UCYN-A3, a newly characterized open ocean sublineage of the symbiotic N_2-fixing cyanobacterium *Candidatus* Atelocyanobacterium thalassa

Francisco M. Cornejo-Castillo[1,2,#], Maria del Carmen Muñoz-Marin[1,3,#], Kendra A. Turk-Kubo[1], Marta Royo-Llonch[2], Hanna Farnelid[1,4], Silvia G. Acinas[2], Jonathan P. Zehr[1]

[1] Department of Ocean Sciences, University of California, Santa Cruz, CA 95064, USA.

[2] Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM), CSIC, Pg. Marítim de la Barceloneta 37-49, 08003 Barcelona, Spain.

[3] Departamento de Bioquímica y Biología Molecular, Edificio Severo Ochoa, planta 1, Universidad de Córdoba, 14071-Córdoba, Spain.

[4] Centre for Ecology and Evolution in Microbial Model Systems (EEMiS), Linnaeus University, 392 34 Kalmar, Sweden.

# These authors contributed equally to this manuscript.

Corresponding author:

Email: b32mumam@uco.es

Abbreviations: N_2, dinitrogen; UCYN-A, unicellular cyanobacterial group A; *nifH*, nitrogenase; DAPI, 4′ , 6- diamidino-2-phenylindole; CARD-FISH, Catalyzed Reporter Deposition-Fluorescence In Situ Hybridization; SAG, single amplified genome.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/1462-2920.14429

Summary

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. However, very little is known about UCYN-A3 beyond the *nifH* sequences from *nifH* gene diversity surveys. In this study, single cell sorting, DNA sequencing, qPCR and CARD-FISH assays revealed discrepancies involving the identification of sublineages, which led to new information on the diversity of the UCYN-A symbiosis. 16S rRNA and *nifH* gene sequencing on single sorted cells allowed us to identify the 16S rRNA gene of the uncharacterized UCYN-A3 sublineage. We designed new CARD-FISH probes that allowed us to distinguish and observe UCYN-A2 in a coastal location (SIO Pier; San Diego) and UCYN-A3 in an open ocean location (Station ALOHA; Hawaii). Moreover, we reconstructed about 13% of the UCYN-A3 genome from *Tara* Oceans metagenomic data. Finally, our findings unveil the UCYN-A3 symbiosis in open ocean waters suggesting that the different UCYN-A sublineages are distributed along different size fractions of the plankton defined by the cell-size ranges of their prymnesiophyte hosts.

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., 2012). 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 (Zehr and Turner, 2001; 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). 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 (Zehr and Turner, 2001; Thompson and Zehr, 2013) which has suggested it has a major role in global N_2 fixation (Moisander et al., 2010; Farnelid et al., 2016; Martínez-Pérez 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). The UCYN-A1 and UCYN-A2 genomes have been sequenced (Tripp et al., 2010; Bombar et al., 2014). Using cell sorting and qPCR it was discovered that the UCYN-A2 sublineage is specifically associated with *B. bigelowii* while UCYN-A1 is associated with a smaller, but closely-related prymnesiophyte (1-3 μm) (Thompson et al., 2012; Thompson et al., 2014). 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 that originally described (Thompson et al., 2014), suggesting that UCYN-A2 symbiosis 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). 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; Martínez-Pérez et al., 2016; Turk Kubo et al., 2017; Gérikas et al., 2018; Stenegren et al., 2018). 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). 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: at Station ALOHA in the NPSG and in 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. The identification of the UCYN-A sublineages was also carried out using double CARD-FISH assays. We used the 16S rRNA data from single amplified genomes previously linked to UCYN-A2 and UCYN-A3 by sequencing also the *nifH* gene to design new CARD-FISH probes that allowed us to determine cell size ranges and morphological features of UCYN-A sublineages. We analyzed size-fractionated metagenome sequence libraries collected during the *Tara* Oceans expedition in the South Atlantic to detect and reconstruct the genome of the UCYN-A3, and to determine the evolutionary relationships of the UCYN-A sublineages with other unicellular N_2-fixing cyanobacteria.

Results

1. Diversity of UCYN-A

1.1. *nifH* gene sequences

UCYN-A diversity was assessed by amplifying *nifH* gene sequences from the SIO Pier and from 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% of total UCYN-A *nifH* sequences. UCYN-A2 sequences were also recovered and the sequence type with the highest relative abundance 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 sublineage 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 abundance (ca. 0.7%).

1.2 Visualization of UCYN-A associations

1.2.1. UCYN-A associations at 2 Pacific Stations (Hawaii and SIO Pier)

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.

At SIO Pier, the UCYN-A1 host averaged 2.3 ± 0.3 µm (n=100 cells) (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 3A, 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 of the whole group of cells (n=100 cells) (Figure 2, 3C, D).

The size of both 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 2, 3A, E). However, at Station ALOHA, using the UCYN-A2 and Host-A2 CARD-FISH probes, the targeted host was significantly 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 2, 3C, G). 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 hypothesize that the UCYN-A2 CARD-FISH probe targets both UCYN-A2 and UCYN-A3.

1.2.2. Design of specific CARD-FISH probes for UCYN-A2 and UCYN-A3

The high abundance of UCYN-A previously detected in metagenomic samples of two *Tara* Oceans stations located in the South Atlantic Ocean (Cornejo-Castillo et al., 2016) encouraged us to perform single cell sorting of seawater samples collected in parallel in these stations. In particular, the single cell sorting of the picoeukaryotic population of the *Tara* station St78_SRF combined with the sequencing of the 16S rRNA and *nifH* genes of the sorted cells allowed us to identify the 16S rRNA gene sequence of the UCYN-A3 sublineage. The UCYN-A3 16S rRNA sequence shares 98.9% and 98.6% identity with the UCYN-A1 and UCYN-2 16S rRNA genes, respectively. Interestingly, the UCYN-A2 CARD-FISH probe matched perfectly with the UCYN-A3 16S rRNA sequences confirming our suspicions about the lack of specificity of the UCYN-A2 CARD-FISH assay (Supplemental Figure 1). We designed two new CARD-FISH probes to distinguish UCYN-A2 and UCYN-A3 sublineages and applied the new CARD-FISH probes in samples collected from Station ALOHA and the SIO Pier (Supplemental Figure 1 and 2). At SIO Pier, both the UCYN-A2 symbiont and its host were detected, but not the UCYN-A3 symbiont. In contrast, at Station ALOHA, we detected the UCYN-A3 with the UCYN-A2 host, but did not detect the UCYN-A2 symbiont with the UCYN-A2 host. Although we detected the UCYN-A2 host in both locations, SIO Pier and ALOHA, we cannot be confident that the UCYN-A2 host CARD-FISH probe would not capture closely related, but distinct lineages of *B. bigelowii.*

1.2.3. New UCYN-A associations

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 (Supplemental Figure 3). 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 (Supplemental Figure 3). 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) (Figure 3A). 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 0.21 cells mL^{-1}, based on CARD-FISH epifluorescence cell counts. Because sequencing of the *nifH* gene detected UCYN-A4 at very low relative sequence abundances (2.9-6.1%), we could hypothesize that this new symbiosis could be UCYN-A4.

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 (Supplemental Figure 4). 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 121 ± 8 *nifH* copies mL^{-1} and 477 ± 37 18S rRNA gene

copies mL^{-1}, respectively, with the host gene copy ratios four times greater than the symbiont (Supplemental Figure 4). Considering that based on the UCYN-A *nifH* libraries the UCYN-A3 sublineage was virtually absent in the SIO samples, the UCYN-A2/A3 qPCR assay was likely detecting and quantifying UCYN-A2. The UCYN-A1 averaged 194 ± 16 *nifH* copies mL^{-1}, but the prymnesiophyte host originally described by Thompson et al. (2012) was not detected using the UCYN-A1 host assay (Supplementary Table 1).

At Station ALOHA, both UCYN-A1 and UCYN-A1 hosts were detected at gene copy ratios averaging 303 ± 14 *nifH* copies mL^{-1} and 335 ± 25 18S rRNA gene copies mL^{-1}, respectively (Supplemental Figure 4). Likewise, the gene copies of UCYN-A2/UCYN-A3 *nifH* gene were on average 60 ± 46 *nifH* copies mL^{-1} and the 18S rRNA gene of the UCYN-A2 host averaged 20 ± 7.9 rRNA gene copies mL^{-1} (Supplemental Figure 4). Considering the results of the UCYN-A *nifH* libraries where the relative abundance of the UCYN-A2 sublineage was extremely low at Station ALOHA, the UCYN-A2/A3 qPCR assay was likely detecting and quantifying UCYN-A3.

2. UCYN-A3 characterization from metagenomes

2.1 Metagenomic detection of a new UCYN-A population.

During the *Tara* Oceans expedition, the distributions of the UCYN-A1 and UCYN-A2 sublineages were 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.

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 2), encompassing the cell size ranges of the small

prymnesiophyte partner (<3 μm) observed in all previous studies (Thompson et al., 2012; Thompson et al., 2014; Cornejo-Castillo et al., 2016; Martínez-Pérez et al., 2016). Interestingly, although reads assigned to the UCYN-A2 sublineage (>95% identity with UCYN-A2 genome) were virtually absent at station TARA_076, we recruited reads showing sequence identities that ranged from 80-95% with the UCYN-A2 genome in metagenomes covering both the 0.8-5 and >0.8 μm size-fractions (Figure 4, Supplementary Table 2). The recruitment of these divergent metagenomic reads occurred within the size-fraction corresponding with the cell size-range of the UCYN-A3 symbiosis (3.6 μm ± 0.7 μm; n=100 cells) measured at Station ALOHA (Figure 2) suggesting that this new divergent UCYN-A genome sequence population was the UCYN-A3 sublineage.

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 5). 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 3). We reconstructed 13% of the UCYN-A3 genome, assuming that the genome size of the UCYN-A3 genome was similar to the genome sizes of the two UCYN-A genomes sequenced to-date (ca. 1.4 Mb). In order to verify the taxonomic affiliation of the reconstructed genes, 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. Only one of the reconstructed UCYN-A3 gene sequences (out of 293) had the UCYN-A1 genome as its best hit. This gene sequence does not have a homologous gene sequence in the UCYN-A2 genome, likely because we are still missing a few genes in the

UCYN-A2 genome because the UCYN-A2 genome is not yet closed. In addition, two reconstructed UCYN-A3 gene sequences do not have homologous gene sequences in the UCYN-A1 genome, suggesting that the UCYN-A3 genome is more similar in terms of genome content to UCYN-A2 than to UCYN-A1. Finally, all of the contigs containing more than one gene had the same gene ordination as in the UCYN-A2 genome.

2.3 Phylogenomic analysis and evolution of the UCYN-A3 sublineage

A phylogenetic tree was constructed in order to place the new UCYN-A3 sublineage in its evolutionary context. Maximum likelihood analysis of a total of 165 protein-coding genes (Supplementary Table 3; 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 (Supplemental Figure 6). Furthermore, UCYN-A3 formed a sub-group with UCYN-A2 (Supplemental Figure 6).

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 4). 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 were not statistically significant (P>0.05, Codon Based Z-test) (Supplementary Table 4).

Discussion

UCYN-A3 characterization

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 the 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 Table 1. 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 morphologies of the UCYN-A1 symbiosis in both environments. The *nifH*:18S rRNA genes ratio in the UCYN-A1 symbiosis was close to 1:1 (~0.9) at Station ALOHA, which is consistent with the 1:1 symbiont-host cell ratio previously observed (Thompson et al., 2014). Zhu and their colleagues suggested a possible correlation between rDNA copy number and organism size (Zhu et al., 2005). Based on this correlation, cells 1 µm in diameter would have around 1 rRNA gene copy number which is consistent with our results.

Based on qPCR and CARD-FISH assays, it originally appeared that UCYN-A2 was present at both stations. However, our findings suggest that the sublineage detected at Station ALOHA was in fact UCYN-A3, not UCYN-A2. First, UCYN-A3 *nifH* amplicon sequences were present at relative abundances of approximately 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 of the UCYN-A2 symbiosis observed at the SIO Pier and at Station ALOHA hinted that the CARD-FISH probes designed to target the UCYN-A2 sublineage also hybridizes to UCYN-A3. For this reason, new specific probes for UCYN-A2 and UCYN-A3 were designed since the 16S rRNA gene sequence of the UCYN-A3 lineages was obtained by combining single cell sorting and sequencing of samples collected from the South Atlantic Ocean by the *Tara* Oceans expedition.

The UCYN-A1 and UCYN-A2 symbioses have been reported in most major ocean basins (Cabello et al., 2015; Martínez-Pérez et al., 2016; Turk Kubo et al., 2017) and UCYN-A3 appears also to have a wide distribution (Turk-Kubo et al., 2017). Therefore using different data sets was necessary to probe the wide dispersion of this new subclade. Moreover, this new data set was essential to design a new CARD-FISH assay allowing us to observe the presence of UCYN-A3 at Station ALOHA as well as its absence at SIO Pier, indicating that the UCYN-A association originally assumed to be UCYN-A2 at Station ALOHA was actually UCYN-A3.

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 in 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.

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 from the *Tara* Oceans metagenomes was the *nifH* gene, which was key to identify this new divergent population as UCYN-A3. However, 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 relationship among UCYN-A sublineages showed that UCYN-A3 is closer to UCYN-A2 than to UCYN-A1, and that UCYN-A3 shares a more recent common ancestor with UCYN-A2 than the ancestor shared with UCYN-A1 (Supplemental Figure 6). 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. 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.

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 and Supplementary Table 5). 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 (Supplemental Figure 7). 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 by CARD-FISH is closely related to the originally detected UCYN-A1 prymnesiophyte host since the Host-A1 probe for CARD-FISH hybridized with the host (Figure 3A, B). Future analysis of 18S metabarcoding will be needed to explore whether UCYN-A1 is associated to a new prymnesiophyte host at the SIO Pier.

In contrast to the 3.3 UCYN-A2 cells per *B. bigelowii* reported by Thompson et al., (2014) (the authors assumed one copy of *nifH* in UCYN-A2 and one copy of the 18S rRNA gene in *B. bigelowii*), in our samples, we observed the opposite pattern, that is, the 18S rRNA gene copy number in the host-A2 (*B. bigelowii*) was almost 4 times greater than the *nifH* gene copy number in UCYN-A2. Zhu et al. (2005) proposed that a protist with the size of the UCYN-A2 host, i.e. around 7 µm, should have approximately 40 copies of the 18S rRNA gene per cell. If we apply this rationale to our case, and taking into account that the 18S rRNA gene copy number of the host averaged 477 ± 37 copies mL^{-1}, we would expect to detect in our samples around 12 UCYN-A2 *nifH* gene copies mL^{-1} (assuming 1 UCYN-A2 cell per host). However, the UCYN-A2 *nifH* gene abundance was 121 ± 8 copies mL^{-1}, suggesting 10 UCYN-A2 cells per A2-host, in agreement with previous observations

(Cornejo-Castillo et al., 2016).

The non-specificity of the UCYN-A2 host qPCR assay could explain the detection of the "UCYN-A2 host" at Station ALOHA (20 copies of the 18S rRNA gene mL^{-1}) despite the absence or extremely low relative abundance of the UCYN-A2 sublineage based on both CARD-FISH and *nifH* sequencing respectively at Station ALOHA (Figure 1). However, these results along with the CARD-FISH analyses might suggest that UCYN-A3 could be associated with a different but closely genetically similar host to the host of UCYN-A2, but with a cell size intermediate between the sizes of the UCYN-A1 and UCYN-A2 hosts (3.61 ± 0.67 µm (Figure 3G, H)). In the future, single cell sorting studies combined with *nifH* and 18S rRNA gene sequencing should provide clues to the characterization of the different partnerships.

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 ecology, diversity and distribution of UCYN-A sublineages. The wide occurrence of multiple UCYN-A sublineages in symbiosis with prymnesiophyte hosts that vary in size suggests that UCYN-A may contribute to marine N_2 fixation in multiple size-fractions.

Moreover, thanks to the design of new CARD-FISH probes, this is the first study reporting microscopic images of the UCYN-A3 sublineage. In addition, a partial genome of the UCYN-A3 sublineage was reconstructed using a novel approach that combines fragment recruitment and genome assembly techniques. With the availability of these UCYNA-3 gene sequences, the recruitment of the whole genome and metatranscriptome 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 and host sublineages, which is critical for elucidating the ecology and the evolution of the UCYN-A symbiosis.

Experimental Procedures

Sampling procedures

Samples were collected from the Scripps Institution of Oceanography (SIO) Ellen Browning Scripps Memorial Pier (32º 52´N, 117º 15.4´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 00´W at 45m) 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 was 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 (Millipore Isopore TM, EMD Millipore, Billerica, MA, USA) with a support filter of 0.8 µm pore-size, 25 mm polycarbonate cellulose acetate membrane filter (Sterlitech Corporation, Kent, WA, USA) 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, Port Washington, NY, USA) using a peristaltic pump. At Station ALOHA, 4 L samples were filtered onto 0.22 µm pore-size Sterivex cartridges (Millipore Corp., Billerica, MA, USA) 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.

Standards for each assay were included as positive controls and to quantify the copy numbers of *nifH*. 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). 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 estimating the rRNA gene copy number based on cell size in the host (Zhu et al., 2005).

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 against UCYN-A, by targeting the 16S rRNA, and against the host, by targeting the 18S rRNA, are compiled in Supplementary Table 1. 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) software.

Sampling, generation and sequencing of single sorted cells

Surface seawater samples (5 m deep, not pre-filtered, station TARA_078) were collected for single cell analysis from the South Atlantic Ocean during the circumnavigation expedition *Tara* Oceans. Replicated 1 mL aliquots of seawater were cryopreserved with 6% glycine betaine (Sigma) and stored at −80ºC. Single cell sorting of the picoeukaryotic community, Multi-displacement Amplification (MDA) of the sorted cells and preliminary 16S rRNA screening with primers 27F (Page et al., 2014) and 907R (Lane, 1991) were performed at the Bigelow Laboratory Single Cell Genomics (scgc.bigelow.org). A nested PCR protocol was used to amplify an ~359 bp region of the nitrogenase *nifH* gene on single sorted cells, using degenerate primers: nifH3, nifH4, nifH1 and nifH2 (Zehr and Turner, 2001). Polymerase Chain Reaction (PCR) products were purified and sequenced by Genoscreen (Lille, France) with OneShot Sanger sequencing.

Accession numbers for PCR products of the UCYN-A3 sublineage: partial 16S rRNA gene (MH807559); partial *nifH* gene (MH815013).

Design of the UCYN-A2 and UCYN-A3 CARD-FISH probes

For the design of specific oligonucleotide probes targeting UCYN-A2 and UCYN-A3 sublineages, 16S rRNA gene sequences including all the defined UCYN-A1 and UCYN-A2

sublineages, and the UCYN-A3 16S rRNA gene sequence obtained from our sorted cells were aligned using MAFFT (Katoh *et al.*, 2002). The newly designed probe UCYN-A2-1137 (5'-CTTCCTAAAGTGCCCACCTT-3') targeted the UCYN-A2 sublineage, while probe UCYN-A3-1137 (5'-CTTCCTAGAGTGCCCACCTT-3') targeted the UCYN-A3 sublineage. UCYN-A2-1137 and UCYN-A2-1137 probes differed in only one position, and required a competitor in order to avoid unspecific hybridizations. Therefore, the labeled probe UCYN-A2-1137 was used in combination with the unlabeled UCYN-A3-1137 oligonucleotide for the detection of UCYN-A2, and *vice versa* for the detection of UCYN-A3. One helper, Helper UCYN-A2A3 (5'- CGGTTTGTCACCGGCAGT -3') was designed to improve the hybridization process for both probes. The specificity of the new probes was checked with the online tool TestProbe (https://www.arb-silva.de/search/testprobe/) and by searching in the GenBank database (http://www.ncbi.nlm.nih.gov/index.html) to detect potential matching sequences in non-target groups.

nifH gene amplicon sequencing and 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 (Zehr and Turner, 2001) 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 acquired from USEARCH v6.1 output files using QIIME scripts, and visualized using the interactive tree of life (iTOL) web tool (Letunic and Bork, 2007). Raw UCYN-A *nifH* sequences are deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject Accession PRJNA488464.

Fragment recruitment and UCYN-A3 genome reconstruction

The analysis of the abundance of UCYN-A based on 16S $_{mi}$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 μ m) 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 4). 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, as described in Caro-Quintero and Konstantinidis (2011). 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 (Dinghua 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.

ACKNOWLEDGEMENTS

We are indebted to the scientists and crew on board of *Tara* Oceans expedition and the C-MORE Cruise. We thank F. Azam (Scripps Institution of Oceanography, UC San Diego) for access to Scripps facilities and 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 also thank M. Hogan for lab and field support and Benjamin Abrams and the UCSC Life Science Microscopy Center for microscopy assistance. This is *Tara* Oceans contribution paper number 81.

J.P.Z was supported by grants from the Simons Collaboration on Ocean Processes and Ecology (SCOPE, grant 329108), the Simons Foundation (grant 545171), and Gordon and Betty Moore Foundation (grant 493.01). Funding was also provided by project MAGGY (CTM2017-87736-R) from the Spanish Ministry of Economy and Competitiveness to SGA. M.M.M. was supported by a Marie Curie International Outgoing Fellowship within the 7th European Community Framework Programme. F.M.C-C was supported by a Marie Curie Individual Global Fellowship within the Horizon 2020 European Framework Programme (project UCYN2PLAST, grant no. 749380). MR-L held a Ph.D. Fellowship FPI (BES-2014-068285) funded by the Spanish Ministry of Economy and Competitivity. H.F. was supported by the Swedish Research Council VR 637- 2013-7502.

Author contributions

M.M.M. and F.M.C-C. designed the study. M.M.M. carried out the sampling at the SIO Pier and Station ALOHA, analyzed CARD-FISH and quantitative PCR assays and extracted DNA for the sequence processing. F.M.C-C. designed and tested the new CARD-FISH probes, performed the fragment recruitment, the UCYN-A3 genome reconstruction and the phylogenomic analysis of UCYN-A sublineages. K.T. designed and analyzed the sequence processing and prepared accompanying figures. M.R-L performed the PCR and sequencing of the single cell sorted cells. H.F. aided sampling UCYN-A samples at the Scripps Pier. SGA contributed to the SAG sequencing and access analysis of *Tara* Oceans-related data. J.P.Z. conceptualized the study. F.M.C-C., M.M.M., K.T., H.F., S.G.A and J.P.Z. drafted and edited the manuscript and figures. All authors read and approved the final manuscript.

Competing financial interests

The authors declare no competing financial interest.

References:

Bombar, D.a.H., P, Sanchez-Baracaldo, P., Carter, B.J., and Zehr, J.P. (2014) Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *ISME Journal* **8**: 250-2542.

Cabello, A.M., Cornejo-Castillo, F.M., 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, M., Short, C., Jenkins, B., Karl, D., and Zehr, J. (2005) Temporal Patterns of Nitrogenase Gene (nifH) Expression in the Oligotrophic North Pacific Ocean. *Appl Environ Microbiol*.

Cornejo-Castillo, F.M., Cabello, A.M., Salazar, G., Sanchez-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**: 11071.

Farnelid, H., Turk-Kubo, K., Muñoz-Marin, M., and Zehr, J. (2016) New insights into the ecology of the globally significant uncultured nitrogen-fixing symbiont UCYN-A. *Aquat Microb Ecol* **77**: 125-138.

Karl, D.M., Church, M.J., Dore, J., Letelier, R., and Mahaffey, C. (2012) Predictable and efficient carbon sequestration in the North Pacific Ocean supported by simbiotic nitrogen fixation. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 1842-1849.

Krupke, A., Musat, N., LaRoche, J., Mohr, W., Fuchs, B.M., Amann, R.I. et al. (2013) In situ identification and N_2 and C fixation rates of uncultivated cyanobacteria populations. *Systematic and Applied Microbiology* **36**: 259-271.

Luo, Y.W., Doney, S.C., Anderson, L.A., 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, C., Dekaezemacker, J., Littmann, S., Yilmaz, P. et al. (2016) Small unicellular diazotrophic symbiont is a key player in the marine nitrogen cycle. *Nature Microbiology*.

McDonald, J., and Kreitman, M. (1991) Adaptative protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652-654.

Moisander, P.H., Beinart, R.A., Hewson, I., White, A.E., Johnson, K.S., Carlson, C.A. et al. (2010) Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* **327**: 1512-1514.

Sohm, J., Webb, E., and Capone, D. (2011) Emerging patterns of marine nitrogen fixation. *Nature Reviews Microbiology* **9**: 499-508.

Tai, V., and Palenik, B. (2009) Temporal variation of *Synechococcus* clades at a coastal Pacific Ocean monitoring site. *Isme Journal* **3**: 903-915.

Thompson, A., Carter, B.J., Turk-Kubo, K., Malfatti, F., Azam, F., and Zehr, J.P. (2014) Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol* **16**: 3238-3249.

Thompson, A.W., and Zehr, J.P. (2013) Cellular interactions: lessons from the nitrogen‐fixing cyanobacteria. *Journal of Phycology* **49**: 1024.

Thompson, A.W., Foster, R.A., Krupke, A., Carter, B.J., Musat, N., Vaulot, D. et al. (2012) Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science* **337**: 1546-1550.

Tripp, H.J., Bench, S.R., Turk, K.A., Foster, R., Desany, B.A., Niazi, F. et al. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**: 90-94.
Turk Kubo, K., Farnelid, H., Shilova, I.N., Henke, B., and Zehr, J.P. (2017) Distinct ecological niches of marine symbiotic N2-fixing cyanobacterium Candidatus Atelocyanobacterium Thalassa sublineages. *J Phycol* **53**: 451-461.
Turk-Kubo, K., Farnelid, H., Shilova, I., Henke, B., and Zehr, J.P. (2017) Distinct ecological niches of marine symbiotic N_2-fixing cyanobacterium *Candidatus* Atelocyanobacterium Thalassa sublineages. *Journal of Phycology* **53**: 451-461.
Zehr, J.P., and Turner, P. (2001) Nitrogen fixation: nitrogenase genes and gene expression. *Methods in Microbiology* **30**: 271-286.
Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79-92.
Bombar, D.a.H., P, Sanchez-Baracaldo, P., Carter, B.J., and Zehr, J.P. (2014) Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *ISME Journal* **8**: 250-2542.
Cabello, A.M., Cornejo-Castillo, F.M., 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, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K. et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335-336 %@ 1548-7091.
Church, M., Short, C., Jenkins, B., Karl, D., and Zehr, J. (2005) Temporal Patterns of Nitrogenase Gene (nifH) Expression in the Oligotrophic North Pacific Ocean. *Appl Environ Microbiol*.
Cornejo-Castillo, F.M., Cabello, A.M., Salazar, G., Sanchez-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**: 11071.
Dinghua, L., Chi-Man, L., Ruibang, L., Kunihiko, S., and Tak-Wah, L. (2015) MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **22**: 2688-2690.
Edgar, R. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford)* **26**: 2460-2461.
Edgar, R., Haas, B., Clemente, J., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford)* **27**: 2194-2200.
Farnelid, H., Turk-Kubo, K., Muñoz-Marin, M., and Zehr, J. (2016) New insights into the ecology of the globally significant uncultured nitrogen-fixing symbiont UCYN-A. *Aquat Microb Ecol* **77**: 125-138.
Gérikas, R.C., Lopes Dos Santos, A., Marie, D., Pereira, B.F., and Vaulot, D. (2018) Small eukaryotic phytoplankton communities in tropical waters off Brazil are dominated by symbioses between Haptophyta and nitrogen-fixing cyanobacteria. *ISME J* **12**: 1360-1374.
Green, S.J., Venkatramanan, R., and Naqib, A. (2015) Deconstructing the Polymerase Chain Reaction: Understanding and Correcting Bias Associated with Primer Degeneracies and Primer-Template Mismatches. *PLoS ONE* **10**: e0128122.
Heller, P., Tripp, H.J., Turk-Kubo, K., and Zehr, J.P. (2014) ARBitrator: a software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank. *Bioinformatics* **30**: 2883-2890.

Karl, D.M., Church, M.J., Dore, J., Letelier, R., and Mahaffey, C. (2012) Predictable and efficient carbon sequestration in the North Pacific Ocean supported by simbiotic nitrogen fixation. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 1842-1849.
Krupke, A., Musat, N., LaRoche, J., Mohr, W., Fuchs, B.M., Amann, R.I. et al. (2013) In situ identification and N_2 and C fixation rates of uncultivated cyanobacteria populations. *Systematic and Applied Microbiology* **36**: 259-271.
Lane, D.J. (1991) 16S/23S rRNA sequencing. In *Nucleic acid techniques in bacterial systematics*. Stackebrandt, E., and Goodfellow, M. (eds). Chichester, UK: John Wiley & Sons, pp. 115-175.
Letunic, I., and Bork, P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **233**: 127-128.
Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar et al. (2004) ARB: a software environment for sequence data. *Nucleic Acids Research* **32**: 1363-1371.
Luo, Y.W., Doney, S.C., Anderson, L.A., 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, C., Dekaezemacker, J., Littmann, S., Yilmaz, P. et al. (2016) Small unicellular diazotrophic symbiont is a key player in the marine nitrogen cycle. *Nature Microbiology*.
McDonald, J., and Kreitman, M. (1991) Adaptative protein evolution at the Adh locus in Drosophila. *Nature* **351**: 652-654.
Moisander, P.H., Beinart, R.A., Voss, M., and Zehr, J.P. (2008) Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *ISME Journal* **2**: 954-967.
Moisander, P.H., Beinart, R.A., Hewson, I., White, A.E., Johnson, K.S., Carlson, C.A. et al. (2010) Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* **327**: 1512-1514.
Page, K.A., Connon, S.A., and Giovannoni, S.J. (2014) Representative freshwater bacterioplankton isolated from Crater Lake, Oregon. *Appl Environ Microbiol* **70**: 6542-6550.
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. *Nature Methods* **9**: 676-682.
Sohm, J., Webb, E., and Capone, D. (2011) Emerging patterns of marine nitrogen fixation. *Nature Reviews Microbiology* **9**: 499-508.
Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688-2690.
Stenegren, M., Caputo, A., Berg, C., Bonnet, S., and Foster, R. (2018) Distribution and drivers of symbiotic and free-living diazotrophic cyanobacteria in the western tropical South Pacific. *Biogeosciences* **15**: 1559-1578.
Tai, V., and Palenik, B. (2009) Temporal variation of *Synechococcus* clades at a coastal Pacific Ocean monitoring site. *Isme Journal* **3**: 903-915.
Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* **30**: 2725-2729.
Thompson, A., Carter, B.J., Turk-Kubo, K., Malfatti, F., Azam, F., and Zehr, J.P. (2014) Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ Microbiol* **16**: 3238-3249.

Thompson, A.W., and Zehr, J.P. (2013) Cellular interactions: lessons from the nitrogen - fixing cyanobacteria. *Journal of Phycology* **49**: 1024.
Thompson, A.W., Foster, R.A., Krupke, A., Carter, B.J., Musat, N., Vaulot, D. et al. (2012) Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga. *Science* **337**: 1546-1550.
Tripp, H.J., Bench, S.R., Turk, K.A., Foster, R., Desany, B.A., Niazi, F. et al. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* **464**: 90-94.
Turk Kubo, K., Farnelid, H., Shilova, I.N., Henke, B., and Zehr, J.P. (2017) Distinct ecological niches of marine symbiotic N2-fixing cyanobacterium Candidatus Atelocyanobacterium Thalassa sublineages. *J Phycol* **53**: 451-461.
Turk-Kubo, K., Farnelid, H., Shilova, I., Henke, B., and Zehr, J.P. (2017) Distinct ecological niches of marine symbiotic N_2-fixing cyanobacterium *Candidatus* Atelocyanobacterium Thalassa sublineages. *Journal of Phycology* **53**: 451-461.
Zehr, J.P., and Turner, P. (2001) Nitrogen fixation: nitrogenase genes and gene expression. *Methods in Microbiology* **30**: 271-286.
Zhang, J., Kobert, K., Flouri, T., and Stamatakis, A. (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics (Oxford)* **30**: 614-620.
Zhu, F., Massana, R., Not, F., Marie, D., and Vaulot, D. (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79-92.

Figures and tables

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.

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.

Figure 3. Micrographs of UCYN-A associations at SIO Pier and at ALOHA Station. Left panels **(A,C,E,G),** correspond to the epifluorescence microscopy images using the double-CARD-FISH assay showing the specificity of symbiont–host pairs. Right panels **(B,D,F,H),** correspond to the 4´-6-diamidino-2-phenylindole signal (DAPI; blue).

SIO Pier: (A,B), UCYN-A1 with its prymnesiophyte partner labeled with the UCYN-A1-732 and UPRYM69 probes with competitors. **(C,D)**, UCYN-A2 association labeled with the UCYN-A2-732 and UBRADO69 probes with competitors. **ALOHA Station**: **(E,F)**, UCYN-A1 with its prymnesiophyte host labeled with the UCYN-A1-732 and UPRYM69 probes with competitors. **(G,H)**, UCYN-A3 association labeled with the UCYN-A2-732 and

UBRADO69 probes with competitors. Based on the sequencing results UCYN-A2 was absent at ALOHA Station, but UCYN-A3 was the second most abundant UCYN-A sublineage. We assume this symbiosis was UCYN-A3, not UCYN-A2.

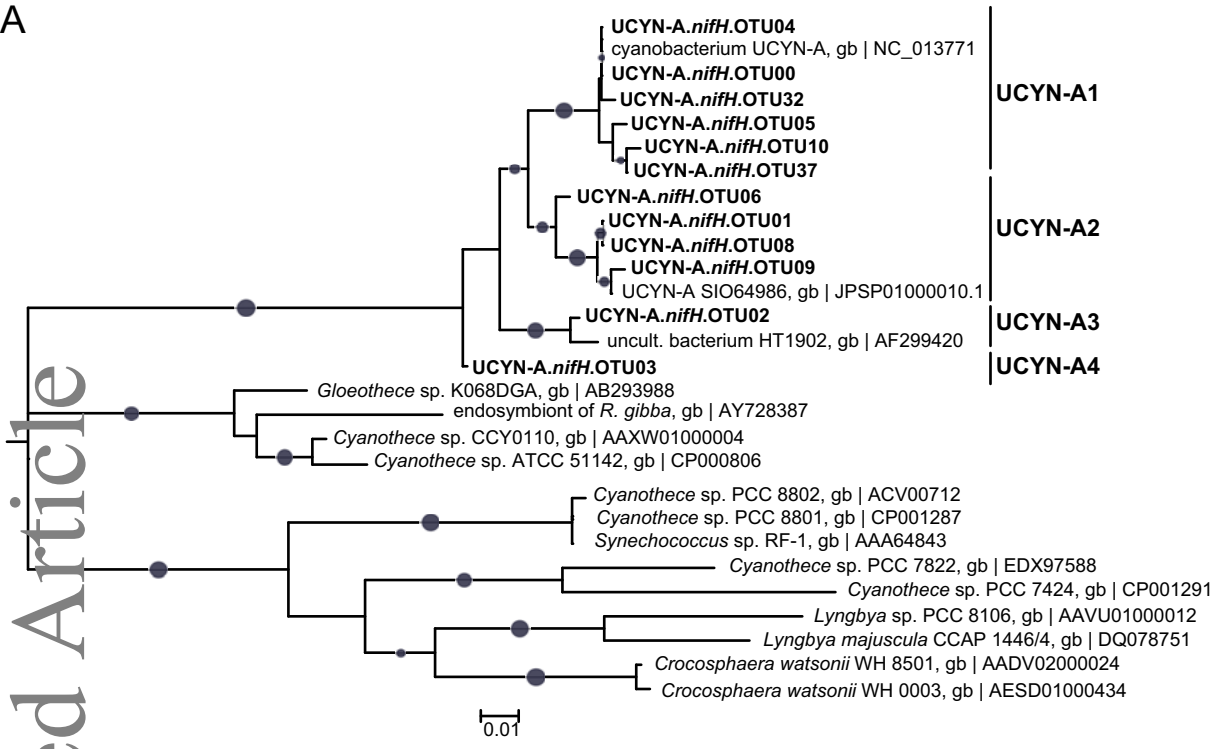
Figure 4: **Metagenome fragment recruitment of UCYN-A lineages in size-fractionated metagenomes from surface waters collected at station TARA_076.** Recruitment plot of metagenomic reads using UCYN-A1 and UCYN-A genomes as reference genomes. Reads are plotted as colored dots, representing the covered genome positions (x axis) and the % of identity with the UCYN-A1 (left) and UCYN-A2 (right) reference genomes (y axis). Reads with identity higher to 95% are considered to represent UCYN-A1 (red dots) or UCYN-A2 (blue dots) populations respectively. Reads with identity lower to 95% were identified as new divergent UCYN-A population, UCYN-A3 (green dots).

Table 1. 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]	+	+
Old UCYN-A2 (CARD-FISH)[4]	+	+
Host-A2 (CARD-FISH)[4]	+	+
New UCYN-A2 (CARD-FISH)[6]	+	-
New UCYN-A3 (CARD-FISH)[6]	-	+
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. (6) This study.

A



B