Title

Distinct ecological niches of marine symbiotic N2-fixing cyanobacterium Candidatus Atelocyanobacterium thalassa sublineages

Journal

Journal of Phycology, 53(2)

ISSN

0022-3646

Authors

Turk-Kubo, Kendra A
Farnelid, Hanna M
Shilova, Irina N
et al.

Publication Date

2017-04-01

DOI

10.1111/jpy.12505

Peer reviewed

DISTINCT ECOLOGICAL NICHES OF MARINE SYMBIOTIC N_2-FIXING CYANOBACTERIUM *CANDIDATUS* ATELOCYANOBACTERIUM THALASSA SUBLINEAGES[1]

Kendra A. Turk-Kubo[2]
Ocean Sciences Department, University of California, Santa Cruz, CA 95064, USA;

Hanna M. Farnelid
Ocean Sciences Department, University of California, Santa Cruz, CA 95064, USA;
Centre for Ecology and Evolution in Microbial Model Systems (EEMiS), Linnaeus University, 392 34 Kalmar, Sweden

Irina N. Shilova
Ocean Sciences Department, University of California, Santa Cruz, CA 95064, USA;

Britt Henke
Ocean Sciences Department, University of California, Santa Cruz, CA 95064, USA;

Jonathan P. Zehr
Ocean Sciences Department, University of California, Santa Cruz, CA 95064, USA;

[1] *Date of submission and acceptance*

[2] Corresponding Author:
Kendra A. Turk-Kubo
Ocean Sciences Department
University of California, Santa Cruz
1156 High Street
Santa Cruz, CA 95064
kturk@ucsc.edu
(831) 459-4490

Running Title: UCYN-A global diversity

Abstract

A recently described symbiosis between the metabolically streamlined nitrogen-fixing cyanobacterium UCYN-A and a single-celled eukaryote prymnesiophyte alga is widely distributed throughout tropical and subtropical marine waters, and is thought to contribute significantly to nitrogen fixation in these regions. Several UCYN-A sublineages have been defined based on UCYN-A nitrogenase (*nifH*) sequences. Due to the low abundances of UCYN-A in the global oceans, currently existing molecular techniques are limited for detecting and quantifying these organisms. A targeted approach is needed to adequately characterize the diversity of this important marine cyanobacterium, and to advance understanding of its ecological importance. We present findings on the distribution of UCYN-A sublineages based on high throughput sequencing of UCYN-A *nifH* PCR amplicons from 78 samples distributed throughout many major oceanic provinces. These UCYN-A *nifH* fragments were used to define oligotypes, alternative taxonomic units defined by nucleotide positions with high variability. The dataset was dominated by a single oligotype associated with the UCYN-A1 sublineage, consistent with previous observations of relatively high abundances in tropical and subtropical regions. However, this analysis also revealed for the first time the widespread distribution of the UCYN-A3 sublineage in oligotrophic waters. Furthermore, distinct assemblages of UCYN-A oligotypes were found in oligotrophic and coastally-influenced waters. This unique dataset provides a framework for determining the environmental controls on UCYN-A distributions and the ecological importance of the different sublineages.

Key index words: *Candidatus* Atelocyanobacterium thalassa, nitrogen fixation, nitrogenase, *nifH*, oligotyping, UCYN-A

Abbreviations: UCYN-A, unicellular cyanobacterial group A; N, Nitrogen; N_2, dinitrogen; *nifH*, nitrogenase; next generation sequencing (NGS); North Pacific Subtropical Gyre (NPSG); California Current System (CCS)

Primary productivity in vast regions of the global ocean is limited by the availability of nitrogen (N) (Gruber and Sarmiento 1997, Karl et al. 1997). Organisms that are capable of fixing dinitrogen (N_2) gas into reduced N, termed diazotrophs, play a critical role in providing new N to oligotrophic oceanic regions. In sunlit surface waters of the marine environment, a diverse assemblage of cyanobacteria that carry out N_2 fixation include *Trichodesmium* spp., diatom-associated *Richelia* strains, and unicellular cyanobacteria such as *Crocosphaera* spp., *Cyanothece* spp., and the uncultivated unicellular cyanobacterial group A (UCYN-A) (see review by Zehr 2011). Originally described from partial *nifH* fragments amplified from the North Pacific (Zehr et al. 1998), UCYN-A has now been detected in all ocean basins and is known to be an important contributor to N_2 fixation in some regions of the North Pacific Subtropical Gyre (Church et al. 2009), and the eastern basin of the Tropical North Atlantic (Montoya et al. 2004, Goebel et al. 2010, Turk et al. 2011, Martinez-Perez et al. 2016).

The emerging picture of UCYN-A diversity has thus far been based on the sporadic detection of UCYN-A resulting from diazotroph diversity surveys. As of May 2015, over a decade after the discovery of UCYN-A, less than 1000 UCYN-A *nifH* sequences had been deposited to the National Center for Biotechnology Information (NCBI) Genbank database. Studies that have used next generation sequencing (NGS) technologies on *nifH* gene fragments amplified using degenerate *nifH* primers have also recently reported UCYN-A sequences (Farnelid et al. 2011,

Bentzon-Tilia et al. 2015, Messer, Doubell et al. 2015, Messer, Mahaffey et al. 2015, Turk-Kubo et al. 2015, Xiao et al. 2015, Doblin et al. 2016, Farnelid et al. 2016). The greater depth of sequence coverage in these studies, compared to traditional clone libraries, may have favored its recovery in unexpected environments such as the Danish Strait (Bentzon-Tilia et al. 2015) and an inverse hypersaline estuary in the South Australian Bight (Messer, Doubell et al. 2015).

Although originally thought to be an organism with low genetic diversity (Tripp et al. 2010), the UCYN-A lineage is now known to be comprised of at least four main sublineages, UCYN-A1, UCYN-A2, UCYN-A3, and UCYN-A4, as defined using nitrogenase (*nifH*) phylogeny of nucleotide sequences (Thompson et al. 2014, Farnelid et al. 2016). Sublineages of marine microorganisms may occupy different ecological niches and their functions may be shaped by environmental factors (*Prochlorococcus* for example; Kent et al. 2016); however, our understanding of the distribution of these UCYN-A sublineages is limited. UCYN-A1 and UCYN-A2 both have greatly reduced genomes and live symbiotically with genetically distinct prymnesiophyte hosts (Tripp et al. 2010, Thompson et al. 2012, Bombar et al. 2014, Thompson et al. 2014). However, very little is known about the UCYN-A3 and UCYN-A4 sublineages beyond where their *nifH* gene sequences have been identified, but there is evidence that all four sublineages may be widely distributed throughout the global oceans (Farnelid et al. 2016). The presence of these sublineages has been overlooked partially due to rare recovery of UCYN-A *nifH* sequences in marine diazotroph diversity studies and the high similarity between all sublineages in amino acid sequences of the highly conserved *nifH* gene (Thompson et al. 2014). Potentially significant differences in sequence identity are also often obscured by phylogenetic analyses that rely on clustering at similarity thresholds. Many open questions remain about these

interesting marine symbioses including the identity of host cells for UCYN-A3 and UCYN-A4 sublineages, the fidelity of host-symbiont associations, and whether additional sublineages are yet to be discovered and described.

Recent advances in next generation sequencing have provided access to unprecedented amounts of sequence data from the ocean ecosystem. The availability of complete UCYN-A1 and nearly complete UCYN-A2 genomes have greatly advanced our ability to visualize this association and to detect UCYN-A in metagenomes and metatranscriptomes (Cabello et al. 2016, Cornejo-Castillo et al. 2016), as well as 16S rRNA amplicon libraries (Martinez-Perez et al. 2016). Nearly full UCYN-A1 and UCYN-A2 genomes have now been assembled from metagenomes obtained from the South Atlantic as part of the recent TARA oceans expedition (Cornejo-Castillo et al. 2016). Furthermore, Cornejo-Castillo et al. (2016) detected the active transcription of several key metabolic genes, including *nifH*, in both sublineages. Recent observations of differences in morphologies (Zehr 2015, Cornejo-Castillo et al. 2016) and cell-specific N_2 fixation rates for the UCYN-A1 and UCYN-A2 sublineages (Martinez-Perez et al. 2016) suggests that different sublineages likely have different impacts on nutrient cycling in the marine environment. Despite these advances, the presence and activity of additional sublineages will remain difficult to detect until genomes from other sublineages are available, due to the low relative abundances of these organisms in a complex microbial ecosystem. Therefore, even with increasing amounts of data available from next generation sequencing studies of nitrogenase diversity, the emerging picture of sublineage biogeography remains patchy.

It can be challenging to reveal ecologically significant patterns in genera of marine microorganisms that appear similar (or even indistinguishable) using conventional molecular markers (*Crocosphaera* for example; Bench et al. 2013, Bench et al. 2016). An emerging approach, oligotyping, provides an alternative to defining operational taxonomic units based on clustering or phylogenetic analyses by defining "oligotypes", highly refined taxonomic units based on nucleotide positions with high variability, or Shannon entropy (Eren et al. 2013). This approach has proven informative at distinguishing potential ecotypes of closely related organisms, such as SAR11 (Eren et al. 2013) and revealing differential responses of *Prochlorococcus* and *Synechococcus* oligotypes to nutrient amendments (Shilova et al. submitted), based on 16S rRNA gene fragments. This is a promising method for investigating UCYN-A diversity considering that differences between sublineages are defined by variability in the wobble positions along their *nifH* gene sequences (Thompson et al. 2014).

In order to investigate the diversity, distribution and ecological significance of UCYN-A sublineages, we defined UCYN-A oligotypes after generating a large dataset of UCYN-A *nifH* gene fragments (hereafter referred to as the UCYN-A *nifH* amplicon dataset). We selected samples from the North and South Atlantic, the North and South Pacific, and the Danish Strait (Table 1) to be screened for the presence of UCYN-A. DNA was extracted using a DNeasy-based protocol described in detail by (Moisander et al. 2008). DNA extracts from the Danish Strait and Sargasso Sea were obtained using protocols described in Bentzon-Tilia et al. (2015) and Farnelid et al. (2011), respectively. We used a nested PCR approach with the first amplification using universal *nifH* primers nifH3/nifH4 (Zehr and Turner 2001). The second amplification used UCYN-A-specific *nifH* primers with 5' common sequence linkers univ_UCYN-A_F_CS1 (5'-

ACA CTG ACG ACA TGG TTC TAC AAG TTT GCA YTG TAA AGC ACA -3') and univ_UCYN-A_R_CS2 (5'- TAC GGT AGC AGA GAC TTG GTC TTC CTT CAC GGA TAG GCA TAG -3'). Reaction conditions and thermocycling parameters are described in Supporting Information Appendix S1. Universal UCYN-A *nifH* primers were designed to target all known UCYN-A sequences deposited to NCBI's Genbank nr/nt database as of March 2015 using Primer 3 (Untergasser et al. 2012). UCYN-A *nifH* fragments were amplified from 78 samples (from a total of 369 samples screened; Figure 1). Libraries were prepared using the dual PCR approach (Green et al. 2015) and sequenced using Illumina MiSeq technology at the DNA Service Facility at the University of Chicago, Illinois.

A total number of 3,078,383 raw paired-end UCYN-A *nifH* reads were obtained. Raw sequence files are archived at NCBI's Sequence Read Archives (SRA) under BioSample Accession numbers SAMN05776250- SAMN05776327. Read counts per sample had high variability, ranging from 22-104,445, presumably reflecting the number of UCYN-A *nifH* gene copies present in each sample. Raw sequences were merged using Paired-End reAd mergeR (PEAR) software (Zhang et al. 2014). Scripts from the Quantitative Insights into Microbial Ecology (QIIME) pipeline (Caporaso, Kuczynski et al. 2010) were used to filter raw sequences for quality, remove chimeric sequences (UCHIME; Edgar et al. 2011), and to determine unique sequences using usearch 6.1 (Edgar 2010). If a sequence was recovered more than 10 times, its representative sequence was imported into ARB (Ludwig et al. 2004), where poor quality (e.g. containing stop codons) and non-*nifH* sequences were removed. Sequences that passed all quality filtering steps (2,044,530 out of 3,078,383) had primer regions trimmed in Galaxy (Afgan et al. 2016), were aligned to a reference alignment available for UCYN-A sequences in a curated

185*nifH* database (Heller et al. 2014) using PyNAST (Caporaso, Bittinger et al. 2010), and prepared
186for oligotyping using custom python and R scripts.
187
188Shannon entropy analysis and oligotyping were performed using the oligotyping pipeline version
1892.2 described by Eren et al. (2013), and 13 positions were selected to define UCYN-A
190oligotypes. Positions with greatest entropy were exclusively wobble bases. Oligotyping analysis
191was carried out using arguments that; 1) identified thirteen positions with greatest entropy (-c
192102, 93, 75, 99, 78, 192, 48, 213, 231, 147, 42, 150, 210 ; Supporting Information Fig. S1); 2)
193allowed for a given oligotype to be present in only one sample (-s 1); 3) required that a given
194oligotype be present at a relative abundance of at least 0.1% in one sample (-a 0.1); and 4)
195required that the most abundant unique sequence defining an oligotype had a sequence count >
196100 across the whole dataset (-M 100). This analysis defined 44 unique oligotypes, which
197represented 99.67% of the sequences submitted for analysis, with a total purity score of 0.87. All
198but one oligotype, oligo7, met the criteria of “convergence” (Eren et al. 2013).
199
200Maximum likelihood trees of representative sequences for each oligotype were calculated in
201MEGA 6 (Tamura et al. 2013) based on the Tamura-Nei model, and node supports were
202determined with 1000 bootstrap replicates. Of the 44 oligotypes, 17 had 100% nucleotide
203similarity to sequences submitted to NCBI’s Genbank database (See sequences with asterisks (*)
204in Fig. 2). UCYN-A oligotype distribution data was analyzed and visualized using the R package
205Phyloseq (McMurdie and Holmes 2013). Data was subsampled using the following criteria: 1)
206removing samples with low (<1000) sequence counts (64/78 samples remained); and 2)
207removing oligotypes that had fewer than 100 total sequences (30/44 oligotypes remained

distributed across 64 samples). Ecological distances between samples was determined using Jaccard and Bray-Curtis ecological indices on both subsampled data and subsampled data rarefied to equal sampling depth. Principal coordinate analysis (PCoA) was performed on the resulting distance matrices, to visualize the dissimilarity between samples and UCYN-A sublineages.

UCYN-A *nifH* sequences from prior studies, compiled as part of a recent review by (Farnelid et al. 2016), were used to explore how well defined oligotypes described the diversity in an independent dataset. This dataset, hereafter referred to as the NGS dataset, was prepared for oligotyping and analyzed in Phyloseq as described above.

The UCYN-A *nifH* amplicon dataset was primarily comprised of four major oligotypes - oligo1, oligo2, oligo3, and oligo4 – which together accounted for 95.9% of all sequences recovered. The remainder of the dataset was comprised of minor oligotypes, oligo5-oligo44, present at low relative abundances across the dataset. Oligo1, which includes the UCYN-A1 genome-derived *nifH* sequence (Zehr et al. 2008, Tripp et al. 2010), dominated the UCYN-A *nifH* amplicon dataset (Fig. 2). In 63 out of the 78 total samples analyzed, oligo1 accounted for over 70% of the sequences recovered (Fig. 3a, Supporting Information Table S1). A majority (19/40) of the minor oligotypes were also phylogenetically affiliated with UCYN-A1 (Fig. 2). The wide distribution of the UCYN-A1 sublineage has been well documented. Its early recovery in clone library-based studies (Zehr et al. 1998, Langlois et al. 2005) led to the design of quantitative PCR-based assays (Church et al. 2005, Langlois et al. 2008) that have since been widely applied in every major ocean basin (e.g. Church et al. 2008, Langlois et al. 2008, Moisander et al. 2008, Bonnet et al.

2312009, Goebel et al. 2010, Bonnet et al. 2015). The high recovery of UCYN-A1-affiliated
232oligotypes was an anticipated result, and is consistent with UCYN-A1 abundances that are
233commonly reported to range between 10^4-10^6 *nifH* copies L^{-1}, and can sometimes be as high as
23410^7 *nifH* copies L^{-1} (Mulholland et al. 2012).
235
236The second most abundant oligotype, oligo2, which differs from oligo1 in 10 of the 13 entropy
237positions (See Supporting Information Fig. S1), clusters with the UCYN-A3 sublineage defined
238by Thompson et al. (2014). Oligo2 was found widely distributed in 55 of the 78 samples
239analyzed, but was recovered in much lower relative abundances than oligo1; on average, oligo2
240comprised 7.4% ± 11.0% of the relative abundances across the dataset, while oligo1 comprised
24178.6% ± 29.3%. In the North Pacific Eddy samples, oligo2 had higher relative abundances at
242mid depths in the water column (30-70 m), and the UCYN-A *nifH* amplicon dataset was
243dominated by surface samples (0-25 m). Hence, if this oligotype resides deeper in the water
244column, relative abundances in this dataset may be underestimated. The highest relative
245abundances for oligo2 were consistently found in Sargasso Sea samples, accounting for between
24622%- 57% of the sequences, but sequence recovery from these samples was generally low
247(Supporting Information Table S1). A total of 7 of the defined oligotypes are phylogenetically
248affiliated to UCYN-A3 (Fig. 2). Very little is known about the UCYN-A3 sublineage, but *nifH*
249sequences have been sporadically reported from different regions, including the Tropical North
250Atlantic (Wheeler, direct submission to Genbank), the South Pacific gyre (Halm et al. 2012), the
251Western South Pacific (Messer, Mahaffey et al. 2015) and the South Australian Bight (Messer,
252Doubell et al. 2015).
253

Oligo3 includes the *nifH* sequence derived from the UCYN-A2 genome (Bombar et al. 2014) and is the third most abundant oligotype. Oligo3 differs from oligo1 in 8 of the 13 entropy positions (See Supporting Information Fig. S1). It was detected in 24 out of 78 samples at relative abundances >0.1%, and the few samples that were dominated by oligo3 (>50% relative abundance) were exclusively found in coastally-influenced waters in the Danish Strait and California Current System (CCS; Fig. 3; Supporting Information Table S1). It has been speculated that the UCYN-A2 sublineage may be a coastally adapted strain (Thompson et al. 2014, Messer, Doubell et al. 2015). However, there have been recent reports that UCYN-A2 is globally distributed and may play a major role in N cycling in both oligotrophic regions and in temperate, high latitude waters (Cabello et al. 2016, Martinez-Perez et al. 2016). Findings from our study do not directly contradict Cabello et al. (2016) and Martinez-Perez et al. (2016). However, the low relative abundance and patchy detection of oligo3 along with the much higher relative abundances of the UCYN-A3 oligotype oligo2 in oligotrophic samples, implies that UCYN-A2 may not be a major sublineage in open ocean regions. Recovery of a *nifH* sequence is currently the only way to confidently determine whether a particular sublineage is present in a given sample. For sublineages other than UCYN-A1 and UCYN-A2, there is currently nothing known about the host cell. 16S rRNA gene sequences are not available, and qPCR assays that differentiate between subclades are not available (Farnelid et al. 2016). Therefore, it is unclear whether studies reporting UCYN-A2 based on distribution data of its *B. bigelowii* host (Cabello et al. 2016) or 16S rRNA gene sequences (Martinez-Perez et al. 2016) are accurately reporting the presence of this sublineage.

Oligo4, the fourth most abundant oligotype, differs from oligo1 in 8 out of 13 entropy positions, clusters with the newly defined UCYN-A4 sublineage (Farnelid et al. 2016), and is one of only two oligotypes associated with this sublineage (Fig. 2). Oligo4 was mainly found in the Danish Strait, at relative abundances as high as 83%. Relative abundances were highest in 0.2-10 μm size fraction samples, but sequences were also found in the 10 μm size fraction. Oligo4 was also found at relative abundances >0.1% in one other sample, Station ALOHA (NP.ALOHA.5.2.HOT5m241), and sequences within the UCYN-A4 sublineage have been reported in the coastal Japan Sea (Accession numbers LC013598, LC013602, LC013603, LC013607; Shiozaki et al. 2015).

The two minor UCYN-A1 oligotypes which are found at high relative abundances in different ocean regions, each differ from the dominant UCYN-A1 oligotype by single entropy positions. One of these minor oligotypes, oligo5, is 100% similar to sequences that have been deposited in Genbank (KF546346.1, KC013065.1, EU187536.1). Oligo5 sequences were present in mid-depth (35-75 m) samples at high relative abundances (up to 25%) at a station situated in an anticyclonic eddy in the North Pacific Subtropical Gyre (NPSG) (NPacEddy.74 samples; Fig. 3A and Supporting Information Table S1). It was also recovered at much lower relative abundances in other NPSG samples taken at Station ALOHA as well as samples from the Coral Sea (Fig. 3A and Supporting Information Table S1). In contrast, a second minor oligotype, oligo6, was present only in four stations in the South Atlantic at high relative abundances (up to ca. 15%; Fig. 3A). It is not yet clear what the ecological relevance of these UCYN-A1 oligotypes may be, yet it is striking that they seem to occupy different regions. A strong seasonal succession between two closely related SAR11 strains (that differ by 2 nucleotides across the V4-V6 16S rRNA gene

region) was revealed using the oligotyping approach (Eren et al. 2013), and with a higher resolution dataset (temporally and/or spatially) changes in UCYN-A oligotypes may reveal similar relationships to environmental parameters.

To evaluate how well the defined oligotypes describe known UCYN-A diversity, the same 13 entropy positions were used to define UCYN-A oligotypes from the NGS dataset (Farnelid et al. 2016) containing *nifH* amplicons from the South Australian Bight (Messer, Doubell et al. 2015), the Arufura and Coral Seas (Messer, Mahaffey et al. 2015), the Danish Strait (Bentzon-Tilia et al. 2015) and the Noumea Lagoon of New Caledonia (Turk-Kubo et al. 2015). The NGS dataset is overwhelmingly comprised of UCYN-A sequences from the Noumea Lagoon (>95% of the 168,022 UCYN-A sequences; Supporting Information Table S2). The resulting purity score (Eren et al. 2013), 0.73, indicates that these 13 entropy positions are well chosen to represent known UCYN-A diversity. The NGS dataset was dominated by three oligotypes, oligo3 (UCYN-A2), oligo43 (UCYN-A2), and oligo1 (UCYN-A1). Oligo3, the same UCYN-A2 oligotype that dominated samples from the Danish Strait and CCS in the UCYN-A *nifH* amplicon dataset, accounted for 57.8% of all sequences (and 59.3% of Noumea Lagoon sequences; Supporting Information Fig. S2). Intriguingly, oligo43, which was a minor oligotype in the UCYN-A *nifH* amplicon dataset, was the second most abundant oligotype recovered in the NGS dataset (22.2%). Differing by the dominant UCYN-A2 oligotype (oligo3) in 8 out of the 13 entropy positions, oligo43 was found exclusively in the Noumea Lagoon samples. In contrast to the UCYN-A *nifH* amplicon dataset, oligo1 only comprised 13.8% of the sequences in the NGS dataset, reflecting the higher relative abundances of UCYN-A2 sublineages from the Noumea Lagoon samples. A total of 7 new oligotypes were defined in the NGS dataset that affiliated

mainly with UCYN-A1 and UCYN-A2 sublineages in the Noumea Lagoon. One of the new oligotypes (oligo48), which was present at low relative abundances, does not cluster with defined lineages (Supporting information Fig. S2).

A clear distinction between UCYN-A populations found in coastally-influenced and oligotrophic regions was revealed based on PCoA on the ecological distance between samples from the UCYN-A *nifH* amplicon dataset. This was observed using both unweighted (Jaccard) and weighted (Bray-Curtis) ecological indices, in all coordinate axes, and using both subsampled data as well as data rarefied to equal sampling depth (Fig. 4A and Supporting Information Fig. S3A). This pattern appears to be driven by a consistent co-occurrence of UCYN-A1 and UCYN-A3 in all oligotrophic samples compared to an occurrence of UCYN-A2, sometimes in the presence of UCYN-A4, in coastally-influenced samples (Fig. 4B and Supporting Information Fig. S3B). Ordination analysis on the NGS dataset also supports co-occurrence of UCYN-A1 with UCYN-A3, as well as the clustering of coastal and oligotrophic samples (Fig. 4D). In this dataset, however, co-occurrence of both UCYN-A1 (oligo1) and UCYN-A2 (oligo3 and oligo43) oligotypes in the Noumea Lagoon samples is clearly seen.

This is the first report of the widespread distribution of the UCYN-A3 sublineage, as well as its co-occurrence with UCYN-A1. These findings indicate that this sublineage lives in an environment now known to be favorable to unicellular diazotrophs, the warm (>20^{o}C), sunlit tropical and subtropical waters of the oligotrophic ocean gyres. In the South Atlantic, this sublineage is present at low abundances, ranging between $7.8x10^{1}$ - $9.2x10^{2}$ *nifH* copies L^{-1}, which is several orders of magnitude less than UCYN-A1 ($3.8x10^{4}$ – $2.0x10^{5}$ *nifH* copies L^{-1};

Supporting Information Fig. S4, and Supporting Information Appendix S2). However, it may be misleading to infer the potential contribution to N_2 fixation rates based on *nifH*-based abundances alone for these symbiotic diazotrophs. An alternate sublineage, currently assumed to be UCYN-A2, fixes N_2 at much higher cell-specific rates than UCYN-A1 in the North Atlantic (Martinez-Perez et al. 2016). Thus, despite being present at lower cellular abundances, its contribution to bulk N_2 fixation rates is similar to UCYN-A1 in this region. Until cell-specific N_2 fixation rates for the UCYN-A3 sublineage are known, its relative contribution to N_2 fixation rates in the oligotrophic gyres remains an open question.

To further test the co-occurrence of UCYN-A1 and UCYN-A3, TARA ocean metagenomic samples from the open-ocean station 78 in the South Atlantic, where Cornejo-Castillo et al. (2016) recruited high percentages of both UCYN-A1 and UCYN-A2 genomes, were screened for the presence of *nifH* oligotypes (See Supporting Appendix S3). Despite the high recruitment of reads to UCYN-A genomes in station 78 samples, only about 100 total *nifH* reads were found that spanned at least half of the fragment used in the oligotyping analysis and they all affiliated with UCYN-A1 and UCYN-A2 oligotypes. It is not possible to conclude that UCYN-A3 was completely absent, but it is clear that UCYN-A1 and UCYN-A2 sublineages were found co-occurring at this station. UCYN-A1 and UCYN-A2 sublineages have been observed to co-occur in coastally influenced regions, such as the North American Mid-Atlantic coastal shelf (Mulholland et al. 2012), the Santa Monica Basin (Hamersley et al. 2011), the Spencer Gulf in the South Australian Bight (Messer, Doubell et al. 2015), coastal Japan (Shiozaki et al. 2015), and the New Caledonia lagoon (Turk-Kubo et al. 2015). Further research is needed to determine whether these sites represent overlapping niches for UCYN-A1 and UCYN-A2 sublineages or

whether the observed UCYN-A diversity in these regions results from mixing oligotrophic and coastal waters. Indeed, it has been suggested that diazotrophs can be transported on currents for large distances while remaining active (Shiozaki et al. 2013).

The results do, however, strongly suggest that the UCYN-A2 sublineage may be more commonly found in coastally-influenced ecosystems. The dominant UCYN-A2 oligotype, oligo3, was found at high relative abundances in geographically distant temperate environments, in the brackish waters of the Danish Strait (DS.GB samples), the CCS transition zone (CCS.6.53198, CCS.7.53199) and in the elevated salinity waters of the Spencer Gulf (Messer, Doubell et al. 2015). The UCYN-A2/*B. bigelowii* association found year-round at the SIO pier is known to be larger than the UCYN-A1 association described in the NPSG (5-10 μm and 2-3 μm, respectively) and with more UCYN-A cells per host cell (Thompson et al. 2014, Zehr 2015, Cornejo-Castillo et al. 2016). These observations are consistent with theories that larger phytoplankton will be found at higher abundances in eutrophic conditions (Irwin et al. 2006).

Conclusions

This study provides unique insights into the global distribution of UCYN-A sublineages with co-occurrences of UCYN-A1/UCYN-A3 observed in open ocean waters and the presence of UCYN-A2, sometimes co-occurring with UCYN-A4, observed in coastal waters. Currently the UCYN-A3 sublineage is known only by its *nifH* gene sequence, and many open questions remain about the evolutionary relationship to UCYN-A1 and UCYN-A2, the identity of its host, and cell-specific N_2 fixation rates, in addition to morphological traits such as the number of symbionts per host. This is the first study that reports the widespread distribution of the UCYN-

A3 sublineage. The emerging ecological niche for UCYN-A3 appears to be in tropical/sub-tropical oligotrophic surface waters throughout the Pacific and Atlantic, and intriguingly, it was always found in the presence of UCYN-A1.

Despite the clear co-occurrence of UCYN-A1/UCYN-A3 and UCYN-A2/UCYN-A4 in these samples, UCYN-A1 and UCYN-A2 are known to co-occur, mainly in coastally-influenced regions, including the Noumea Lagoon in New Caledonia, and the TARA station 78 in the South Atlantic. More high resolution temporal and/or spatial data (both lateral and depth), with parallel measurements of environmental data is needed in these regions to better characterize patterns of co-occurrence, and begin to understand the environmental factors that influence UCYN-A sublineage distributions.

No UCYN-A sequences were recovered from the Eastern Tropical South Pacific (ETSP), a region where discrepancies between N_2 fixation rates and diazotroph abundances led to the speculation that cyanobacterial phylotypes, like UCYN-A, could be difficult to detect (Turk-Kubo et al. 2014). It is unlikely that the paradox of N_2 fixation in this region can be attributed to undetected UCYN-A sublineages. However, the targeted UCYN-A *nifH* PCR assay described in this study has the promise to help identify regions where UCYN-A sublineages have been overlooked using qPCR and metagenomics/metatranscriptomics.

Applying an oligotyping approach to UCYN-A *nifH* fragments, provides a new, standardized framework for characterizing sublineage distributions. Even though a vast majority of the recovered sequences in this dataset as a whole were attributed to a single UCYN-A1 oligotype,

distinct, sample-specific differences in sublineage distributions were discovered. Furthermore, the presence of oligotypes not affiliated with defined UCYN-A lineages also hints to a greater diversity yet to be discovered.

The approach also uncovered several minor oligotypes, that appeared to be potentially significant members of the UCYN-A community in distinct regions, a finding that would have been missed using a cluster-based approach. It remains an open question what the broader genomic diversity may be between distinct oligotypes from the same UCYN-A sublineage, and even within a single oligotype, as well as whether these oligotypes are associated with the same prymnesiophyte hosts, and have similar morphological characteristics.

Acknowledgements

We gratefully acknowledge Ryan Paerl, Julie Robidart, Sam Wilson, Andy Rees, and Sophie Bonnet for collecting samples used in this study, Ed Boring (UCSC) for bioinformatics assistance, and Dr. Stefan Green and his staff at the DNA Services Facility and the University of Illinois, Chicago for next generation sequencing consultation. We would also like to thank Dr. Lasse Riemann (University of Copenhagen) for providing us with DNA extracts from previous research projects in the Sargasso Sea and Danish Strait, and the Hawaii Ocean Time Series crew and staff. This study uses CTD/sea-surface data from the Atlantic Meridional Transect Consortium (NER/0/5/2001/00680), provided by the British Oceanographic Data Centre and supported by the Natural Environment Research Council. This work was supported in part by a grant from the Simons Foundation (SCOPE Award ID 329108, J.Z.), and awards from NSF's

Dimensions in Biodiversity (Award #1241221) and Biological Oceanography (Award #1559165). HF is supported by the Swedish Research Council VR 637-2013-7502.

Conflict of Interest

The authors declare no conflict of interest.

References

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C. & Grüning, B.. 2016. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res.*, 44:W3-W10.

Bench, S. R., Frank, I., Robidart, J. & Zehr, J. P. 2016. Two subpopulations of *Crocosphaera watsonii* have distinct distributions in the North and South Pacific. *Environ. Microbiol.*, 18: 514-524.

Bench, S. R., Heller, P., Frank, I. E. & Arciniega, M. 2013. Whole genome comparison of six *Crocosphaera watsonii* strains with differing phenotypes. *J. Phycology,* 49:786-801.

Bentzon-Tilia, M., Traving, S. J., Mantikci, M., Knudsen-Leerbeck, H., Hansen, J. L. S., Markager, S. & Riemann, L. 2015. Significant N_2 fixation by heterotrophs, photoheterotrophs and heterocystous cyanobacteria in two temperate estuaries. *ISME J.*, 9:273-285.

Bombar, D., Heller, P., Sanchez-Baracaldo, P., Carter, B. J. & Zehr, J. P. 2014. Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria. *ISME J.*, 8:2530-2542.

Bonnet, S., Biegala, I. C., Dutrieux, P., Slemons, L. O. & Capone, D. G. 2009. Nitrogen fixation in the western equatorial Pacific: Rates, diazotrophic cyanobacterial size class distribution, and biogeochemical significance. *Global Biogeochem. Cy.,* 23:GB3012.

Bonnet, S., Rodier, M., Turk-Kubo, K.A., Germineaud, C., Menkes, C., Ganachaud, A., Cravatte, S., Raimbault, P., Campbell, E., Quéroué, F. & Sarthou, G. 2015. Contrasted geographical distribution of N_2 fixation rates and *nifH* phylotypes in the Coral and Solomon Seas (southwestern Pacific) during austral winter conditions. *Global Biogeochem. Cy.,* 29:1874-1892.

Cabello, A.M., Cornejo-Castillo, F.M., Raho, N., Blasco, D., Vidal, M., Audic, S., De Vargas, C., Latasa, M., Acinas, S.G. & Massana, R. 2016. Global distribution and vertical patterns of a prymnesiophyte-cyanobacteria obligate symbiosis. *ISME J.,* 10:693-706.

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L. & Knight, R. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics,* 26:266-267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G., Goodrich, J.K., Gordon, J.I. & Huttley, G.A. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods,* 7:1548-7091.

Church, M. J., Bjorkman, K. M., Karl, D. M., Saito, M. A. & Zehr, J. P. 2008. Regional distributions of nitrogen-fixing bacteria in the Pacific Ocean. *Limnol. Oceanog.,* 53:63-77.

Church, M. J., Jenkins, B. D., Karl, D. M. & Zehr, J. P. 2005. Vertical distributions of nitrogen-fixing phylotypes at Stn ALOHA in the oligotrophic North Pacific Ocean. *Aquat. Microb. Ecol.,* 38:3-14.

Church, M. J., Mahaffey, C., Letelier, R. M., Lukas, R., Zehr, J. P. & Karl, D. M. 2009. Physical forcing of nitrogen fixation and diazotroph community structure in the North Pacific subtropical gyre. *Global Biogeochem. Cy.,* 23:GB2020.

Cornejo-Castillo, F.M., Cabello, A.M., Salazar, G., Sánchez-Baracaldo, P., Lima-Mendez, G., Hingamp, P., Alberti, A., Sunagawa, S., Bork, P., De Vargas, C. & Raes, J. 2016. Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nature Communication,* 7:11071.

Doblin, M.A., Petrou, K., Sinutok, S., Seymour, J.R., Messer, L.F., Brown, M.V., Norman, L., Everett, J.D., McInnes, A.S., Ralph, P.J. & Thompson, P.A. 2016. Nutrient uplift in a cyclonic eddy increases diversity, primary productivity and iron demand of microbial communities relative to a western boundary current. *PeerJ,* 4:e1973.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics,* 26:1367-4803.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics,* 27:2194-2200.

Eren, A. M., Maignien, L., Sul, W. J., Murphy, L. G., Grim, S. L., Morrison, H. G. & Sogin, M. L. 2013. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution,* 4:1111-1119.

Farnelid, H., Andersson, A.F., Bertilsson, S., Al-Soud, W.A., Hansen, L.H., Sørensen, S., Steward, G.F., Hagström, Å. and Riemann, L. 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., Muñoz-Marin, M. & Zehr, J. 2016. New insights into the ecology of the globally significant uncultured nitrogen-fixing symbiont UCYN-A. *Aquat. Microb. Ecol.,* 77:135-138.

Goebel, N.L., Turk, K.A., Achilles, K.M., Paerl, R., Hewson, I., Morrison, A.E., Montoya, J.P., Edwards, C.A. & Zehr, J.P. 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.,* 12:3272-3289.

Green, S. J., 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.

Gruber, N. & Sarmiento, J. L. 1997. Global patterns of marine nitrogen fixation and denitrification. *Global Biogeochem. Cy.,* 11:235-266.

Halm, H., Lam, P., Ferdelman, T.G., Lavik, G., Dittmar, T., LaRoche, J., D'Hondt, S. & Kuypers, M.M. 2012. Heterotrophic organisms dominate nitrogen fixation in the South Pacific Gyre. *ISME J.*, 6:1238-1249.

Hamersley, M. R., Turk, K. A., Leinweber, A., Gruber, N., Zehr, J. P., Gunderson, T. & Capone, D. G. 2011. Nitrogen fixation within the water column associated with two hypoxic basins in the Southern California Bight. *Aquat. Microb. Ecol.,* 63:193-205.

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

Irwin, A. J., Finkel, Z. V., Schofield, O. M. & Falkowski, P. G. 2006. Scaling-up from nutrient physiology to the size-structure of phytoplankton communities. *J. Plankton Res.,* 28:459-471.

Karl, D., Letelier, R., Tupas, L., Dore, J., Christian, J. & Hebel, D. 1997. The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature,* 388:533-538.

Kent, A. G., Dupont, C. L., Yooseph, S. & Martiny, A. C. 2016. Global biogeography of *Prochlorococcus* genome diversity in the surface ocean. *ISME J.*, 10:1856-1865.

Langlois, R. J., Hummer, D. & LaRoche, J. 2008. Abundances and distributions of the dominant *nifH* phylotypes in the Northern Atlantic Ocean. *Appl. Environ. Microb.,* 74:1922-1931.

Langlois, R. J., LaRoche, J. & Raab, P. A. 2005. Diazotrophic diversity and distribution in the tropical and subtropical Atlantic Ocean. *Appl. Environ. Microb.,* 71:7910-7919.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Buchner, A., Lai, T., Steppi, S., Jobb, G., Förster, W., & Brettske, I., et al. 2004. ARB: a software environment for sequence data. *Nucleic Acids Res., 32*(4):1363-1371.

Martinez-Perez, C., Mohr, W., Löscher, C.R., Dekaezemacker, J., Littmann, S., Yilmaz, P., Lehnen, N., Fuchs, B.M., Lavik, G., Schmitz, R.A., LaRoche, J. & M.M.M. Kuypers. 2016. The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nature Microbiology*, 1: 16163.

McMurdie, P. J. & Holmes, S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE,* 8:e61217.

Messer, L. F., Doubell, M., Jeffries, T. C., Brown, M. V. & Seymour, J. R. 2015. Prokaryotic and diazotrophic population dynamics within a large oligotrophic inverse estuary. *Aquat. Microb. Ecol.,* 74:1-15.

Messer, L.F., Mahaffey, C., Robinson, C.M., Jeffries, T.C., Baker, K.G., Isaksson, J.B., Ostrowski, M., Doblin, M.A., Brown, M.V. & Seymour, J.R. 2015. High levels of heterogeneity in diazotroph diversity and activity within a putative hotspot for marine nitrogen fixation. *ISME J.,* 10:1499-1513.

Moisander, P. H., Beinart, R. A., Voss, M. & Zehr, J. P. 2008. Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *ISME J.,* 2:954-967.

Montoya, J. P., Holl, C. M., Zehr, J. P., Hansen, A., Villareal, T. A. & Capone, D. G. 2004. High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature,* 430:1027-1031.

Mulholland, M.R., Bernhardt, P.W., Blanco-Garcia, J.L., Mannino, A., Hyde, K., Mondragon, E., Turk, K., Moisander, P.H. & Zehr, J.P. 2012. Rates of dinitrogen fixation and the abundance of diazotrophs in North American coastal waters between Cape Hatteras and Georges Bank. *Limnol. Oceanogr.,* 57:1067-1083.

Shilova, I. N., Mills, M. M., Robidart, J. C., Turk-Kubo, K. A., Björkman, K.M., Kolber, Z., Church, M.J., Arrigo, K.R. & Zehr, J. P. Differential effects of nitrate, ammonium and urea as N sources for microbial communities in the North Pacific Ocean. Submitted to *Limnol. Oceanogr.*

Shiozaki, T., Kodama, T., Kitajima, S., Sato, M. & Furuya, K. 2013. Advective transport of diazotrophs and importance of their nitrogen fixation on new and primary production in the western Pacific warm pool. *Limnol. Oceanogr.* 58:49-60.

Shiozaki, T., Nagata, T., Ijichi, M. & Furuya, K. 2015. Nitrogen fixation and the diazotroph community in the temperate coastal region of the northwestern North Pacific. *Biogeosciences, 12*(15):1726-4170.

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

Thompson, A.W., Foster, R.A., Krupke, A., Carter, B.J., Musat, N., Vaulot, D., Kuypers, M.M. & Zehr, J.P. 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.A., Desany, B.A., Niazi, F., Affourtit, J.P. & Zehr, J.P. 2010. Metabolic streamlining in an open ocean nitrogen-fixing cyanobacterium. *Nature*, 464:90-94.

Turk, K.A., Rees, A.P., Zehr, J.P., Pereira, N., Swift, P., Shelley, R., Lohan, M., Woodward, E.M.S. & Gilbert, J. 2011. Nitrogen fixation and nitrogenase (*nifH*) expression in tropical waters of the eastern North Atlantic. *ISME J.*, 5:1201-1212.

Turk-Kubo, K. A., Frank, I. E., Hogan, M. E., Desnues, A., Bonnet, S. & Zehr, J. P. 2015. Diazotroph community succession during the VAHINE mesocosms experiment (New Caledonia Lagoon). *Biogeosciences*, 12:7435-7452.

Turk-Kubo, K. A., Karamchandani, M., Capone, D. G. & Zehr, J. P. 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. *Envion. Microbiol.,* 16:3095-3114.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. & Rozen, S. G. 2012. Primer3—new capabilities and interfaces. *Nucleic Acids Res.,* 40:e115-e115.

Xiao, P., Jiang, Y., Liu, Y., Tan, W., Li, W. & Li, R. 2015. Re-evaluation of the diversity and distribution of diazotrophs in the South China Sea by pyrosequencing the *nifH* gene. *Mar. Freshwater Res.,* 66:681-691.

Zehr, J. 2015. How single cells work together: Are single-celled symbioses organelle evolution in action? *Science*, 349:1163-1164.

Zehr, J., Mellon, M. & Zani, S. 1998. New nitrogen-fixing microorganisms detected in oligotrophic oceans by amplification of nitrogenase (*nifH*) genes. *Appl. Environ. Microb.,* 64:3444.

Zehr, J. P. 2011. Nitrogen fixation by marine cyanobacteria. *Trends Microbiol.,* 19:162-173.

Zehr, J.P., Bench, S.R., Carter, B.J., Hewson, I., Niazi, F., Shi, T., Tripp, H.J. & Affourtit, J.P. 2008. Globally distributed uncultivated oceanic N_2-fixing cyanobacteria lack oxygenic photosystem II. *Science, 322*:1110-1112.

Zehr, J. P. & Turner, P. J. 2001. Nitrogen fixation: nitrogenase genes and gene expression. In J. H. Paul (Ed.), *Methods in Microbiology: Marine Microbiology*, Vol. 30, pp. 271-286. London: Academic Press, Ltd.

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

614Table 1: Sample sets screened for the presence of UCYN-A. Region names are used in Figure
6153A-B, and sample names prefixes are used in Fig. 2. psu – practical salinity units, unk. –
616unknown.

617

Region	Region Name	Sample prefix	cruise(s) or sampling description	No. samples screened	No. samples + UCYN-A *nifH*	depth ranges (m)	temp. range (°C)	sal. Range (psu)	Ref. for original diazotroph diversity study
California Current System	CCS	CCS	Controlled, Agile, and Novel Observing Network (CANON) Initiative	53	8	0 - 80	13.0-13.1	33.2-33.5	*This study*
Roskilde Fjord	Danish Strait	DS.RF	Danish Marine monitoring program	7	5	0	2.5-9	unk.	Bentzon-Tilia et al. 2015
Great Belt Strait	Danish Strait	DS.GB	Danish Marine monitoring program	6	6	0	0.4-17	11-17	Bentzon-Tilia et al. 2015
Sargasso Sea	Sarg	Sarg	R/V Atlantic Explorer cruise X0804	61	4	1-200	18.5-26.6	unk.	Farnelid et al. 2011
North Pacific	NPac	NPac	Nutrient Effects on Marine microOrganisms (NEMO)	30	17	5-115	20.3-25.5	33.6-35.6	*This study*
North Pacific Subtropical Gyre Eddy	NPacEddy	NPacEddy	HOE-Legacy 2 (KM1215)	5	5	35-100	22.9-26.6	35.3-35.5	*This study*
North Pacific, Station ALOHA	NPac	NP.ALOHA	Hawaii Ocean Time Series (HOT) 240 & 241	3	3	5-70	22.8-23.5	35.1-35.3	*This study*
North Atlantic	NAtl	NAtl	Atlantic Meridional Transect 19 & 20	31	2	0	27.3[a]	36.6[a]	*This study*
South Atlantic	SAtl	SAtl	Atlantic Meridional Transect 19 & 20	22	10	0-6	20.9-25.3[b]	36.0-37.2[b]	*This study*
Coral Sea	CoralSea	CoralSea	Bifurcation	18	18	5-80	22.2-26.0	34.7-35.4	Bonnet et al. 2015
Monterey Bay	---	---	Monterey Bay Time Series (MBTS)	104	0	0-30	---	---	*This study*

53
54

Eastern Tropical South Pacific	---	---	AT1561	29	0	5-145	---	---	Turk-Kubo et al. 2014

[a] Data available for only 1/2 samples

[b] Data available for only 5/10 samples

Figure 1. Map of UCYN-A *nifH* amplicon dataset sample sites. Region abbreviations and basic environmental parameters are detailed in Table 1.

Figure 2. Maximum likelihood (ML) tree of partial *nifH* nucleotide sequences (248 positions) containing representative sequences of each defined UCYN-A oligotype. The ML tree was calculated in MEGA 6 (Tamura et al. 2013) based on the Tamura-Nei model, and node support was determined with 1000 bootstrap replicates. Oligotypes with representative sequences that have 100% nucleotide similarity to sequences submitted to NCBI's Genbank database are marked with an asterisk (*). Sequence counts for each oligotype, integrated across the whole dataset, are displayed in the barplot at the right. UCYN-A sublineages are defined as in Thompson et al. (2014) and Farnelid et al. (2016), and two potentially new sublineages are identified, UCYN-A5 and UCYN-A6.

Figure 3: Global distribution of UCYN-A oligotypes. A) The number of sequences and the distribution of UCYN-A oligotypes, colored by sublineage, after subsampling as described in the text. B) The relative abundance of oligotypes oligo4 (UCYN-A4), oligo5 (UCYN-A1), and oligo6 (UCYN-A1) make up a large proportion of UCYN-A sequences in distinct regions. CCS –

California Current system; NPSG – North Pacific Subtropical Gyre; NAtl – North Atlantic; SAtl – South Atlantic.

Figure 4: Principal Coordinate Analysis (PCoA) using the Jaccard ecological index to determine dissimilarity between samples for both UCYN-A *nifH* amplicon (A-B) and NGS datasets (C-D). Both datasets were transformed to equal sampling depth after subsampling as described in the text. In the UCYN-A *nifH* amplicon dataset, coastal (triangle) and oligotrophic (square) samples form distinct and separate clusters (A), and strong co-occurrence patterns are seen between UCYN-A1/UCYN-A3 and UCYN-A2/UCYN-A4 (B). Ordination analysis on the NGS dataset support distinct differences between oligotrophic and coastal samples (C) and support the co-occurrence of UCYN-A1/UCYN-A3 (D). Similar clustering is found in Axis.2 vs Axis.3 and Axis.1 vs Axis.3, and using the Bray-Curtis ecological index for both datasets (See Supporting Information Fig. S3). Region names displayed in (A) are detailed in Table 1. Region names displayed in (C) are: ArafuraCoralSea – South Pacific; Great Belt and Roskilde Fjord – Danish Strait; SouthAust – South Australian Bight; NewCaledonia – Noumea Lagoon mesocosms.

Table S1A. Environmental data for UCYN-A *nifH* amplicon dataset samples. Psu – practical salinity units; ddm – decimal degree minutes; unk – unknown; na – not applicable; coast – coastal; open – oligotrophic open ocean.

Table S2. UCYN-A oligotype distributions for each study included in NGS dataset, compiled as part of a recent review by Farnelid et al. 2016. The number of samples included from a given study are indicated at the head of each column.

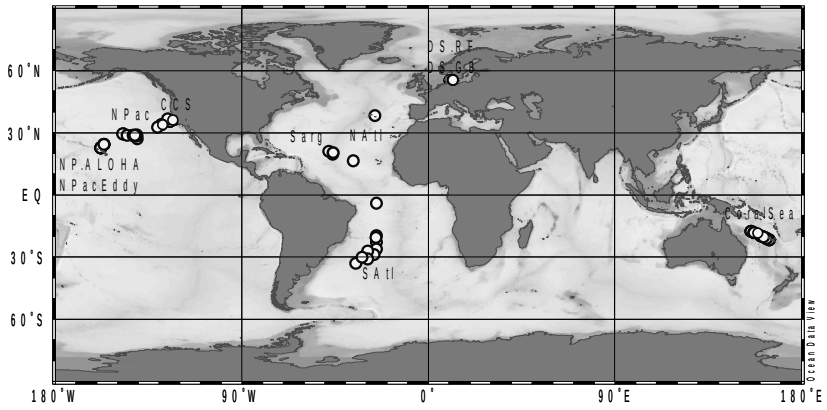
Figure S1. Shannon entropy analysis displaying positions with highest entropy across the entire UCYN-A *nifH* amplicon dataset. Representative sequences for the 6 most abundance oligotypes are overlaid on the Shannon entropy plot. Modified from output files from the oligotyping pipeline (merenlab.org/software/oligotyping/; Eren et al. 2013).

Figure S2. Counts of UCYN-A oligotype sequences from the NGS dataset.

Figure S3. Principal Coordinate Analysis (PCoA) using the Bray-Curtis ecological index to determine dissimilarity between samples for both UCYN-A *nifH* amplicon (A-B) and NGS datasets (C-D). Both datasets were transformed to equal sampling depth after subsampling as described in the text. In the UCYN-A *nifH* amplicon dataset, coastal (triangle) and oligotrophic (square) samples form distinct and separate clusters (A), and strong co-occurrence patterns are seen between UCYN-A1/UCYN-A3 and UCYN-A2/UCYN-A4 (B). Ordination analysis on the NGS dataset support distinct differences between oligotrophic and coastal samples (C) and support the co-occurrence of UCYN-A1/UCYN-A3 (D). Similar clustering is found in Axis.2 vs Axis.3 and Axis.1 vs Axis.3. Region names displayed in (A) are defined in Table 1 in the main text. Region names displayed in (A) are detailed in Table 1. Region names displayed in (C) are: ArafuraCoralSea – South Pacific; Great Belt and Roskilde Fjord – Danish Strait; SouthAust – South Australian Bight; NewCaledonia – Noumea Lagoon mesocosms.

Figure S4. UCYN-A1 and UCYN-A2/UCYN-A3 *nifH*-based abundances in the Atlantic during Atlantic Meridional Transect (AMT) cruises AMT-19 (2009) and AMT-20 (2010). Samples that were included in the UCYN-A *nifH* amplicon dataset are indicated with arrows.

Figure 1



685

686

687

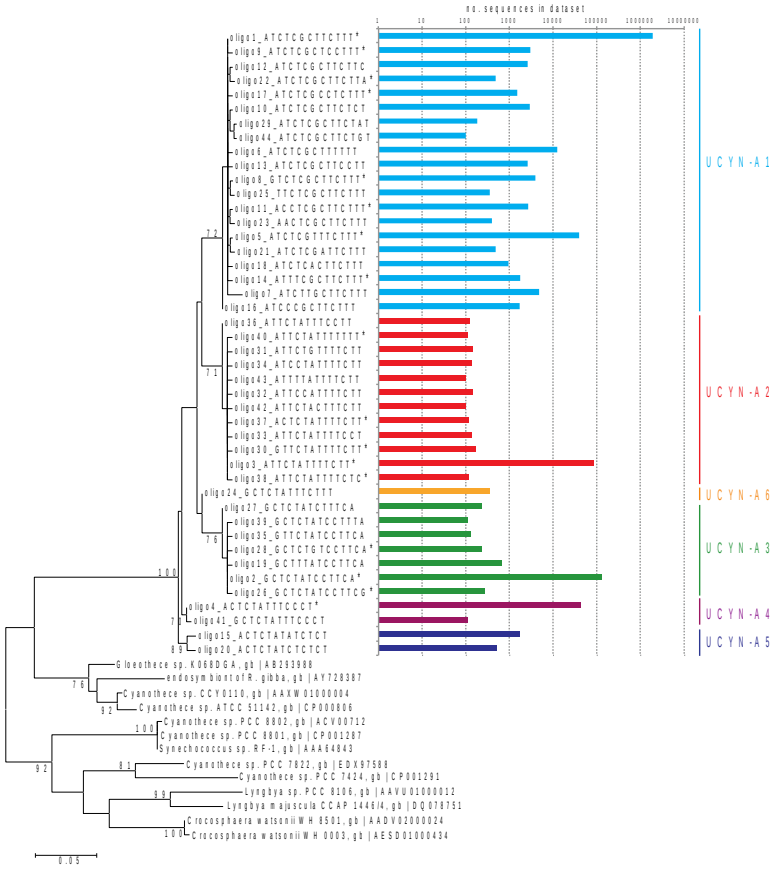
688

689

690

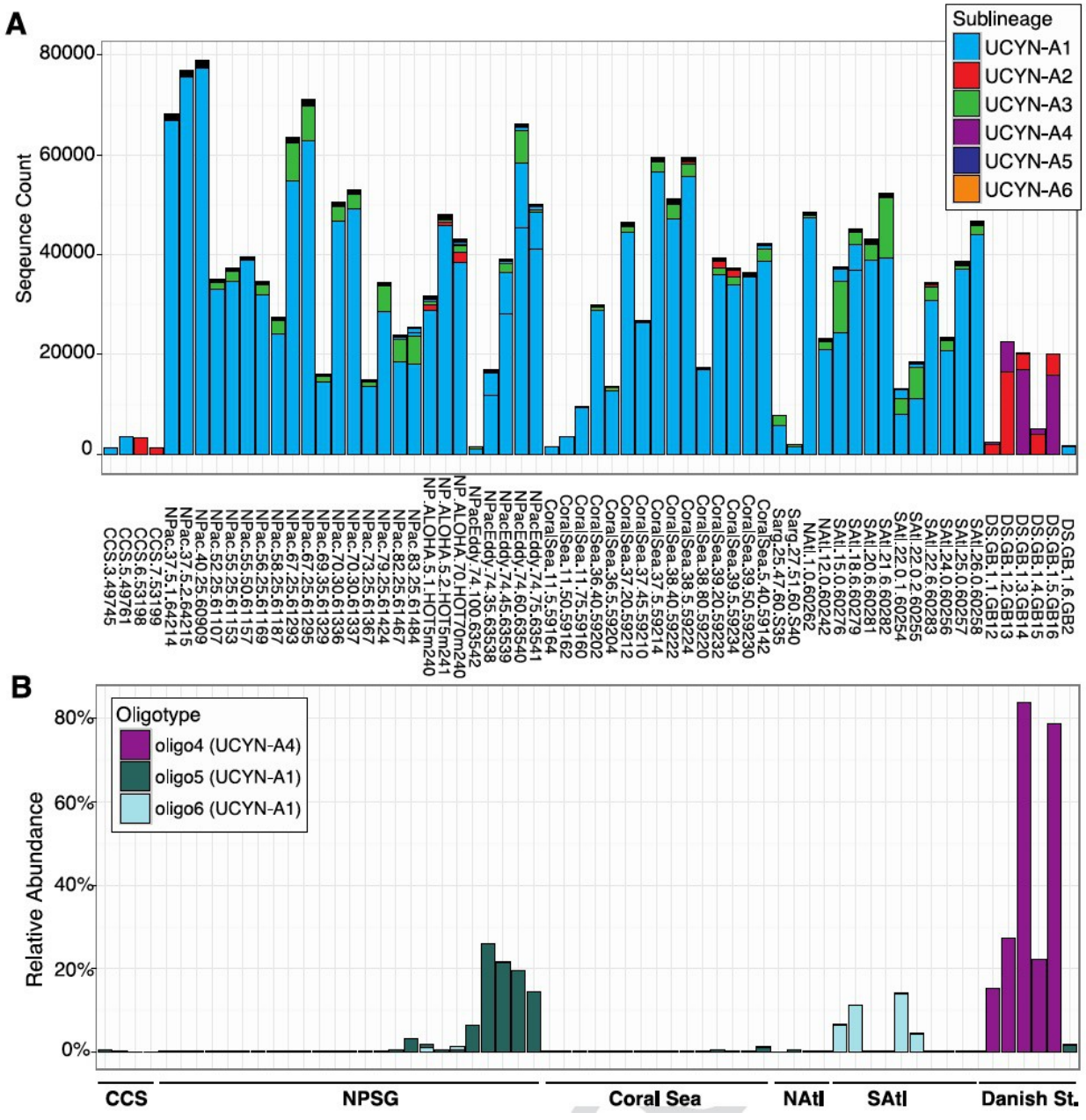
691

Figure 2

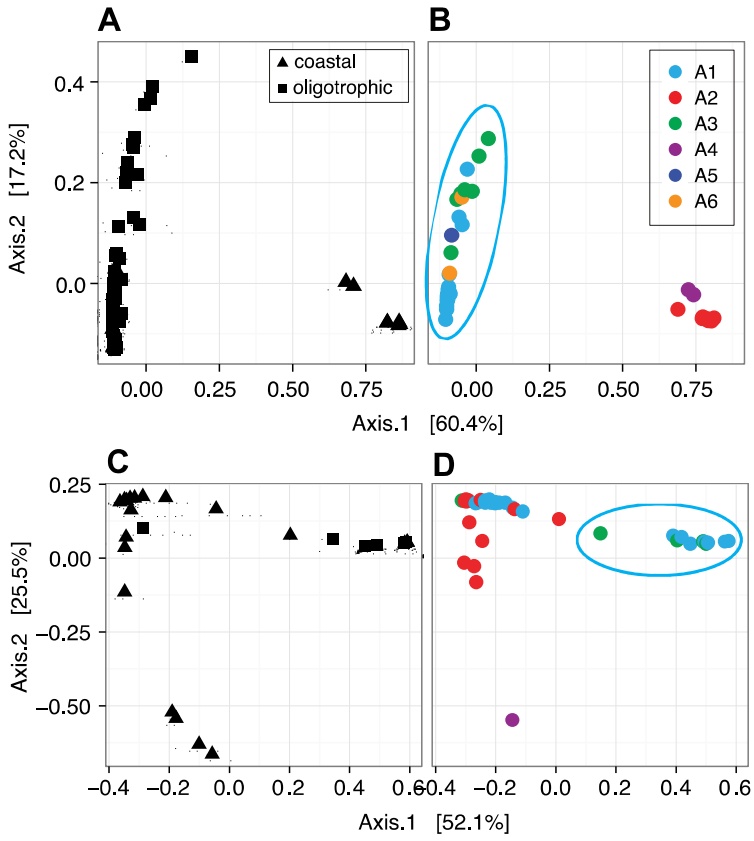


692

67
68


A
Sublineage
UCYN-A1
UCYN-A2
UCYN-A3
UCYN-A4
UCYN-A5
UCYN-A6
Seqeunce Count
80000
60000
40000
20000
0
CCS.3.49745
CCS.5.49761
CCS.6.53198
CCS.7.53199
NPac.37.5.1.64214
NPac.37.5.2.64215
NPac.40.25.60909
NPac.52.25.61107
NPac.55.25.61153
NPac.55.50.61157
NPac.56.25.61169
NPac.58.25.61187
NPac.67.25.61293
NPac.67.25.61295
NPac.69.35.61329
NPac.70.30.61336
NPac.70.30.61337
NPac.73.25.61367
NPac.79.25.61424
NPac.82.25.61467
NPac.83.25.61484
NP.ALOHA.5.1.HOT5m241
NP.ALOHA.5.2.HOT5m241
NP.ALOHA.70.HOT70m240
NPacEddy.74.100.63542
NPacEddy.74.35.63538
NPacEddy.74.45.63539
NPacEddy.74.60.63540
NPacEddy.74.75.63541
CoralSea.11.5.59164
CoralSea.11.50.59162
CoralSea.11.75.59160
CoralSea.36.40.59202
CoralSea.36.5.59204
CoralSea.37.20.59212
CoralSea.37.45.59210
CoralSea.38.40.59222
CoralSea.38.5.59224
CoralSea.38.80.59220
CoralSea.39.20.59232
CoralSea.39.5.59234
CoralSea.5.40.59142
Sarg.25.47.60.S35
Sarg.27.51.60.S40
NAtl.1.0.60262
NAtl.12.0.60242
SAtl.15.0.60276
SAtl.18.6.60279
SAtl.20.6.60281
SAtl.21.6.60282
SAtl.22.0.1.60254
SAtl.22.0.2.60255
SAtl.22.6.60283
SAtl.24.0.60256
SAtl.25.0.60257
SAtl.26.0.60258
DS.GB.1.1.GB12
DS.GB.1.2.GB13
DS.GB.1.3.GB14
DS.GB.1.4.GB15
DS.GB.1.5.GB16
DS.GB.1.6.GB2
B
Oligotype
oligo4 (UCYN-A4)
oligo5 (UCYN-A1)
oligo6 (UCYN-A1)
Relative Abundance
80%
60%
40%
20%
0%
CCS
NPSG
Coral Sea
NAtl
SAtl
Danish St.

693


A
B
C
D
coastal
oligotrophic
A1
A2
A3
A4
A5
A6
Axis.2 [17.2%]
Axis.1 [60.4%]
Axis.2 [25.5%]
Axis.1 [52.1%]

694

71
72