diazotrophic groups showed contrasting habitat preferences. More importantly, we uncovered novel diversity that had remained unnoticed in all previous primer-based studies, since we demonstrate that more than half of the detected *nifH* variants showed mismatches in the primer-binding site with the primers used, specially with non-cyanobacterial diazotrophs. Therefore, our findings suggest that most diazotroph diversity studies may be disregarding an important fraction of the heterotrophic nitrogen-fixing community members, and may be one of the reasons explaining why studies rarely find a link between the taxonomic composition of these diazotrophs and nitrogen fixation rates. In addition, it implies that marine heterotrophic diazotrophic prokaryotes may be much more important than previously thought.

Among the diazotrophs detected in Chapter 1, the most abundant was the unicellular symbiotic cyanobacterium *C.* Atelocyanobacterium thalassa (UCYN-A), a key player in the marine nitrogen cycle. However, the metagenomic approach used here did not allow to differentiate between UCYN-A sublineages because they have small nucleotide divergences than can be hidden behind the 95% identity clustering threshold applied for the construction of the OM-RGC. Thus, in **Chapter 2** and **Chapter 3**, we explored in detail aspects related to the ecology, diversity and evolution of this remarkable microorganism.

Luckily, UCYN-A can be detected with the commonly used primer-based approaches, so its distribution has been widely studied. However, UCYN-A has circumvented all the attempts to keep and maintain it in culture to date and therefore, all current knowledge on UCYN-A biology has been conditioned to the fortune of detecting it in environmental samples. In this sense, we were lucky to find this cyanobacterium in high enough abundances in a couple of stations located in the South Atlantic Ocean, so the next two chapters focused on the exploration of UCYN-A at these two stations. In **Chapter 2**, we could detect this organism using not only metagenomic approaches but also visualization techniques like the CARD-FISH assay. This allowed us to unveil for the first time that UCYN-A1 and UCYN-A2 lineages live in symbiosis with two distinct prymnesiophyte partners with different cell size and that both symbiotic systems were lineage specific. Our analyses inferred a streamlined genome expression towards nitrogen fixation in both UCYN-A lineages and revealed a strong purifying selection in UCYN-A1 and UCYN-A2 with a diversification process about 91 Mya, in the late Cretaceous.

Finally, in **Chapter 3** we focused on the study UCYN-A3, another sublineage of which very little was known besides the existence of its *nifH* sequences. Using several methods (PCR, qPCR, CARD-FISH and metagenomes) to gain insight into the ecology of the different UCYN-A sublineages led to different interpretations of its ecology, revealing thus new information on the diversity of the UCYN-A symbiosis: for example, we could visually identify for the first time UCYN-A3 and its association with an alga of different size. Moreover we reconstructed a significant fraction of the UCYN-A3 genome establishing that this sublineage constitutes a new UCYN-A genomic species. Overall, Chapters 2 and 3 have largely expanded our knowledge of the ecology and evolution of UCYN-A lineages: We demonstrated that different UCYN-A lineages invested their genetic machinery to fix nitrogen for their respective hosts, whom they displayed partner fidelity, revealed new UCYN-A genomic species and placed the nitrogen fixation in novel planktonic compartments distributed along different size fractions of the plankton in accordance with the cell-size range of the lineage-specific UCYN-A hosts.

Overall, therefore, this thesis has significantly contributed to expand the knowledge on the marine nitrogen fixation unveiling new diazotrophic diversity and new planktonic compartments that can potentially contribute to a better understanding of the marine nitrogen cycle.

Methodological contributions of this thesis

Although numerous PCR primers have been designed to amplify *nifH* (Gaby and Buckley, 2012), the diversity of diazotrophs is still poorly described and many organisms remain to be discovered. Metagenomic analyses are PCR-independent and, therefore, not biased by primers designed on the basis of expectations of sequence conservation providing, thus, a potential source of new diazotrophic diversity. Our metagenomic analysis in Chapter 1 has helped to increase the repertoire of *nifH* genes allowing, moreover, the evaluation *in silico* of the nifH primers. Primer binding analysis showed that primers extensively used in marine diazotrophic diversity surveys (nifH1-nifH4) (Zehr *et al.*, 1998; Zani *et al.*, 2000) do probably not recover any of the *nifH* gene sequences of Cluster I (Gammaproteobacteria) and Cluster III (Deltaproteobacteria and Firmicutes) diazotrophs recruited in our study. In order to verify this affirmation, we searched the closest *nifH* gene sequences in NCBI generated via primer-based surveys (Table 1). Our results indicated that these primers would probably miss 100% of Cluster III diazotrophs detected in our study. Our contribution in this sense has been to modify the nifH4 primer in order to capture all that newly detected *nifH* diversity. This improvement of the nifH4 primer will hopefully help to have a more realistic view of the marine diazotrophic diversity in future PCR-based studies.

In Chapter 2 we designed new CARD-FISH probes to distinguish for the first time close but different lineages of the UCYN-A clade. The previous CARD-FISH probes did not allowed to distinguish them at such specific level and, consequently, had led to erroneous interpretations of the ecology of UCYN-A, like for instance to associate different UCYN-A lineages to different growth states (Krupke *et al.*, 2014) or to hypothesize a non-symbiotic particle-attached state of UCYN-A (Benavides *et al.*, 2013). In fact, our CARD-FISH probes have been successfully used in other studies that have contributed significantly to our current knowledge of the ecology of UCYN-A (Cabello *et al.*, 2015; Martínez-Pérez *et al.*, 2016).

OM-RGC gene ID	Closest *nifH* gene sequence in NCBI (nr/nt) from			
	primer-based surveys		non primer-based surveys	
Cluster I - Gammaproteobacteria	% identity	[ACCN]	% identity	[ACCN]
OM-RGC.v1.007436991	97%	[GU192754.1]	97%	[KF800062.1]
OM-RGC.v1.007483613	99%	[KX502144.1]	99%	[CP002622.1]
OM-RGC.v1.007595848	99%	[HM801587.1]	88%	[KJ021873.1]
OM-RGC.v1.007601814	98%	[HQ586568.1]	89%	[KJ021873.1]
OM-RGC.v1.007647800	99%	[KF151548.1]	89%	[KJ021873.1]
OM-RGC.v1.007667460	94%	[KF151548.1]	91%	[KJ021873.1]
OM-RGC.v1.011403932	99%	[KP260438.1]	91%	[CP001614.2]
OM-RGC.v1.026833116	100%	[HQ455956.1]	-	-
OM-RGC.v1.028674582	97%	[KF151548.1]	88%	[AF216883.1]
OM-RGC.v1.037241215	79%	[HQ224035.1]	83%	[CP002436.1]
Cluster III - Deltaproteobacteria				
OM-RGC.v1.007482987	89%	[JN601414.1]	100%	[LT907975.1]
OM-RGC.v1.008529501	87%	[HM750631.1]	78%	[CP000096.1]
OM-RGC.v1.008691244	88%	[HQ190142.1]	100%	[CP002364.1]
OM-RGC.v1.008734443	87%	[AY040518.1]	81%	[CP017237.1]
OM-RGC.v1.008759637	95%	[LC063964.1]	94%	[AP010904.1]
OM-RGC.v1.031135473	-	-	77%	[CP000096.1]
OM-RGC.v1.032496133	80%	[KX867946.1]	86%	[CP003220.1]
Cluster III - Firmicutes				
OM-RGC.v1.010396209	77%	[KF847385.1]	-	-
OM-RGC.v1.013419284	81%	[HQ224439.1]	80%	[CP003167.1]
OM-RGC.v1.031513582	76%	[HQ223500.1]	73%	[CP000254.1]

Table 1. **Comparison of *nifH* gene sequences recruited from OM-RGC and NCBI.** For each *nifH* gene sequence recruited from the OM-RGC, the identity (at the nucleotide level) with the closest *nifH* gene sequence found in NCBI (*nr/nt database*) is shown.

In Chapter 3, we designed a new strategy for genome reconstruction that could be applied in any kind of metagenomic studies. In our attempt to reconstruct the genome of close related UCYN-A lineages, we found that the current assembly methodology was not able to generate genomic contigs of the less abundant genomic species (in this case the less abundant was UCYN-A3 against UCYN-A1). We do not know what the reason behind this limitation is, but we guess that metagenomic reads belonging to both lineages did not show enough sequence divergence among lineages and, consequently, the assembly algorithm could not resolve the UCYN-A lineages as separate entities. Therefore, we decided to isolate the metagenomic read population belonging to each different lineage before performing the assembly process via fragment recruitment analysis (Suppl. Fig. 1 in Chapter 3). By these means, it is possible to explore not only whether a divergent genomic species is present in a particular sample (via metagenomic fragment recruitment) but also, whether the divergent species is active or not (via metatranscriptomic fragment recruitment, see example in Fig. 1).

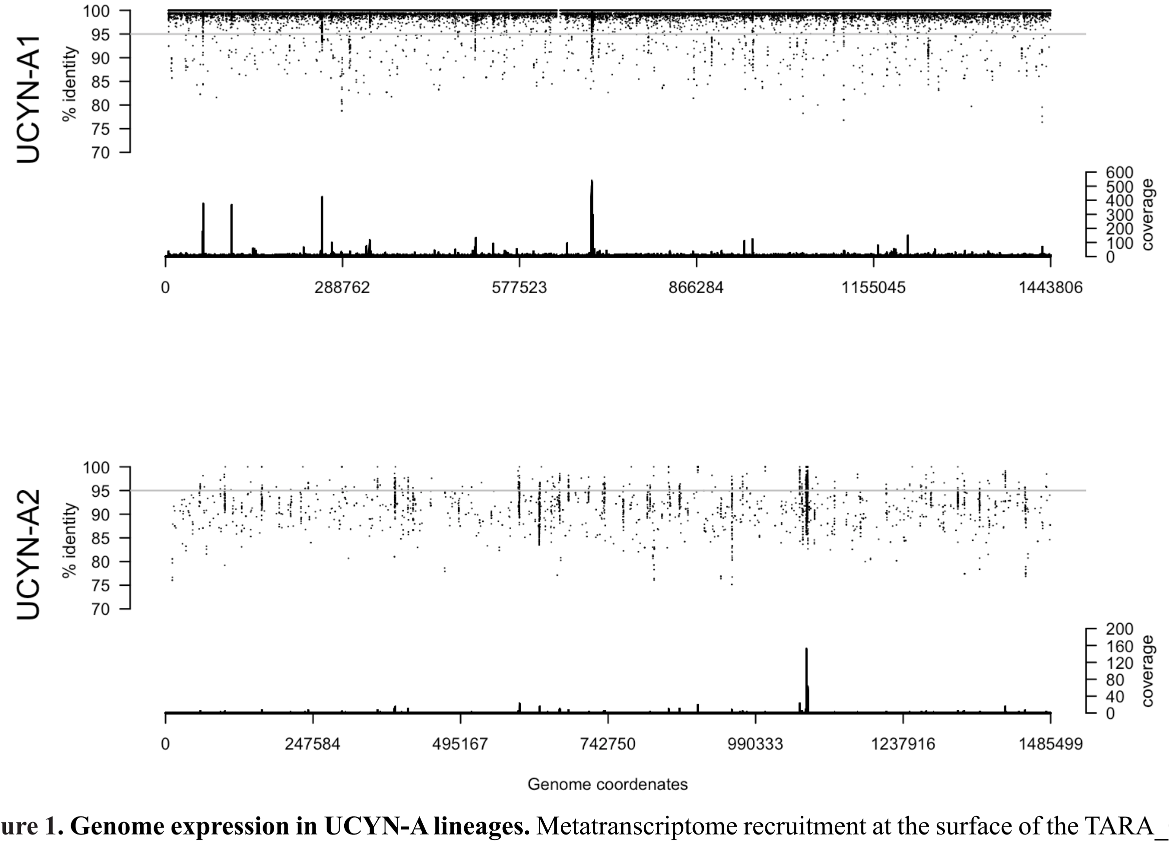


Figure 1. Genome expression in UCYN-A lineages. Metatranscriptome recruitment at the surface of the TARA_076 station of UCYN-A1 and UCYN-A3 transcripts in the >0.8 µm size fraction. We considered as UCYN-A3 those transcripts that showed less than 95% identity with the UCYN-A2 genome (this criteria is explained in Chapter 3). Transcripts are plotted as black dots representing the covered genome positions and the % of identity with the closest reference. A horizontal gray line set at 95% identity shows the threshold used to count the number of times, or coverage, that a gene was expressed.

The diazotrophic (rare) biosphere

Rare microorganisms may contribute significantly to oceanic geochemical processes (Giovannoni and Stingl, 2005; Campbell *et al.*, 2011). Our analysis from Chapter 1 provided a 'snapshot' of the abundance and diversity of diazotrophs in the global ocean. This 'snapshot' revealed that diazotrophs are often found in the rare biosphere, which is commonly assumed to be below the 0.1% relative abundance cut-off (Pedrós-Alió, 2012) with some punctual exceptions. It has been previously shown that both biotic and abiotic factors influence the abundance dynamics of rare microorganisms over time, which likewise abundant species, can also display different temporal dynamics presumably representing different life strategies (Lynch and Neufeld, 2015) (Fig. 2). Indeed, some of the diazotrophs detected here could be fitted in some of the categories depicted in Fig. 2. For example, UCYN-A, based on the 16S rRNA gene marker, was placed in the fourth position of the rank-abundance curve of the South Atlantic station TARA_078 (surface), with 1.28% of reads, whilst it was absent in most other stations (data extracted from Sunagawa *et al.*, 2015). This TARA_078 station had the peculiarity of being located within a 3-year-old Agulhas ring, an anti-cyclonic eddy characterized by a warm salty core and a 100 meters deeper pycnocline than surrounding waters (Villar *et al.*, 2015). It has been suggested that nitrogen fixation is enhanced inside anti-cyclonic eddies (Fong *et al.*, 2008; Church *et al.*, 2009) and it could be the cause behind the high abundance of UCYN-A in this station. Another example of a rare taxon becoming abundant was found in one of the Cluster I gammaproteobacterial diazotroph (OM-RGC.v1.007667460). This *nifH* variant maintained its relative abundance below 0.1% across all samples with two exceptions, in mesopelagic TARA_102 and surface TARA_110. These samples were likely influenced by the oxygen minimum zone (OMZ) off Peru, where nitrogen fixation has been demonstrated to occur, suggesting a close spatial coupling of N-input and N-loss processes (Loescher *et al.*, 2014). The third *nifH* variant that was above the 0.1% relative abundance cut-off was another Cluster I gammaproteobacteria (OM-RGC.v1.007601814) that, in only one sample (TARA_132 mesopelagic), reached a relative abundance of 0.37%. We do not know whether the diazotrophs mentioned above are periodically or occasionally recruited from the rare biosphere but, in any case, it has been demonstrated that, at least in the case of UCYN-A (Zehr *et al.*, 2016; Martínez-Pérez *et al.*, 2016), they are important nitrogen-fixing players. The rest of diazotrophs always showed relative abundances below 0.1%, so they might represent permanently rare, or transiently rare species that bloom under particular situations that we did not capture with our sampling strategy (Fig. 2). We did not find any biotic or abiotic factor explaining the abundance patterns of the detected diazotrophs, even though previous studies have shown that factors such as temperature, iron, or phosphorous can affect the presence and/or abundance of diazotrophs at the large spatial scale (Stal, 2009; Sohm *et al.*, 2011). Certainly, studies addressing the temporal variability of members within diazotrophic communities are needed in order to achieve a better understanding of the abundance and activity dynamics of nitrogen-fixing microorganisms.

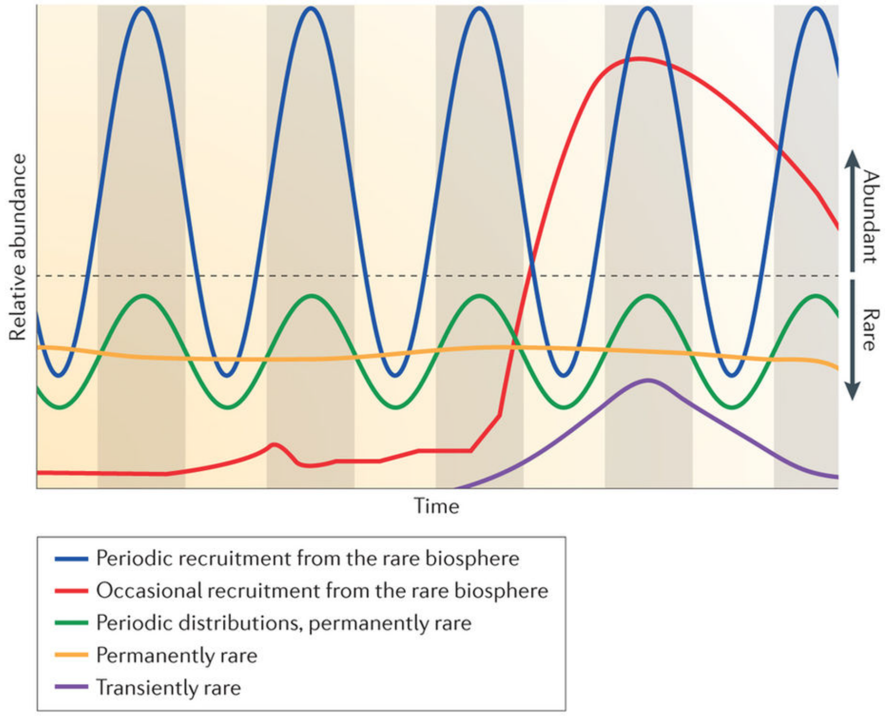


Figure 2. Hypothetical temporal abundance profiles for rare biosphere microorganisms. The scheme shows examples of different hypothetical temporal abundance dynamics of members within the rare biosphere, some of which may represent the behaviour of the different diazotrophic phylotypes described in Chapter 1. The grey shading represents seasonal variations (From Lynch and Neufeld, 2015).

Nitrogen fixation in different planktonic compartments

Nitrogen fixation has traditionally been placed in large planktonic size fractions, mainly occupied by *Trichodesmium* and the heterocystous endosymbiont Richelia. However, the discovery of marine unicellular cyanobacterial diazotrophs was crucial to redraw the size class where nitrogen fixation takes place in oligotrophic waters (Thompson and Zehr, 2013). The size class where nitrogen fixation occurs is important because it determines for example the efficiency of sinking and sequestration of the fixed nitrogen, and amount of fixed carbon in the case of photosynthetic diazotrophs or diazotrophs associated with photosynthetic hosts in deep waters (Karl *et al.*, 2012). In chapter 1 we targeted only the smallest planktonic fraction (0.2-3 μm), unveiling new diazotrophs never observed before and suggesting that this size fraction of the plankton may also contain potential new nitrogen-fixers. It is unlikely that they live in a permanent free-living state since nitrogen fixation requires the absence of oxygen. In this sense, the agglomerates of live and/or dead particulate organic material are potential nitrogen-fixation compartments and indeed, the colonization of a particle can be made from a previous free-living state. Thus, the diazotrophs detected in Chapter 1 might be potential particle colonizers. These particles range in size from less than one micrometre to several centimetres, serving as vehicles for nitrogen and carbon exportation from the sunlit upper waters to the sea floor (Alldredge *et al.*, 1986). Moreover, particles are nutrient-rich environments (compared with their surrounding waters) and, consequently, bacteria frequently colonize them, generating 'hot spots' for microbial activity (Simon *et al.*, 2002). Indeed, the high microbial respiration can lead to low oxygen concentrations or even ephemeral anaerobic conditions in the interior of larger particles (Ploug *et al.*, 1997), a suitable environment for nitrogen fixation. It has been suggested that the presence of cluster III *nifH* genes (associated with anaerobic diazotrophs)(Zehr *et al.*, 2003) in the oxygenated water column could be explained by the presence of anaerobic microsites, within particles for example, where nitrogen fixation could be performed (Farnelid *et al.*, 2011; Benavides *et al.*, 2015). With the aim of demonstrating that the diazotrophs detected in the free-living fraction could transitively be associated or attached to particles, we analysed the genome content of a large *nifH*-gene containing contig, specifically it was a 156 Kb contig containing the OM-RGC.v1.007667460 *nifH* gene assigned to gammaproteobateria (Cluster I) (data not shown). Among the 148 predicted genes, we found, for example, all the genes needed to form the Type IV pilus system involved in the attachment to host cells or surfaces (Craig *et al.*, 2004). Interestingly, we also found genes involved in chemotaxis nearby flagellar protein-coding genes, suggesting that this diazotroph can actively 'swim' toward a particle or toward another organism and establish a physical linkage in its surface via the Type IV pilus system. Cyanobacterial diazotrophs have also been visualized on large particles (Bonnet *et al.*, 2009) and, even a particle-attached lifestyle has been suggested for UCYN-A (Le Moal and Biegala, 2009; Benavides *et al.*, 2011). However, this particle-attached state of UCYN-A was observed using a CARD-FISH probes that hybridize also with other unicellular diazotrophs (Nitro821 probe) (Le Moal and Biegala, 2009) and, it is likely that they were observing *Crocosphaera* attached to particles instead of UCYN-A.

In Chapters 2 and 3 we did observe that different UCYN-A clades occupy overlapping but different size fractions, in accordance to the cell size range of their specific hosts. Thus, our results unveil new UCYN-A lineage-specific nitrogen fixation compartments. Furthermore, we demonstrated that these lineages were active, as shown in Chapter 2 for UCYN-A1 and UCYN-A2 lineages (Cornejo-Castillo *et al.*, 2016) and here (in Fig. 1) for UCYN-A3 lineage. The increasing number of UCYN-A lineages being found, together with their partner fidelity, suggests that we may be soon discovering new lineages occupying new compartments. For example, active UCYN-A *nifH* gene transcripts have been also observed in the gut of copepods, suggesting that new nitrogen can be directly transferred to higher trophic levels (Scavotto *et al.*, 2015). We performed a search in NCBI (nr/nt database) to explore whether 16S rRNA gene of UCYN-A had been found in previous studies associated with other organisms. Surprisingly, UCYN-A 16S rRNA gene sequences, specifically UCYN-A2 sequences, had been amplified from dinoflagellates (*Histioneis* spp.) isolated from the Pacific Ocean (Foster *et al.*, 2006). However, this finding was unnoticed because the 16S rRNA gene sequence of UCYN-A was known just a few years ago (Zehr *et al.*, 2008) and the authors assumed that the 16S rRNA gene sequence belong to *Cyanothece* spp. (Foster *et al.*, 2006). These findings suggest that UCYN-A can occupy a variety of planktonic compartments where nitrogen fixation has not been considered. A deeper analysis of the genome content of these diazotrophs together with the application of visualization tools would help to uncover new nitrogen fixation compartments and to understand the lifestyle of marine diazotrophs.

The curious case of UCYN-A: towards a 'nitrogen-fixing' organelle?

How single cells work together is one of the key questions in evolutionary biology (Zehr, 2015), yet the difficulty to detect single-celled symbioses in nature makes the advance in this field extremely challenging. Symbiotic interactions are key drivers of ecological diversification and evolutionary innovation on Earth (Margulis and Fester, 1991; Moran, 2007; Guerrero *et al.*, 2013; López-García *et al.*, 2017). For example, one of the major innovations in nature was postulated in the Theory of Endosymbiosis by Linn Margulis, which posits that plastids and mitochondria in eukaryotes originated from bacterial endosymbionts giving rise to the eukaryotic photoautotrophic lineages (Margulis, 1971a, 1971b). Primary cyanobacterial-like plastids have experienced modifications in their genomes during the evolution towards a plastid lifestyle, but the underlying processes remain difficult to address because they occur over geological time scales. Thus, the study of present-day symbiotic associations between unicellular eukaryotes and prokaryotes might be an alternative way to understand how symbioses are established.

Of the few currently known single-celled associations, those involving diazotrophic microorganisms are essential in marine biogeochemical cycles. Given that the nitrogen fixation is performed exclusively by prokaryotes together with the importance of nitrogen as a limiting nutrient in the world's oceans productivity (Karl *et al.*, 2002), eukaryotes harbouring nitrogen-fixing symbionts might succeed in nitrogen-depleted environments such as the photic ocean. Nevertheless our knowledge of the distribution of such symbiotic relationships across the global ocean is very limited, and thus we do not know which processes or circumstances promote these types of relationships.

As in primary cyanobacterial-like plastids, the UCYN-A genome underwent a strong genome reduction and, additionally, lacks metabolic pathways typical for cyanobacteria, including the tricarboxylic acid cycle or the oxygenic photosystem II (Zehr *et al.*, 2008; Tripp *et al.*, 2010). Interestingly, the UCYN-A symbiosis can be traced back to the late Cretaceous (91 Mya) (Cornejo-Castillo *et al.*, 2016), which is significantly more recent than the origin of mitochondria or plastids. The presence of a double membrane together with the migration of genes from the endosymbiont genome to the host genome has been used as defining features to distinguish between true plastids and endosymbionts (Theissen and Martin, 2006; Nakayama and Archibald, 2012). Therefore, it will be important to sequence the genome of the UCYNA- host and to determine whether UCYN-A is surrounded by the host membrane (and thus a true endosymbiont) or is attached to the external surface (Zehr *et al.*, 2016) (Fig. 3). In conclusion, the parallelism between the endosymbiosis that originated the plastids and the UCYN-A symbiosis makes this system a unique model to gain insight into the evolution of plastids, and further poses the question of whether we are currently witnessing an evolutionary process that will eventually lead to the establishment of a nitrogen-fixing organelle.

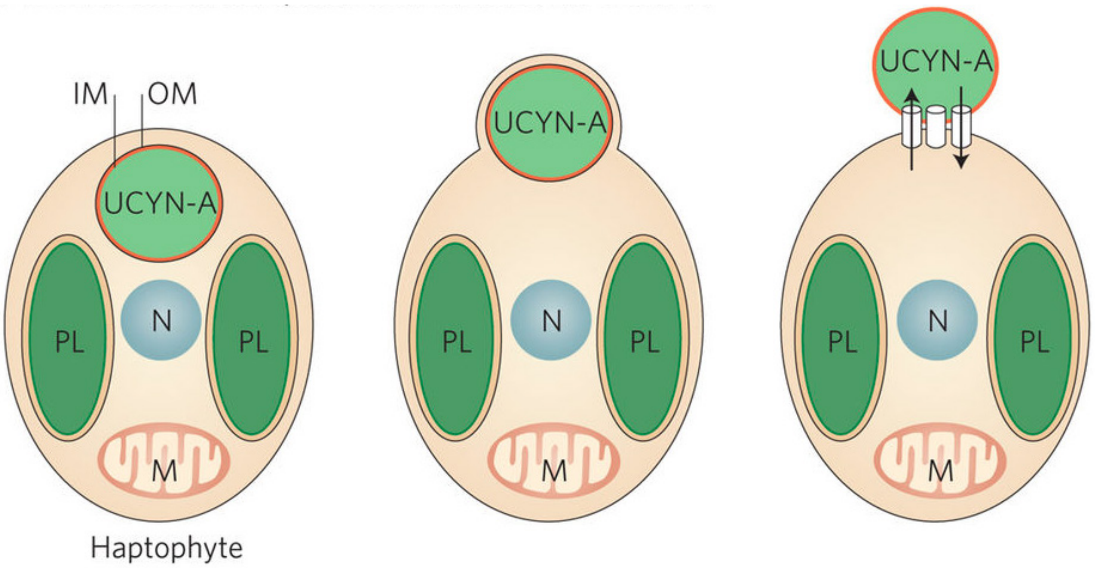
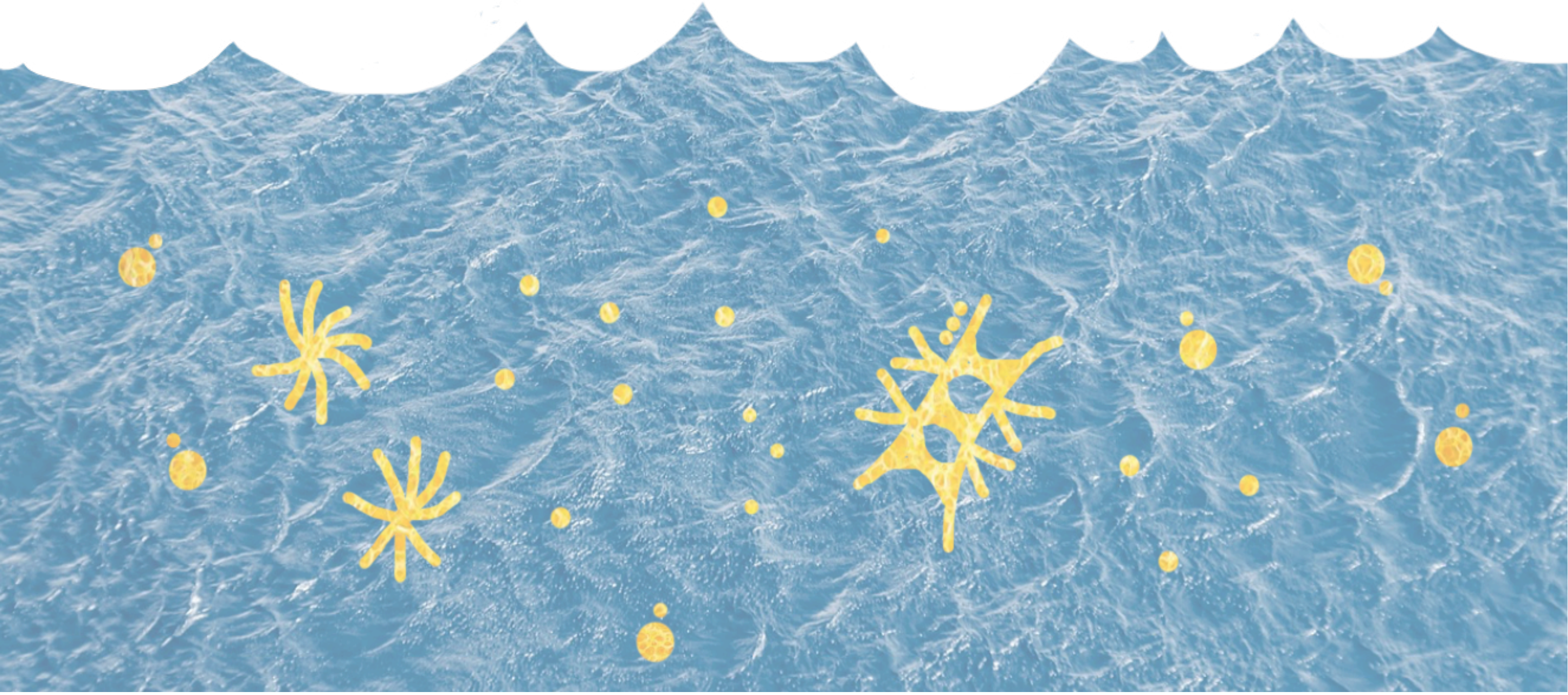


Figure 3. Possible models of symbiotic interactions between UCYN-A and its haptophyte host. N, host nucleus; PL, plastids; M, mitochondrion; OM, outer membrane; IM, inner membrane. (From Zehr *et al*., 2016).

Future directions for the study of the UCYN-A symbiosis

The application of novel molecular approaches has accelerated the advent of microbial symbiosis research (McFall-Ngai, 2008). This is the case of the UCYN-A symbiosis, discovered through culture-independent methods. The combination of flow cytometric–sorting and high-throughput sequencing unveiled an streamlined genome of UCYN-A, suggesting a symbiotic lifestyle due to the lack of biosynthetic pathways for essential nutrients such as fixed carbon or essential amino acids (Tripp *et al*., 2010). Similarly, nano-scale secondary ion mass spectrometry (nanoSIMS) was key to elucidate the basis of the symbiosis in terms of metabolic exchange. This revealed that in the UCYN-A1 and UCYN-A2 symbiosis the cyanobacteria offers fixed nitrogen to its host and, in return, receives fixed carbon from the alga (Thompson *et al*., 2012; Martínez-Pérez *et al*., 2016). However, of the four different lineages of UCYN-A (Thompson *et al*., 2014; Farnelid *et al*., 2016) identified based on nitrogenase gene (*nif*H) phylogenetic trees, at least two differ in the number of cells per algal host and in the host cell size, suggesting that nutrient requirements and exchange patterns may differ between lineages, yet this is still unknown. Investigating the nutrient exchange of new UCYN-A symbioses is thus key to understand the interplay between both host-symbiont and symbiont-symbiont in those cases with a cell ratio different than 1:1. Moreover, since the algal host and UCYN-A are photosynthetic and photoheterotrophic organisms, respectively, these exchanges will likely vary over the diel cycle. In particular, nitrogen fixation needs an oxygen-isolated environment to happen because the main enzyme of this process, the nitrogenase, is inactivated by oxygen, which may explain why UCYN-A expresses the nitrogenase genes during the day (given that it does not evolve O_2) (Zehr, 2011). However, the association with an oxygen-evolving partner could make the nitrogenase enzyme in UCYN-A not completely safe from oxygen, opening new unknowns on the nitrogen fixation process. Both carbon and nitrogen fixation processes should be highly synchronized since the transfer of carbon from the alga to UCYN-A requires the previous transfer of nitrogen from UCYN-A (Krupke *et al.*, 2015), but so far no study has explored the diel variations in the different processes resulting from these symbiotic interactions. In addition, although this is an obligate symbiosis (Cabello *et al.*, 2015), the mechanisms driving the host-symbiont cell division, particularly in those cases with a host-symbiont cell ratio different than 1:1, are yet to be resolved. All this suggests a very tight and strongly regulated coupling between partners to succeed as a symbiotic entity, but the processes behind are still poorly understood. The future research on UCYN-A should go towards gaining knowledge of the ecology and the evolution of the UCYN-A nitrogen-fixing symbiosis to, eventually, get a better understanding of the mechanisms conducting to the plastid formation. Single-cell sorting, quantitative isotopic techniques, epifluorescence microscopy and high throughput sequencing of DNA and RNA performed on natural and experimentally-derived samples will be crucial to investigate how UCYN-A interacts with its hosting-cell at the genetic and biochemical level, and to understand the mechanisms and the environmental drivers that promote these types of symbiotic interactions in nature.

Conclusions


"

CONCLUSIONS

1. Our metagenomic-based approach of the *Tara* Oceans dataset has increased the current repertoire of known *nifH* gene sequences unveiling that the nifH primers used in most diversity studies of marine diazotrophic diversity have led to an inaccurate view of their diversity and abundance patterns. This biased view is especially noticeable in heterotrophic diazotrophs, particularly in gammaproteobacterial (Cluster I) and deltaproteobacterial (Cluster III) diazotrophs, and less important in cyanobacterial diazotrophs.

2. Diazotrophs were widely distributed in the global ocean, but were commonly found within the rare biosphere community (<0.1% relative abundance) with some punctual exceptions, showing significant higher relative abundances in mesopelagic (0.07% of the prokaryotic community) than in surface (0.04%) or DCM waters (0.02%). Moreover, some diazotrophic clusters showed contrasting habitat preferences, i.e., photic waters for cyanobacterial diazotrophs and mesopelagic waters for gamma- and alphaproteobacterial diazotrophs.

3. The *nifH* gene sequences were mainly associated with gammaproteobacterial (63% of *nifH* sequences), cyanobacterial (16%) and delta-proteobacterial (12%) diazotrophs. The most abundant individual diazotroph detected in the dataset was the unicellular symbiotic cyanobacterium *C. Atelocyanobacterium thalassa* (UCYN-A), which accounted for up to 0.6 % of the prokaryotic community in surface waters of the South Atlantic Ocean.

4. The combined analysis of co-occurring UCYN-A lineages in the South Atlantic Ocean demonstrated that UCYN-A display partner fidelity with their prymnesiophyte hosts, i.e., *B. bigelowii* in the case of UCYN-A2 and a closely-related prymnesiophyte in the case of UCYN-A1. We showed that the two UCYN-A lineages displayed different symbiotic organizations: whereas the UCYN-A1 lineage has one or two separate cells per host, the UCYN-A2 lineage may harbor up to 10 cells per host within a common symbiotic structure.

5. Although the UCYN-A1 lineage was in general more abundant than UCYN-A2, an increase in abundance of the UCYN-A2 lineage was observed from smaller (0.2-3 μm) to larger (5-20 μm) size fractions, likely explained by the difference in cell size of their prymnesiophyte partners.

6. UCYN-A lineages dedicate a large transcriptional investment to fix nitrogen coupled to the generation of reducing power and the ATP synthesis. In both UCYN-A1 and UCYN-A2 lineages, the nitrogen fixation operon, including the *nifH* gene, was the most highly expressed gene-cluster accounting for a quarter of the total transcripts. Despite UCYN-A1 being more abundant than UCYN-A2, the expressed *nifH* transcripts per cell were almost 2 times higher for UCYN-A2 (648.33) than for UCYN-A1 (396.60), which may reflect differential nutrient requirements for growth of their specific partners.

7. We did not detect signs of large-scale positive selection, i.e. no apparent strong adaptation to novel niches in UCYN-A lineages, suggesting that the evolutionary forces for niche adaptation would act on the prymnesiophyte partners rather than on UCYN-A, and that the symbionts were genetically adapted to their hosts before they were separated by speciation. Our results indicate that UCYN-A1 and UCYN-A2 lineages diverged around 91 million years ago

(Mya), i.e. during the late Cretaceous, after the low nutrient regime period occurred during the Jurassic.

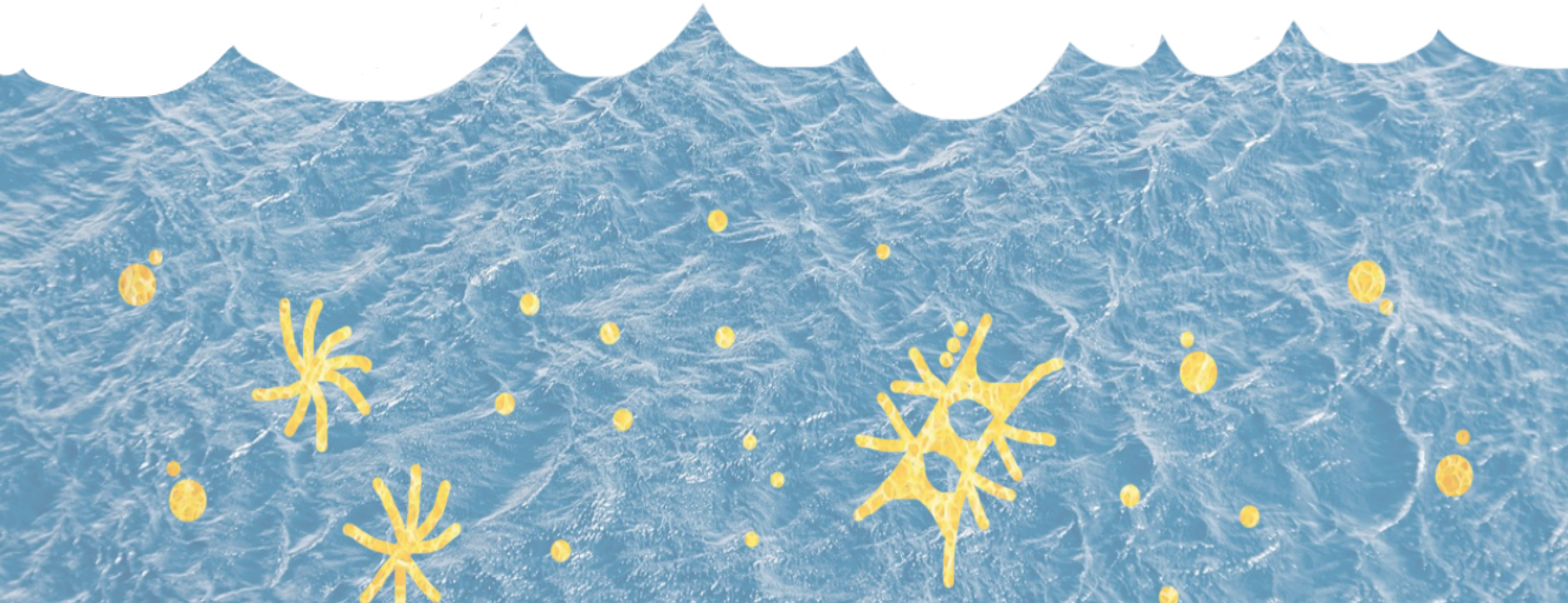
8. Different methods (CARD-FISH counts, together with the UCYN-A *nifH* gene sequences and qPCR ratios) showed some discrepancies in the identification of each UCYN-A lineage and led us to conclude that the UCYN-A association originally assumed to be UCYN-A2 at Station ALOHA is actually UCYN-A3.

9. The partial reconstruction of the genome of UCYN-A3 revealed this lineage as a new UCYN-A genomic species. Microscopic visualization showed that the size of UCYN-A3 cells and of their host, as well as the number of cyanobacterial cells per host are different from that of the better characterized lineages (UCYN-A1 and UCYN-A2), thus occupying a new planktonic compartment.

10. The existence of multiple UCYN-A sublineages and prymensiophyte hosts that vary in size has implications for N_2 fixation rates, since they could occupy size-fractions that are not considered in diazotroph diversity studies, consequently, UCYN-A may play a larger role in N_2 fixation than previously thought.

References

(General Introduction)



REFERENCES (GENERAL INTRODUCTION)

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF. (2005). PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Appl Environ Microbiol* **71**(12):8966-8969.

Arrigo KR. (2005). Marine microorganisms and global nutrient cycles. *Nature* **437**:349–355.

Benavides M, Moisander PH, Berthelot H, Dittmar T, Grosso O, Bonnet S. (2015). Mesopelagic N2 Fixation Related to Organic Matter Composition in the Solomon and Bismarck Seas (Southwest Pacific). *PLoS One* **10**:e0143775.

Bench SR, Frank I, Robidart J, Zehr JP. (2016). Two subpopulations of Crocosphaera watsonii have distinct distributions in the North and South Pacific. *Environ Microbiol* **18**: 514-524.

Bench SR, Heller P, Frank I, Arciniega M, Shilova IN, Zehr JP. (2013). Whole genome comparison of six Crocosphaera watsonii strains with differing phenotypes. *J Phycol* **49**:786–801.

Bench SR, Ilikchyan IN, James Tripp H, Zehr JP. (2011). Two strains of crocosphaera watsonii with highly conserved genomes are distinguished by strain-specific features. *Front Microbiol* **2**. doi:10.3389/fmicb.2011.00261.

Bergman B, Sandh G, Lin S, Larsson J, Carpenter EJ. (2013). Trichodesmium - a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol Rev* **37**:286–302.

Berman-Frank I, Lundgren P, Falkowski P. (2003). Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Res Microbiol* **154**:157–164.

Biegala IC, Raimbault P. (2008). High abundance of diazotrophic picocyanobacteria (<3 μm) in a Southwest Pacific coral lagoon. *Aquat Microb Ecol* **51**:45–53.

Bird C, Martinez JM, O'Donnell AG, Wyman M. (2005). Spatial distribution and transcriptional activity of an uncultured clade of planktonic diazotrophic γ-proteobacteria in the Arabian Sea. *Appl Environ Microbiol* **71**:2079–2085.

Blais M, Tremblay J-E, Jungblut AD, Gagnon J, Martin J, Thaler M, *et al.* (2012). Nitrogen fixation and identification of potential diazotrophs in the Canadian Arctic. *Global Biogeochem Cycles* **26**. doi:10.1029/2011GB004096.

Bombar D, Heller P, Sanchez-Baracaldo P, Carter BJ, Zehr JP. (2014). Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *ISME J*. http://dx.doi.org/10.1038/ismej.2014.167.

Bombar D, Paerl RW, Riemann L. (2017). Marine Non-Cyanobacterial Diazotrophs: Moving beyond Molecular Detection. *Trends Microbiol* **24**:916–927.

Bonnet S, Biegala IC, Dutrieux P, Slemons LO, Capone DG. (2009). Nitrogen fixation in the western equatorial Pacific: Rates, diazotrophic cyanobacterial size class distribution, and biogeochemical significance. *Global Biogeochem Cycles* **23**. doi:10.1029/2008GB003439.

Bonnet S, Dekaezemacker J, Turk-Kubo KA, Moutin T, Hamersley RM, Grosso O, *et al.* (2013). Aphotic N2 fixation in the eastern tropical South Pacific Ocean. *PLoS One* **8**. doi:10.1371/journal.pone.0081265.

Bürgmann H, Widmer F, Von Sigler W, Zeyer J. (2004). New Molecular Screening Tools for Analysis of Free-Living Diazotrophs in Soil. *Appl Environ Microbiol* **70**:240–247.

Capone DG, Montoya JP. (2001). Nitrogen fixation and denitrification. *Methods Microbiol* **30**:501–515.

Capone G, Bronk DA, Mulholland MR, Carpenter EJ. (2008). Nitrogen in the Marine Environment. doi:10.1016/B978-0-12-372522-6.X0001-1.

Carpenter EJ. (2002). Marine cyanobacterial symbioses. *Biol Environ* **102**:15–18.

Carpenter EJ. (1983). Physiology and ecology of marine Oscillatoria (Trichodesmium).

Carpenter EJ, Capone DG. (2008). Nitrogen Fixation in the Marine Environment. doi:10.1016/B978-0-12-372522-6.00004-9.

Carpenter EJ, Janson S. (2000). Intracellular cyanobacterial symbionts in the marine diatom Climacodium frauenfeldianum (Bacillariophyceae). *J Phycol* **36**:540–544.

Carpenter EJ, Price CC. (1976). Marine oscillatoria (Trichodesmium): explanation for aerobic nitrogen fixation without heterocysts. *Science (80-)* **191**:1278–1280.

Church MJ, Jenkins BD, Karl DM, Zehr JP. (2005). Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat Microb Ecol* **38**:3–14.

Church MJ, Short CM, Jenkins BD, Karl DM, Zehr JP. (2005). Temporal patterns of nitrogenase gene (nifH) expression in the oligotrophic North Pacific Ocean. *Appl Environ Microbiol* **71**:5362–5370.

Díez B, Bergman B, Pedrós-Alió C, Antó M, Snoeijs P. (2012). High cyanobacterial nifH gene diversity in Arctic seawater and sea ice brine. *Environ Microbiol Rep* **4**:360–366.

Dore JE, Letelier RM, Church MJ, Lukas R, Karl DM. (2008). Summer phytoplankton blooms in the oligotrophic North Pacific Subtropical Gyre: Historical perspective and recent observations. *Prog Oceanogr* **76**:2–38.

Dugdale RC, Menzel DW, Ryther JH. (1961). Nitrogen fixation in the Sargasso Sea. *Deep Sea Res* **7**:297–300.

Falcon LI, Cipriano F, Chistoserdov AY, Carpenter EJ. (2002). Diversity of diazotrophic unicellular cyanobacteria in the tropical North Atlantic Ocean. *Appl Environ Microbiol* **68**:5760–5764.

Farnelid H, Andersson AF, Bertilsson S, Al-Soud WA, Hansen LH, Sørensen S, *et al.* (2011). Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* **6**:e19223.

Farnelid H, Bentzon-Tilia M, Andersson AF, Bertilsson S, Jost G, Labrenz M, *et al.* (2013). Active nitrogen-fixing heterotrophic bacteria at and below the chemocline of the central Baltic Sea. *ISME J* **7**:1413–1423.

Fay P. (1992). Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol Rev* **56**:340–373.

Fernandez C, Farías L, Ulloa O. (2011). Nitrogen fixation in denitrified marine waters. *PLoS One* **6**. doi:10.1371/journal.pone.0020539.

Fogg GE. (1978). Nitrogen Fixation in the Oceans. *Ecol Bull* 11–19.

Fogg GE. (1942). Studies on Nitrogen Fixation by Blue-Green Algae. *J Exp Biol* **19**:78–87.

Fong AA, Karl DM, Lukas R, Letelier RM, Zehr JP, Church MJ. (2008). Nitrogen fixation in an anticyclonic eddy in the oligotrophic North Pac. *ISME J* **2**:663–676.

Foster RA, Kuypers MMM, Vagner T, Paerl RW, Musat N, Zehr JP. (2011). Nitrogen fixation and transfer in open ocean diatom-cyanobacterial symbioses. *ISME J* **5**:1484–1493.

Foster RA, Subramaniam A, Mahaffey C, Carpenter EJ, Capone DG, Zehr JP. (2007). Influence of the Amazon River plume on distributions of free-living and symbiotic cyanobacteria in the western tropical north Atlantic Ocean. *Limnol Oceanogr* **52**:517–532.

Foster RA, Subramaniam A, Zehr JP. (2009). Distribution and activity of diazotrophs in the Eastern Equatorial Atlantic. *Environ Microbiol* **11**:741–750.

Foster RA, Sztejrenszus S, Kuypers MMM. (2013). Measuring carbon and N2 fixation in field populations of colonial and free-living unicellular cyanobacteria using nanometer-scale secondary ion mass spectrometry. *J Phycol* **49**:502–516.

Giovannoni SJ, Stingl U. (2005). Molecular diversity and ecology of microbial plankton. *Nature* **437**:343–348.

Gradoville MR, Bombar D, Crump BC, Letelier RM, Zehr JP, White AE. (2017). Diversity and activity of nitrogen-fixing communities across ocean basins. *Limnol Oceanogr* **62**:1895–1909.

Grokopf T, Mohr W, Baustian T, Schunck H, Gill D, Kuypers MMM, *et al.* (2012). Doubling of marine dinitrogen-fixation rates based on direct measurements. *Nature* **488**:361–364.

Gruber N, Sarmiento JL. (1997). Global patterns of marine nitrogen fixation and denitrification. *Global Biogeochem Cycles* **11**:235–266.

Hagino K, Onuma R, Kawachi M, Horiguchi T. (2013). Discovery of an endosymbiotic nitrogen-fixing cyanobacterium UCYN-A in Braarudosphaera bigelowii (Prymnesiophyceae). *PLoS One* **8**:e81749.

Halm H, Lam P, Ferdelman TG, Lavik G, Dittmar T, Laroche J, *et al.* (2011). Heterotrophic organisms dominate nitrogen fixation in the South Pacific Gyre. *ISME J* 1238–1249.

Hardy RWF, Holsten RD, Jackson EK, Burns RC. (1968). The Acetylene-Ethylene Assay for N2 Fixation: Laboratory and Field Evaluation. *Plant Physiol* **43**:1185–1207.

Heller P, Tripp HJ, Turk-Kubo K, Zehr JP. (2014). ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank . *Bioinformatics* **30**:2883–2890.

Horner-Devine MC, Martiny AC. (2008). Biogeochemistry: News about nitrogen. *Science (80-)* **320**:757–758.

Howurth RW, Butler T, Lunde K, Swaney D, Chu CR. (1993). Turbulence and planktonic nitrogen fixation: A mesocosm experiment. *Limnol Oceanogr* **38**:1696–1711.

Hynes AM, Webb EA, Doney SC, Waterbury JB. (2012). Comparison of cultured trichodesmium (cyanophyceae) with species characterized from the field. *J Phycol* **48**:196–210.

Karl D, Michaels A, Bergman B, Capone D, Carpenter E, Letelier R, *et al.* (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58(1)**:47–98.

Krupke A, Musat N, Laroche J, Mohr W, Fuchs BM, Amann RI, *et al.* (2013). In situ identification and N_2 and C fixation rates of uncultivated cyanobacteria populations. *Syst Appl Microbiol* **36**:259–71.

Langlois R, Großkopf T, Mills M, Takeda S, LaRoche J. (2015). Widespread Distribution and Expression of Gamma A (UMB), an Uncultured, Diazotrophic, γ-Proteobacterial nifH Phylotype. *PLoS One* **10**:e0128912.

Langlois RJ, Hümmer D, LaRoche J. (2008). Abundances and distributions of the dominant nifH phylotypes in the Northern Atlantic Ocean. *Appl Environ Microbiol* **74**:1922–1931.

Loescher CR, Groskopf T, Desai FD, Gill D, Schunck H, Croot PL, *et al.* (2014). Facets of diazotrophy in the oxygen minimum zone waters off Peru. *ISME J* **8**:2180–2192.

Martínez-Pérez C, Mohr W, Löscher CR, Dekaezemacker J, Littmann S, Yilmaz P, *et al.* (2016). The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol* **1**. doi:10.1038/nmicrobiol.2016.163.

Le Moal M, Biegala IC. (2009). Diazotrophic unicellular cyanobacteria in the northwestern Mediterranean Sea: A seasonal cycle. *Limnol Oceanogr* **54**:845–855.

Mohr W, Großkopf T, Wallace DWR, LaRoche J. (2010). Methodological underestimation of oceanic nitrogen fixation rates. *PLoS One* **5**:1–7.

Moisander PH, Serros T, Paerl RW, Beinart RA, Zehr JP. (2014). Gammaproteobacterial diazotrophs and nifH gene expression in surface waters of the South Pacific Ocean. *ISME J* **8**:1962–1973.

Moisander PH, Zhang R, Boyle EA, Hewson I, Montoya JP, Zehr JP. (2012). Analogous nutrient limitations in unicellular diazotrophs and Prochlorococcus in the South Pacific Ocean. *ISME J* **6**:733–744.

Monteiro FM, Follows MJ, Dutkiewicz S. (2010). Distribution of diverse nitrogen fixers in the global ocean. *Global Biogeochem Cycles* **24**. doi:10.1029/2009GB003731.

Montoya JP, Voss M, Kähler P, Capone DG. (1996). A simple, high-precision, high-sensitivity tracer assay for N2 fixation. *Appl Environ Microbiol* **62**:986–993.

Neveux J, Lantoine F, Vaulot D, Marie D, Blanchot J. (1999). Phycoerythrins in the southern tropical and equatorial Pacific Ocean: Evidence for new cyanobacterial types. *J Geophys Res Ocean* **104**:3311–3321.

Orcutt KM, Rasmussen U, Webb EA, Waterbury JB, Gundersen K, Bergman B. (2002). Characterization of Trichodesmium spp. by genetic techniques. *Appl Environ Microbiol* **68**:2236–2245.

Rahav E, Bar-Zeev E, Ohayon S, Elifantz H, Belkin N, Herut B, *et al.* (2013). Dinitrogen fixation in aphotic oxygenated marine environments. *Front Microbiol* **4**. doi:10.3389/fmicb.2013.00227.

Raymond J, Siefert JL, Staples CR, Blankenship RE. (2004). The Natural History of Nitrogen Fixation. *Mol Biol Evol* **21**:541–554.

Scharek R, Tupas LM, Karl DM. (1999). Diatom fluxes to the deep sea in the oligotrophic North Pacific gyre at Station ALOHA. *Mar Ecol Prog Ser* **182**:55–67.

Sherman, L.A., Meunier, P., and Colón-López, M.S. (1998). Diurnal rhythms in metabolism: a day in the life of a unicellular, diazotrophic cyanobacterium. *Photosynth Res* **58**: 25–42.

Shiozaki T, Bombar D, Riemann L, Hashihama F, Takeda S, Yamaguchi T, *et al.* (2017). Basin scale variability of active diazotrophs and nitrogen fixation in the North Pacific, from the tropics to the subarctic Bering Sea. *Global Biogeochem Cycles* **31**:996–1009.

Shiozaki T, Ijichi M, Kodama T, Takeda S, Furuya K. (2014). Heterotrophic bacteria as major nitrogen fixers in the euphotic zone of the Indian Ocean. *Global Biogeochem Cycles* **28**:1096–1110.

Simpson FB, Burris RH. (1984). A nitrogen pressure of 50 atmospheres does not prevent evolution of hydrogen by nitrogenase. *Science (80-)* **224**:1095 LP-1097.

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K, Nikolausz M. (2007). Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60**:341–350.

Sohm JA, Edwards BR, Wilson BG, Webb EA. (2011). Constitutive extracellular polysaccharide (EPS) production by specific isolates of Crocosphaera watsonii. *Front Microbiol* **2**. doi:10.3389/fmicb.2011.00229.

Sohm JA, Webb EA, Capone DG. (2011). Emerging patterns of marine nitrogen fixation. *Nat Rev Microbiol* **9**:499–508.

Staal M, Hekkert STL, Brummer GJ, Veldhuis M, Sikkens C, Persijn S, *et al.* (2007). Nitrogen fixation along a north-south transect in the eastern Atlantic Ocean. *Limnol Oceanogr* **52**:1305–1316.

Stal LJ. (2009). Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature?: Minireview. *Environ Microbiol* **11**:1632–1645.

Staples CR, Lahiri S, Raymond J, Von Herbulis L, Mukhophadhyay B, Blankenship RE. (2007). Expression and association of group IV nitrogenase NifD and NifH homologs in the non-nitrogen-fixing archaeon Methanocaldococcus jannaschii. *J Bacteriol* **189**:7392–7398.

Stein LY, Klotz MG. (2017). The nitrogen cycle. *Curr Biol* **26**:R94–R98.

Stewart WDP. (1965). Nitrogen Turnover in Marine and Brackish HabitatsI. Nitrogen Fixation. *Ann Bot* **29**:229–239.

Subramaniam A, Yager PL, Carpenter EJ, Mahaffey C, Björkman K, Cooley S, *et al.* (2008). Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci U S A* **105**:10460–10465.

Suzuki MT, Giovannoni SJ. (1996). Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* **62**:625–630.

Taniuchi Y, Chen Y-LL, Chen H-Y, Tsai M-L, Ohki K. (2012). Isolation and characterization of the unicellular diazotrophic cyanobacterium Group C TW3 from the tropical western Pacific Ocean. *Environ Microbiol* **14**:641–654.

Thompson A, Carter BJ, Turk-Kubo K, Malfatti F, Azam F, Zehr JP. (2014). Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol* n/a-n/a.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, *et al.* (2012). Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science (80-)* **337**:1546–1550.

Thompson AW, Zehr JP. (2013). Cellular interactions: Lessons from the nitrogen-fixing cyanobacteria. *J Phycol* **49**:1024–1035.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**:90–94.

Turk-Kubo KA, Karamchandani M, Capone DG, Zehr JP. (2014). 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.

Villareal TA. (1992). Marine Nitrogen-Fixing Diatom-Cyanobacteria Symbioses. In:*Marine Pelagic Cyanobacteria: Trichodesmium and other Diazotrophs*, Carpenter, EJ, Capone, DG, & Rueter, JG (eds), Springer Netherlands: Dordrecht, pp. 163–175.

Villareal TA, Carpenter EJ. (2003). Buoyancy Regulation and the Potential for Vertical Migration in the Oceanic Cyanobacterium Trichodesmium. *Microb Ecol* **45**:1–10.

Waksman SA, Hotchkiss M, Carey CL. (1933). Marine bacteria and their role in the cycle of life in the sea: II. bacteria concerned with the cycle of nitrogen in the sea. *Biol Bull* **65**:137–167.

Webb EA, Ehrenreich IM, Brown SL, Valois FW, Waterbury JB. (2009). Phenotypic and genotypic characterization of multiple strains of the diazotrophic cyanobacterium, Crocosphaera watsonii, isolated from the open ocean. *Environ Microbiol* **11**:338–348.

Young JPW. (2005). The Phylogeny and Evolution of Nitrogenases BT - Genomes and Genomics of Nitrogen-fixing Organisms. In: Palacios, R & Newton, WE (eds), Springer Netherlands: Dordrecht, pp. 221–241.

Zehr JP. (2011). Nitrogen fixation by marine cyanobacteria. *Trends Microbiol* **19**:162–173.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, *et al.* (2008). Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**:1110–2.

Zehr JP, Bench SR, Mondragon EA, Mccarren J, Delong EF. (2007). Low genomic diversity in tropical oceanic N 2 -fixing cyanobacteria. *Sci York* **104**.

Zehr JP, Jenkins BD, Short SM, Steward GF. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ Microbiol* **5**:539–554.

Zehr JP, Kudela RM. (2011). Nitrogen Cycle of the Open Ocean: From Genes to Ecosystems. *Ann Rev Mar Sci* **3**:197–225.

Zehr JP, McReynolds LA. (1989). Use of degenerate oligonucleotides for amplification of the nifH gene from the marine cyanobacterium Trichodesmium thiebautii. *Appl Environ Microbiol* **55**:2522–2526.

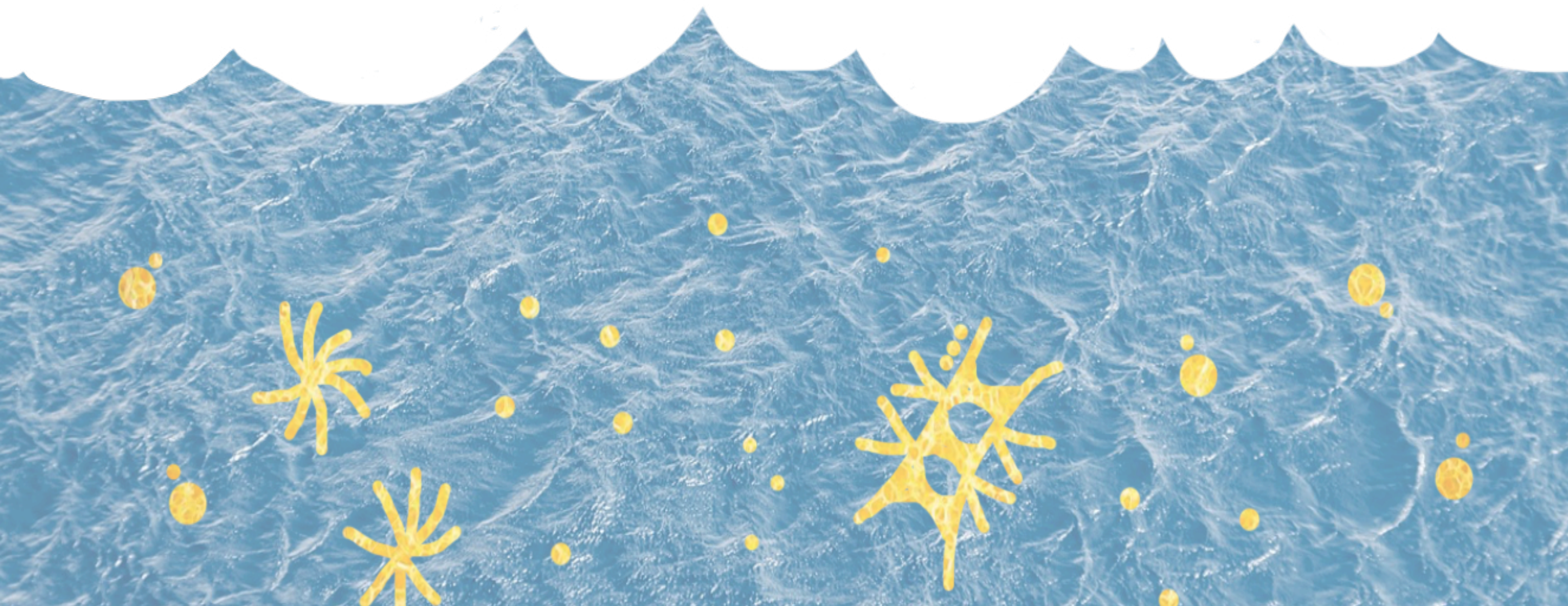
Zehr JP, Mellon MT, Zani S. (1998). New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (nifH) genes. *Appl Environ Microbiol* **64**:3444–3450.

Zehr JP, Shilova IN, Farnelid HM, Muñoz-Maríncarmen MDC, Turk-Kubo KA. (2016). Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. *Nat Microbiol* **2**. doi:10.1038/nmicrobiol.2016.214.

Zhang Y, Zhao Z, Sun J, Jiao N. (2011). Diversity and distribution of diazotrophic communities in the South China Sea deep basin with mesoscale cyclonic eddy perturbations. *FEMS Microbiol Ecol* **78**:417–427.

References

(General Discussion)



REFERENCES (GENERAL DISCUSSION)

Alldredge AL, Cole JJ, Caron DA. (1986). Production of heterotrophic bacteria inhabiting macroscopic organic aggregates (marine snow) from surface waters. *Limnol Oceanogr* **31**:68–78.

Benavides M, Arístegui J, Agawin NSR, Cancio JL, Hernàndez-León S. (2013). Enhancement of nitrogen fixation rates by unicellular diazotrophs vs. Trichodesmium after a dust deposition event in the Canary Islands. *Limnol Oceanogr* **58**:267–275.

Benavides M, Moisander PH, Berthelot H, Dittmar T, Grosso O, Bonnet S. (2015). Mesopelagic N2 Fixation Related to Organic Matter Composition in the Solomon and Bismarck Seas (Southwest Pacific). *PLoS One* **10**:e0143775.

Benavides M, NSR A, Arístegui J, Ferriol P, Stal L J. (2011). Nitrogen fixation by Trichodesmium and small diazotrophs in the subtropical northeast Atlantic . *Aquat Microb Ecol* **65**:43–53.

Bonnet S, Biegala IC, Dutrieux P, Slemons LO, Capone DG. (2009). Nitrogen fixation in the western equatorial Pacific: Rates, diazotrophic cyanobacterial size class distribution, and biogeochemical significance. *Global Biogeochem Cycles* **23**. doi:10.1029/2008GB003439.

Cabello AM, Cornejo-Castillo FM, Raho N, Blasco D, Vidal M, Audic S, *et al.* (2015). Global distribution and vertical patterns of a prymnesiophyte-cyanobacteria obligate symbiosis. *ISME J.* **10**:693-706.

Church MJ, Mahaffey C, Letelier RM, Lukas R, Zehr JP, Karl DM. (2009). Physical forcing of nitrogen fixation and diazotroph community structure in the North Pacific subtropical gyre. *Global Biogeochem Cycles* **23**:n/a-n/a.

Cornejo-Castillo FM, Cabello AM, Salazar G, Sánchez-Baracaldo P, Lima-Mendez G, Hingamp P, *et al.* (2016). Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat Commun* **7**. doi:10.1038/ncomms11071.

Craig L, Pique ME, Tainer JA. (2004). Type IV pilus structure and bacterial pathogenicity. *Nat Rev Microbiol* **2**:363.

Farnelid H, Andersson AF, Bertilsson S, Al-Soud WA, Hansen LH, Sørensen S, *et al.* (2011). Nitrogenase gene amplicons from global marine surface waters are dominated by genes of non-cyanobacteria. *PLoS One* **6**:e19223.

Farnelid H, Turk-Kubo K, Del Carmen Muñoz-Marín M, Zehr JP. (2016). New insights into the ecology of the globally significant uncultured nitrogen-fixing symbiont UCYN-A. *Aquat Microb Ecol* **77**:128–138.

Fong AA, Karl DM, Lukas R, Letelier RM, Zehr JP, Church MJ. (2008). Nitrogen fixation in an anticyclonic eddy in the oligotrophic North Pacific Ocean. *ISME J* **2**:663–676.

Foster RA, Collier JL, Carpenter EJ. (2006). Reverse transcription PCR amplification of cyanobacterial symbiont 16S rRNA sequences from single non-photosynthetic eukaryotic marine planktonic host cells. *J Phycol* **42**:243–250.

Gaby JC, Buckley DH. (2012). A Comprehensive Evaluation of PCR Primers to Amplify the nifH Gene of Nitrogenase. *PLoS One* **7**:e42149.

Giovannoni SJ, Stingl U. (2005). Molecular diversity and ecology of microbial plankton. *Nature* **437**:343–348.

Guerrero R, Margulis L, Berlanga M. (2013). Symbiogenesis: The holobiont as a unit of evolution. *Int Microbiol* **16**:133–143.

Hazen TC, Dubinsky EA, DeSantis TZ, Andersen GL, Piceno YM, Singh N, *et al.* (2010). Deep-Sea Oil Plume Enriches Indigenous Oil-Degrading Bacteria. *Science (80-)* **330**:204–208.

Karl D, Michaels A, Bergman B, Capone D, Carpenter E, Letelier R, *et al.* (2002). Dinitrogen fixation in the world's oceans. *Biogeochemistry* **57/58(1)**:47–98.

Karl DM, Church MJ, Dore JE, Letelier RM, Mahaffey C. (2012). Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc Natl Acad Sci* **109**:1842–1849.

Krupke A, Lavik G, Halm H, Fuchs BM, Amann RI, Kuypers MMM. (2014). Distribution of a consortium between unicellular algae and the N2 fixing cyanobacterium UCYN-A in the North Atlantic Ocean. *Environ Microbiol*. doi:10.1111/1462-2920.12431.

Krupke A, Mohr W, LaRoche J, Fuchs BM, Amann RI, Kuypers MMM. (2015). The effect of nutrients on carbon and nitrogen fixation by the UCYN-A-haptophyte symbiosis. *ISME J* **9**:1635–1647.

Loescher CR, Groskopf T, Desai FD, Gill D, Schunck H, Croot PL, *et al.* (2014). Facets of diazotrophy in the oxygen minimum zone waters off Peru. *ISME J* **8**:2180–2192.

López-García P, Eme L, Moreira D. (2017). Symbiosis in eukaryotic evolution. *J Theor Biol* **434**:20–33.

Lynch MDJ, Neufeld JD. (2015). Ecology and exploration of the rare biosphere. *Nat Rev Microbiol* **13**:217–229.

Margulis L. (1971a). Symbiosis and evolution. *Sci Am* **225**:48–57.

Margulis L. (1971b). The origin of plant and animal cells. *Am Sci* **59**:230–235.

Margulis L, Fester R. (1991). Bellagio conference and book. Symbiosis as Source of Evolutionary Innovation: Speciation and Morphogenesis. Conference--June 25-30, 1989, Bellagio Conference Center, Italy. *Symbiosis* **11**:93–101.

Martínez-Pérez C, Mohr W, Löscher CR, Dekaezemacker J, Littmann S, Yilmaz P, *et al.* (2016). The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat Microbiol* **1**. doi:10.1038/nmicrobiol.2016.163.

McFall-Ngai M. (2008). Are biologists in 'future shock'? Symbiosis integrates biology across domains. *Nat Rev Microbiol* **6**:789–792.

Le Moal M, Biegala IC. (2009). Diazotrophic unicellular cyanobacteria in the northwestern Mediterranean Sea: A seasonal cycle. *Limnol Oceanogr* **54**:845–855.

Moran NA. (2007). Symbiosis as an adaptive process and source of phenotypic complexity. *Proc Natl Acad Sci U S A* **104**:8627–8633.

Pedrós-Alió C. (2012). The Rare Bacterial Biosphere. *Ann Rev Mar Sci* **4**:449–466.

Ploug H, Kühl M, Buchholz-Cleven B, Jørgensen BB. (1997). Anoxic aggregates - An ephemeral phenomenon in the pelagic environment? *Aquat Microb Ecol* **13**:285–294.

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

Simon M, Grossart H-P, Schweitzer B, Ploug H. (2002). Microbial ecology of organic aggregates in aquatic ecosystems. *Aquat Microb Ecol* **28**:175–211.

Sohm JA, Webb EA, Capone DG. (2011). Emerging patterns of marine nitrogen fixation. *Nat Rev Microbiol* **9**:499–508.

Stal LJ. (2009). Is the distribution of nitrogen-fixing cyanobacteria in the oceans related to temperature?: Minireview. *Environ Microbiol* **11**:1632–1645.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, *et al.* (2015). Structure and function of the global ocean microbiome. *Science (80-)* **348**. doi:10.1126/science.1261359.

Thompson A, Carter BJ, Turk-Kubo K, Malfatti F, Azam F, Zehr JP. (2014). Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol* n/a-n/a.

Thompson AW, Foster RA, Krupke A, Carter BJ, Musat N, Vaulot D, *et al.* (2012). Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science (80-)* **337**:1546–1550.

Thompson AW, Zehr JP. (2013). Cellular interactions: Lessons from the nitrogen-fixing cyanobacteria. *J Phycol* **49**:1024–1035.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, *et al.* (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**:90–94.

Villar E, Farrant GK, Follows M, Garczarek L, Speich S, Audic S, *et al.* (2015). Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science (80-)* **348**. doi:10.1126/science.1261447.

Zani S, Mellon MT, Collier JL, Zehr JP. (2000). Expression of nifH genes in natural microbial assemblages in Lake George, New York, detected by reverse transcriptase PCR. *Appl Environ Microbiol* **66**:3119–3124.

Zehr JP. (2015). How single cells work together. *Science (80-)* **349**:1163–1164.

Zehr JP. (2011). Nitrogen fixation by marine cyanobacteria. *Trends Microbiol* **19**:162–73.

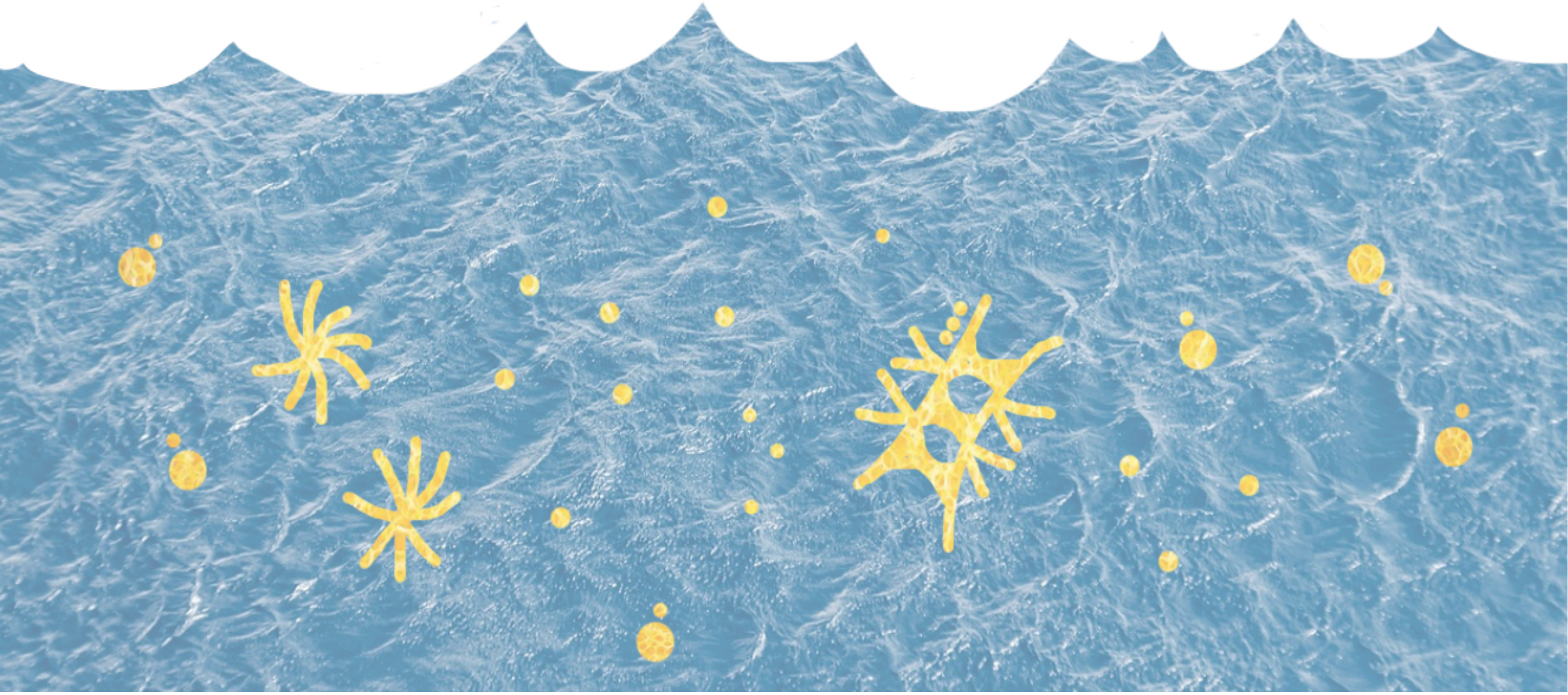
Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, *et al.* (2008). Globally distributed uncultivated oceanic N2-fixing cyanobacteria lack oxygenic photosystem II. *Science* **322**:1110–2.

Zehr JP, Jenkins BD, Short SM, Steward GF. (2003). Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Environ Microbiol* **5**:539–554.

Zehr JP, Mellon MT, Zani S. (1998). New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (nifH) genes. *Appl Environ Microbiol* **64**:3444–3450.

Zehr JP, Shilova IN, Farnelid HM, Muñoz-Maríncarmen MDC, Turk-Kubo KA. (2016). Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. *Nat Microbiol* **2**. doi:10.1038/nmicrobiol.2016.214.

Acknowledgements



AGRADECIMIENTOS

Lo pensé, mira que lo pensé, sabía que el mayor problema al que me iba a enfrentar si la tesis se alargaba era a éste…cómo acordarme de tanta gente con la que he compartido esto que, según diferentes autores, se define como etapa, trabajo, tortura, calvario, parto, sin vivir, agonía, muerte a pellizcos, o simplemente tesis. Por contextualizar, mientras voy haciendo memoria, gran parte de esta tesis habla del beneficio que obtienen algunos individuos o algunas especies al asociarse. Esta interacción se puede categorizar dependiendo del grado en que las partes se vean beneficiadas (o incluso perjudicadas). En este sentido, mi interacción con las personas que me han acompañado durante este corto periodo evolutivo podría entrar dentro de lo que se conoce como comensalismo o, en ciertas ocasiones, incluso como parasitismo. De entre todo el elenco de personas parasitadas, la que se ha llevado la peor parte probablemente hayas sido tú, Silvia Acinas. De ti he aprendido muchísimo, tanto de tu presencia como de tus ausencias, de tus aciertos y de tus errores, de tus victorias y de tus derrotas y, sobre todo, de tu vocación y de tu entusiasmo. Gracias por haber confiado en mi y por haberme abierto tantas puertas, por haberme dado la posibilidad de conocer tantos países a través de la ciencia, de participar en campañas oceanográficas, en definitiva, de conocer a tanta gente. Tú me has enseñado a dar mis primeros pasos en esto de la ciencia… espero de corazón que podamos seguir caminando juntos en el futuro.

En este aprender a caminar me he agarrado de la mano de muchas personas, muchas de ellas afincadas en esta tribu o comuna perrofláutica que es el ICM (a veces pienso que en lugar de estar en un centro de investigación estoy en una de las sedes de la CUP). *“¡¡Como en el ICM en ningún sitio!!”* esto lo he escuchado de muchos de vosotros y, aunque estoy seguro de que esto es así, ahora que me voy supongo que lo viviré en mis propias carnes. Es verdad que queda mucho por mejorar a nivel institucional pero, a nivel personal y profesional, este sitio es un *bastinazo* (buscar en el diccionario andaluz de la lengua). Podría empezar a escribir una lista infinita enumerando los responsables que hacen que esto sea así, pero esto encarecería mucho la impresión del documento y, como *la pela es la pela*, para reducir gastos y en un alarde de irresponsabilidad, destacaré a los que se me vengan a la mente así a bote pronto.

Vane y Clara, sólo conozco a dos criaturas que juntas tengan más peligro que vosotras, los velociraptors de Jurassic Park. Gracias por la guía práctica titulada *Lo que realmente hay que saber para hacer biología molecular*. Irene, gracias por enseñarme a hacer visible lo invisible sin dejar ni un porta fuera de su sitio (me gustaría ver el cajón de los cubiertos de tu casa, sólo por curiosidad). Gracias de verdad a todas las técnicas y técnicos, sin vuestra paciencia y buenas maneras esto no tiraría para delante. Sois la balsa de madera que nos permite flotar en nuestros inicios.

¡A los jefes y jefas! A los peces gordos del departamento, ¡y no me refiero sólo a ti Pep Gasol! (sorry Pep). Realmente es envidiable el chiringuito que habéis montado. Cuando os veo pienso, yo también quiero montar un centro de investigación con mis amigos. Es súper estimulante estar rodeado de tanto conocimiento con patas, y la sinergia que promovéis (y provocáis) entre todos apoya aún más me teoría de que el ICM es como una comuna hippie.

También me gustaría agradecer mi paso por el ICM al colectivo definido como *the white walkers*, es decir, gente paliducha con el ánimo más o menos mermado que, incansablemente, con recompensa o sin ella, intenta atravesar el muro. Este calificativo normalmente se atribuye a los estudiantes de doctorado pero, cuando uno observa con atención, ve que los postdocs también lo sufren. Así, dependiendo de si se trata de un colectivo o de otro, el muro podría representar 'la tesis' o la siguiente convocatoria para tener el quinto postdoc o, con mucha suerte, una plaza fija en un centro que no se aleje más de 3000 Km del sitio donde realmente quieres estar. Gracias a todos vosotros, compañeros y compañeras, amigos y amigas, a los pasados y a los presentes, gracias por las conversaciones más o menos profundas (incluidas las científicas) regadas con café, vino, cerveza o chupito de licor de yerbas, gracias por vuestras sonrisas, por vuestra empatía y por hacer que alguien de fuera se sienta como en su propia casa. De los que aún andan por aquí, les doy especialmente las gracias a Marta S. y a Pablo S., ¡los maduritos interesantes! Gracias por demostrar continuamente, sobre todo tú Pablo, que hay vida después de los 40, sois un ejemplo a seguir. A Eli Alacid, la pareja perfecta de volley-pista, volley-playa, carreras nocturnas, gyn-tonics espontáneos, flamenco y de lo que haga falta, pero sobre todo gran amiga… ¡mucha suerte en tu camino! De los que más o menos se fueron, me gustaría destacar a Guillem, compañero de aventuras, sobre todo durante los comienzos, sin ti esto no hubiese sido lo mismo. A Massimo, gracias por compartir tu visión artística de la ciencia, y a Sara Z. por tu perseverancia frente a las adversidades, sois la pareja perfecta. A Juancho, por tantas conversaciones vitales. A Sarah J., por tu energía incombustible. A Mariona, por tu complicidad en forma de Mario Benedetti. A María de la Fuente, por compartir tus pies con el mundo. A Pati J., Aránzazu, Hugo, Rachele, Roy, Elena, Andrea, Cristina, Marta Royo, Isabel F., Elisa F., Ramiro, Caterina, Mireia… sois tantos con los que he compartido tantos momentos durante estos años y a los que me gustaría dedicaros unas palabras de agradecimiento que se me hace imposible hacerlo como se merece.

Cuando uno se embarca en este viaje de la tesis, a veces, con suerte y sobre todo con dosis industriales de biodramina, también se embarca en el sentido más estricto del verbo embarcarse. Y sí, entre vómito y vómito, conoces a gente estupenda que durante un corto periodo de tiempo se vuelve tu hermano o tu amiga del alma. Juan, Laura y Quique, fue un auténtico placer atravesar el Atlántico sur con vosotros, cada vez que hago memoria me muero de la risa. Loïc, Danu, Bianca, Luis, Gabriele et Sarah S. (my sister!), merci d'avoir fait le voyage sur le bateau Tara encore plus magique. I would like to thank people that I have met during my stays abroad, specially to Laurence and Greg during my time in Roscoff (France), and to Lucas Stal, Veronique, and to the 'De Keete' community, also known as the Big Brother House in Yerseke (The Netherlands).

Y como no todo es trabajo en esta vida, sobre todo en este país (sólo hay que ver la tasa de desempleo), también quiero hacer un agradecimiento especial a mi familia flamenca de Barcelona, a todos estos *catalufos* con alma chirigotera que son capaces de viajar a Cádiz en platillo volante y que hacen que cada viernes se convierta en una fiesta por alegrías. Y en especial me gustaría agradecerte a ti, Isabelle Laudenbach, esta faceta de mi vida en Barcelona, eres una persona realmente especial. Me alegro mucho de que nuestras vidas se hayan cruzado.

También querría agradecer a mis amigos de San Fernando su apoyo y los consejos que me habéis dado durante estos años, en especial a David, a Lauzara y a Álvaro. Y también a las personas especiales que vas conociendo en el transcurso de tu vida académica y que se quedan por siempre a tu lado. Hablo de ti, Ana Fernández, gracias por todos los cafés (con licor o sin licor) que compartimos en Santiago, por estar disponible siempre y por hacer que los meses/años que pasamos sin hablarnos no se noten.

Y hablando de personas especiales… Clara Ruiz, reencontrarme contigo ha sido lo mejor que me ha pasado en mucho tiempo. Consigues mezclar vocación y devoción en la dosis justa para no echar en falta ni aborrecer ninguna de ellas. Gracias por ayudarme tanto, en esta etapa final y en las anteriores, y por confiar en mí. Tu sonrisa incombustible, tu carácter balsámico y tus ganas de todo hacen que cada día sea especial. Pero sobre todo, gracias por quererme y entenderme.

A mi familia, por supuesto, en especial a mis padres, Isabel y Paco, un ejemplo de lo que es luchar y querer a los hijos, bueno, concretamente a cinco hijos. Esta tesis es vuestra. Sin libros ni guías habéis conseguido plasmar en nosotros el amor y el respeto por las personas. Una familia de siete personas viviendo del sueldo de un obrero… está claro que el dinero no da la felicidad, sino que sois vosotros los que, con más imaginación que medios, nos la habéis dado. A mis hermanos y hermanas, Isa, Juan Antonio, Manolo e Inma (fuego, tierra, agua y aire), gracias por entender mis ausencias y sobre todo por vuestro apoyo y vuestra fuerza en la distancia. Os quiero con locura.

A la familia Cabello por supuesto, también os quería agradecer enormemente vuestro apoyo incondicional, Antonio (padre e hijo) y Ana María, siempre he sentido vuestro cariño hacia mí. Sé que siempre podré contar con vosotros, sois unas personas fenomenales. Gracias por preocuparos por mí y por cuidarme durante este tiempo. ¡Ah! Y por llenar mi coche de jamón, solomillo, sandías, naranjas, etc, etc. ¡Cada viaje de Sevilla a Barcelona nos daba para vivir dos meses!

Y hablando de viajes… de Sevilla a Santiago, de Santiago a Barcelona, de Barcelona a Bilbao, ¡¡y de Bilbao a Santa Cruz (CA)!! Gracias a ti Ana Mari. ¡*Pa habernos matao!* Gracias por crecer conmigo, por quererme, por acompañarme, por hacerme feliz, por ser mi amiga incondicional, gracias por todo. Nada de esto habría sido posible ni tendría sentido sin ti . Eres la persona más grande que he conocido, me siento muy afortunado por haber caminado contigo todo este trecho y por la oportunidad que nos brinda el destino de caminar un cachito más. Te deseo todo lo mejor en la vida porque te lo mereces.

En definitiva, gracias a todos por vuestro cariño, hacéis que me sienta la persona más afortunada del mundo.