Comparative genomics reveals surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria

Permalink
https://escholarship.org/uc/item/7th6b8rt

Journal
ISME Journal, 8(12)

ISSN
1751-7362

Authors
Bombar, D
Heller, P
Sanchez-Baracaldo, P
et al.

Publication Date
2014

DOI
10.1038/ismej.2014.167

Peer reviewed
Comparative genomics reveal surprising divergence of two closely related strains of uncultivated UCYN-A cyanobacteria

Deniz Bombar, Philip Heller, Patricia Sanchez-Baracaldo, Brandon J. Carter, Jonathan P. Zehr

Running title: UCYN-A genome comparison

1Ocean Sciences Department, University of California Santa Cruz, 1156 High Street, CA 95064, USA

2Biomolecular Engineering Department, University of California Santa Cruz, 1156 High Street, CA 95064, USA

3Schools of Biological and Geographical Sciences, University of Bristol, UK

*these authors contributed equally to this study

Corresponding author: Jonathan P. Zehr, Ocean Sciences Department, University of California Santa Cruz, 1156 High Street, CA 95064, USA, Tel.: +1 831 459 3128 Fax: +1 831 459 4882. Email: jpzehr@gmail.com
Abstract

Marine planktonic cyanobacteria capable of fixing molecular nitrogen (termed “diazotrophs”) are key in biogeochemical cycling, and the nitrogen fixed is one of the major external sources of nitrogen to the open ocean. Candidatus Atelocyanobacterium thalassa (UCYN-A) is a diazotrophic cyanobacterium known for its widespread geographic distribution in tropical and subtropical oligotrophic oceans, unusually reduced genome, and symbiosis with a single-celled Prymnesiophyte alga. Recently a novel strain of this organism was also detected in coastal waters sampled from the Scripps Institute of Oceanography (SIO) pier. We analyzed the metagenome of this UCYN-A2 population by concentrating cells by flow cytometry. Phylogenomic analysis provided strong bootstrap support for the monophyly of UCYN-A (here called UCYN-A1) and UCYN-A2 within the marine Crocosphaera sp. and Cyanothece sp. clade. UCYN-A2 shares 1159 of the 1200 UCYN-A1 protein coding genes (96.6%) with high synteny, yet the average amino acid sequence identity between these orthologs is only 86%. UCYN-A2 lacks the same major pathways and proteins that are absent in UCYN-A1, suggesting that both strains can be grouped at the same functional and ecological level. Our results suggest that UCYN-A1 and UCYN-A2 had a common ancestor and diverged after genome reduction. These two variants may reflect adaptation of the host to different niches, which could be coastal and open ocean habitats.

Key words: cyanobacteria/genome reduction/Nitrogen fixation/symbiosis/marine

Subject category: • Evolutionary genetics
Introduction

Marine pelagic cyanobacteria play a major role in biogeochemical cycling of carbon and nitrogen (N) in the ocean. *Prochlorococcus* and *Synechococcus* together are the most abundant phototrophic prokaryotes on Earth, and are responsible for a major fraction of oceanic carbon fixation (Partensky et al., 1999; Scanlan and West, 2002; Scanlan, 2003; Johnson et al., 2006). Likewise, cyanobacteria capable of fixing molecular N (“diazotrophs”) dominate global oceanic $\text{N}_2$ fixation, although they are typically orders of magnitude less abundant than *Prochlorococcus* or *Synechococcus* (Zehr and Paerl, 2008; Zehr and Kudela, 2011; Voss et al., 2013). Together with upward fluxes of deep-water $\text{NO}_3^-$ to the surface ocean, diazotrophs supply the N requirement of primary productivity and quantitatively balance losses by sinking of organic material, which can sequester $\text{CO}_2$ from the atmosphere to deep waters (Karl et al., 1997; Sohm et al., 2011).

There are several groups of quantitatively significant diazotrophic cyanobacteria in the open ocean, all of which thrive mainly in tropical and subtropical latitudes (Stal, 2009). Traditionally, the filamentous, aggregate-forming cyanobacterium *Trichodesmium* sp. was viewed as the most important oceanic $\text{N}_2$ fixer, based on its wide distribution and direct measurements of its $\text{N}_2$ fixation capacity (Dugdale et al., 1961; Capone et al., 1997; Bergman et al., 2013). Other diazotroph cyanobacteria discovered in early microscopic studies are the filamentous heterocyst-forming types of the *Richelia* and *Calothrix* lineages, which live in symbioses with several different diatom species (Villareal, 1992; Janson et al., 1999; Foster and Zehr, 2006). More recently, molecular approaches resulted in the discovery of unexpected and unusual cyanobacteria involved in oceanic $\text{N}_2$ fixation (Zehr et al., 1998; 2001). These have usually been grouped as “unicellular” diazotrophic
cyanobacteria, but among them, different types have very different lifestyles, with *Crocosphaera watsonii* being a photosynthetic and mostly free-living cell (but see Foster et al. (2011)), while UCYN-A (*Candidatus* Atelocyanobacterium thalassa) is a photoheterotroph that is symbiotic with prymnesiophyte algae (Thompson et al., 2012). While the major biogeochemical role of all diazotrophic cyanobacteria is to provide new N to the system, their different lifestyles suggest important differences regarding their distribution in the ocean, and the fate of the fixed N and carbon (Glibert and Bronk, 1994; Scharek et al., 1999; Mulholland, 2007).

As a diazotrophic cyanobacterium, UCYN-A (termed UCYN-A1 from here on) is remarkable in several ways. Although somewhat closely related to *Cyanothece* sp. strain ATCC 51142, the UCYN-A1 genome is only 1.44 Mb and lacks many genes including whole metabolic pathways and proteins, such as the oxygen-evolving photosystem II and Rubisco, i.e. features that normally define cyanobacteria (Tripp et al., 2010). The recent identification of a symbiotic eukaryotic prymnesiophyte partner, to which UCYN-A1 provides fixed N while receiving carbon in return, is the first known example of a symbiosis between a cyanobacterium and a prymnesiophyte alga (Thompson et al., 2012). Further, UCYN-A1 can be detected in colder and deeper waters compared to other major N₂ fixers like *Trichodesmium* sp. and *Crocosphaera watsonii* (Needoba et al., 2007; Langlois et al., 2008; Rees et al., 2009; Moisander et al., 2010; Diez et al., 2012), and is also abundant in some coastal waters (Mulholland et al., 2012).

There is now evidence that there are at least three *nifH* lineages of UCYN-A in the ocean (Thompson et al., 2014). These different clades were previously unrecognized because their *nifH* amino acid sequences are nearly identical, with sequence variation...
primarily only occurring in the third base pair of each codon (Thompson et al., 2014). It is unknown whether these strains represent different metabolic variants of UCYN-A, analogous to observations in free-living cyanobacteria like *Prochlorococcus* and *Synechococcus*, which have extensive heterogeneity in their genome contents that enable them to occupy different niches along gradients of nutrients and light (Moore et al., 1998; Ahlgren et al., 2006; Kettler et al., 2007). Phylotype “UCYN-A2” shares only 95% *nifH* nucleotide similarity with UCYN-A1, and was discovered to be abundant and actively expressing *nifH* off of the Scripps Institute of Oceanography (SIO) pier. This habitat seems to generally lack UCYN-A1 and has environmental conditions that clearly differ from the tropical/subtropical oligotrophic open-ocean during large parts of the year (Chavez et al., 2002). UCYN-A2 is associated with a prymnesiophyte host that is closely related to but not identical to the UCYN-A1 host (Thompson et al., 2014). Interestingly, the known 18S rRNA gene sequences of the UCYN-A2 host generally fall into a ‘coastal’ cluster while the UCYN-A1 host sequences almost exclusively cluster with sequences recovered from open ocean environments (Thompson et al., 2014). Further, both UCYN-A1 and its host appear significantly smaller than UCYN-A2 and its host (Thompson et al., 2014). Based on these findings, Thompson et al. suggested that UCYN-A1 could be an oligotrophic open ocean ecotype, whereas UCYN-A2 could possibly be more adapted to coastal waters.

The present study represents the first opportunity to characterize the metabolic potential of a new clade of UCYN-A, by analyzing the metagenome of a UCYN-A2 population sampled from waters off the SIO pier. This enabled us to test whether habitat differences, or a distinct symbiont-host relationship, are reflected in genome features that
distinguish UCYN-A2 from UCYN-A1, and whether UCYN-A2 has the same lack of
genes as UCYN-A1. With the availability of the new UCYN-A2 metagenome, it was also
possible to perform phylogenomic analyses (including 135 proteins), to determine
whether UCYN-A2 and UCYN-A1 form a monophyletic group, and to establish how
these two organisms are related to other cyanobacteria.

Material & Methods

Sampling

After the initial detection of a new nifH phylotype similar to UCYN-A1 in coastal
waters off Scripps Pier and its classification as a new strain (UCYN-A2, Thompson et al.,
2014), we used the previously described cell-sorting approach (Zehr et al., 2008;
Thompson et al., 2012) to obtain cell sorts enriched in UCYN-A2 for genome
sequencing. Surface water samples (10L) were taken at Scripps Pier with a bucket, gently
poured into a polypropylene bottle, and immediately transferred to the laboratory at
Scripps. The sample was then concentrated by gentle vacuum filtration through a 0.22
µm pore size polycarbonate filter and cells resuspended by vortexing the filter in 50 mL
of sterile-filtered seawater. The concentrate was flash-frozen in liquid N and shipped to
UCSC.

Fluorescence activated cell sorting (FACS) and nifH quantitative PCR and
genome amplification
The concentrated seawater samples were thawed at room temperature and briefly vortexed again immediately prior to cell sorting. Seawater samples were pre-filtered using 50 μm mesh size CellTrics filters (Partec, Swedesboro, NJ, USA) to prevent clogging of the nozzle (70 μm-diameter) with large particles. Samples were analyzed in logarithmic mode with an Influx Cell Sorter (BD Biosciences, San Jose, CA, USA). Flow cytometry sorting gates were defined using forward scatter (FSC, a proxy for cell size) and chlorophyll fluorescence at 692 nm (Fig. 1). Chlorophyll autofluorescence was excited using a 200 mW, 488 nm Sapphire laser (Coherent, Santa Clara, CA, USA).

A UCYN-A2-specific QPCR assay (Thompson et al., 2014) was used to screen sorted events within each gate (between 100-200 events). Cells were sorted directly into aliquots of 10 uL 5 kDa filtered nuclease-free water, and then amended with QPCR 1x Universal PCR master mix (Applied Biosystems, Foster City, CA, USA) to a total reaction volume of 25 uL, including UCYN-A2 specific forward- and reverse primers (0.4 μM final concentration), as well as TaqMan probes (0.2 μM final concentration). QPCR reactions were conducted in a 7500 Real-Time PCR instrument (Applied Biosystems, Foster City, CA, USA). Reaction- and thermal cycling conditions were carried out as described previously (Thompson et al., 2014; Moisander et al., 2010).

Abundances of nifH gene copies were quantified relative to standard curves comprised of amplification of linearized plasmids containing inserts of the target nifH gene, and abundances of gene copies per sample calculated as described by Short and Zehr (2005). Standards were made from serial dilutions of plasmids in nuclease free water (range: 1-10^3 nifH gene copies per reaction), with 2 μL of each dilution added up to 25 μL qPCR
(total volume) mixtures. Duplicates of each standard were included with each set of samples run on the qPCR instrument, as well as at least two no-template controls.

Using this approach, we detected a sort region relatively enriched in UCYN-A2 but still containing other organisms besides the target (Fig. 1). This region appears to include single UCYN-A2 cells rather than populations in the picoeukaryote size fraction as described in Thompson et al. (2014) (Fig. 1). The disruption of the UCYN-A symbiotic association appears to be a typical result of the concentration- and freezing protocol (Thompson et al., 2012), and proved advantageous for our genome amplification and assembly. A sample taken on May 31, 2011 was used to obtain a cell sort enriched in UCYN-A2 for genome amplification (Fig. 1). Approximately 3.5*10^4 events were sorted into a 1.5 mL microcentrifuge tube containing 90 µL of TE buffer. Cells were pelleted at 14,000 rpm (21,000 x g) for approximately 45 min and the supernatant was discarded.

We used a Qiagen REPLI-g Midi kit for cell lysis and amplification of genomic DNA, following the manufacturer's recommendations with few modifications. Briefly, the pelleted cells were resuspended in 3.5 µL PBS buffer and 3.5 µL buffer D2 (0.09M DTT), incubated at 65°C for 5 min, and immediately stored on ice after adding the kit-provided “stop buffer”. The amplification reaction was carried out in a thermal cycler at 30 ºC for 6 hours after addition of 40 µL Repli G mastermix to the tube. The quality, size and quantity of the amplified DNA was checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, USA) and again quantified using Pico Green (Invitrogen Corp., Carlsbad, USA). The suitability of this sample for a genome-sequencing run was indicated by the presence of 10^6 nifH gene copies of UCYN-A2 per µL, measured by QPCR.
Illumina sequencing

Library preparation and paired-end sequencing were performed at the BioMicro Center of the Massachusetts Institute of Technology (MIT, http://openwetware.org/wiki/BioMicroCenter:Sequencing). The DNA sample was split into two equal aliquots and prepared for sequencing using the SPRIworks system (Beckman Coulter Genomics, Danvers, USA) with 150-350 bp and 250-550 bp inserts. Ligated libraries were amplified and molecular barcodes added. Samples were pooled and sequenced on an Illumina MiSeq v1 flowcell with 151 bp of sequence read in each direction. Fastq files (illumina v1.5) were prepared and separated into the individual libraries allowing one mismatch with the barcode sequences. Post-run quality control includes confirmation of low sequencing error rates by analyzing phiX spike sequences, checking for significant contamination from human, mouse, yeast and E. coli, and confirming the presence of only the expected barcodes.

Please see the Supplemental Material section for a detailed description of sequence assembly, annotation, and phylogenomic analyses. This sequencing project has been deposited at DDBJ/EMBL/GenBank under the organism name “Candidatus Atelocyanobacterium thalassa isolate SIO64986”, accession number JPSP01000000.

Results

The aligned UCYN-A2 scaffolds to the UCYN-A1 reference chromosome covered nearly the entire UCYN-A1 sequence (Fig. 2). For the majority of the adjacent pairs of scaffolds, the last gene of the upstream scaffold and the first gene of the
downstream scaffold matched consecutive genes in the gene order of UCYN-A1 (30 cases), thereby conserving and extending the high synteny seen across the alignments. In the remaining cases, adjacent scaffold ends carried partial genes that matched different parts of the same gene in UCYN-A1 (43 partial genes in UCYN-A2 matching to 21 genes in UCYN-A1).

Overall, the UCYN-A2 draft genome is highly similar to UCYN-A1 in gene content, synteny, and basic genome features including GC content (31%), percent of coding DNA (79.3 %), codon usage (supplemental Fig. 4), and overall gene count including two rRNA operons (Fig. 2, Table 1). There is 99% 16S rRNA gene sequence identity between both genomes. Seven RNA genes in UCYN-A2 had very similar but un-annotated sequences in UCYN-A1 (91-100% nucleotide identity over 97-100% of the query sequence), and some annotated matching sequences exist in other cyanobacteria such as *Calothrix* sp. PCC7507 and *Cyanothece* sp. 8801, 8802, and 51142. These consist of one additional tRNA gene for methionine and six RNA genes annotated as non-coding RNA (ncRNA) with unknown functions (“other RNA genes” in Table 1).

A total of 1159 of the 1200 UCYN-A1 proteins (Tripp et al., 2010) have closely matching sequences in UCYN-A2, i.e. 96.6 % of UCYN-A1’s genes are shared with UCYN-A2. For these 1159 genes, the average amino acid sequence identity is 86.3% (range 51-100%, Fig. 2). The most conserved genes (≥ 95% identity) include housekeeping genes (ribosomal proteins, NADH dehydrogenase, ATP synthase), Photosystem I subunits, and proteins involved in N₂ fixation (*nif* cluster).

The previously described UCYN-A1 genome was unusual and had extensive genome reduction, lacking the genes encoding Photosystem II, Rubisco, biosynthesis
pathways for several amino acids and purines, as well as the TCA cycle and other key metabolic pathways (Zehr et al. 2008, Tripp et al. 2010). The genes missing in the UCYN-A1 genome were also absent in the UCYN-A2 draft genome. In addition to the analysis of all rejected contigs, we used TBLASTN to search the full set of unassembled sequencing reads for all 114 *Cyanothece* sp. 51142 genes reported missing in UCYN-A1 (Tripp et al. 2010), to test whether some of these genes might have escaped assembly. Subject reads were compared to GenBank using BLASTN against the nt database, and taxonomy was retrieved for the top 20 hits for each read. Matching reads were found for only 13 different genes out of these 114 query genes (18 total hits, incl. 5 PSII genes). Seven hits had 98-100% identity to known organisms (*Synechococcus*, *Pelagomonas*, *Thalassiosira pseudonana*), and four hits to an uncultured marine prokaryote. The remaining 7 hits had maximal identity ranging between 79% and 89% to sequences from other organisms (*Galdieria*, *Aureococcus*, *Acaryochloris*, *Flavobacterium*, *Nitrosomonas*, and *Monosiga*).

Apart from the 1159 genes shared by UCYN-A2, there are 41 UCYN-A1 genes (including 25 hypothetical proteins) that appear to be pseudogenes in UCYN-A2. These pseudogenes were either neighboring partial genes that aligned consecutively to a full ORF of a UCYN-A1 gene, with interrupting stop codons and/or insertions between them (a total of 21 partial genes in UCYN-A2 matching to 8 genes in UCYN-A1, not counting genes at scaffold ends; Table 2), or short, un-annotated sequences that match only parts of UCYN-A1 genes (remaining 33 UCYN-A1 genes). Although the evidence for pseudogenes was strong, as the UCYN-A2 sequences were from good assemblies that yielded high-coverage scaffolds, we additionally used PCR to amplify across nine
random examples of these pseudogenes, confirming that the interrupting stop codons were present and were not artifacts of assembly (see Supplemental Material for details). The genome comparison revealed that such pseudogenes also exist in UCYN-A1 (Table 2).

An interesting difference between both genomes is that for all UCYN-A1 genes, at least short, unannotated remnants or pseudogenes can be found in UCYN-A2, while in turn UCYN-A2 possesses 31 genes, of which 15 are hypothetical proteins, for which no traces (pseudogenes or gene remnants) were found in UCYN-A1, indicating that they have been completely lost from the genome (Table 2). The loss of these genes has in most cases resulted in further genome compaction in UCYN-A1, i.e. they appear fully excised instead of being replaced by non-coding DNA (examples shown in Fig. 3). The majority of these unique UCYN-A2 genes had top BLASTP similarity to genes in different Cyanothece sp. (16 genes) or in other Cyanobacteria (5 genes), while 10 short hypothetical proteins (27-63 amino acids) had no clear phylogenetic affiliation.

In addition to interrupted genes, we note 132 genes that show differences in amino acid length compared to orthologs in the other genome, i.e. they appear truncated at either the C- or N-terminal end of the protein. For UCYN-A2, this was also confirmed for a few examples by PCR amplification (Supplemental Material). Some of these truncated genes might be pseudogenes as well. Thirteen genes in UCYN-A1 and 14 genes in UCYN-A2 had less than 75% of the amino acids in the comparable protein sequence in the other strain. A comparison of the ortholog pairs of UCYN-A1 and UCYN-A2 to orthologs in Cyanothece sp. 51142 showed that the truncated versions of the genes almost exclusively occur in one of the UCYN-A strains, but not in Cyanothece sp. 51142, while
the gene length of the longer ortholog in UCYN-A1/A2 correlated well with the gene length in *Cyanothece* sp. 51142 (Fig. 4 A). Interestingly, UCYN-A1 generally possessed the shortest versions of the gene among these three genomes (Fig. 4 B).

Overall, both genomes show extremely similar genome reduction, but there are some differences regarding which genes have become pseudogenes, and UCYN-A1 appears to have a higher level of reduction, with fully excised genes at several loci and overall greater truncation of genes. Functions affected by gene deletions or pseudogenization differ for UCYN-A1 and UCYN-A2 (Table 2), with the latter genome e.g. retaining genes involved in cell wall synthesis, vitamin import, and detoxification of active oxygen species such as H$_2$O$_2$.

Maximum likelihood analyses confirmed that both UCYN-A strains belong to a well-supported monophyletic group of marine planktonic cyanobacteria containing *Crocosphaera* sp., *Cyanothece* sp. and other unicellular N$_2$ fixing cyanobacteria (Sanchez-Baracaldo et al, 2014). The results of the analyses strongly support that UCYN-A2 and UCYN-A1 form a monophyletic group that is a sister group to *Crocosphaera* sp. and *Cyanothece* sp. (Bootstrap support 100; Fig. 5). This clade of marine unicellular N$_2$ fixers belongs to the previously described SPM group (Sanchez-Baracaldo et al, 2005) containing *Synechocystis*, *Pleurocapsas*, and *Microcystis* (Fig. 5).

**Discussion**

UCYN-A is likely one of the major oceanic N$_2$ fixers given that it has a wider geographic distribution than *Trichodesmium* sp., diatom symbionts, or *Crocosphaera* sp., and can be highly abundant at certain times and places (Church et al., 2009; Moisander et
The symbiotic relationship of UCYN-A with a eukaryotic, possibly calcifying prymnesiophyte raises many important questions about the variability and regulation of \( \text{N}_2 \) fixation in UCYN-A, the fate of the fixed N (and C) in the planktonic food web, the role of UCYN-A in element export to the deep ocean, and its susceptibility to ocean acidification (Thompson et al., 2012). Further, the recently recognized \( nifH \) sequence diversity in the UCYN-A clade suggests that there could be different ecotypes of UCYN-A in the ocean, which could potentially be very different in terms of genome composition and physiology (Thompson et al., 2014). The genome comparison in this study addresses this question, with the surprising discovery that both types have very similar gene content, genome reduction, but also substantially divergent DNA sequences.

UCYN-A2 has very similar gene content to UCYN-A1 and also lacks photosystem II genes, RuBisCo, TCA cycle components and other pathways. It therefore represents a second, independently verified example of this kind of genome reduction in UCYN-A symbionts. Together with the highly conserved gene order, which implies gene function conservation, this suggests that UCYN-A1 and UCYN-A2 have similar functions and metabolic interactions in the symbiosis with their haptophyte hosts. Although it can be difficult to confirm that genes are missing in unclosed genomes, we base the claim on several independent lines of evidence: 1) Many scaffolds ended with partial genes that mapped to a single UCYN-A1 gene, or ended with full genes that matched and preserved the gene order in UCYN-A1, suggesting that breaks between scaffolds were not due to missing sequence. 2) Even though there is variability in genome sequence coverage (26.7 on average, supplemental Fig. 2), it is highly unlikely that there would be no coverage at all for the long stretches of target genome needed to
contain the many missing genes in UCYN-A1. 3) The rejected contigs had a GC content of 44.7% (very different from the 31% found in UCYN-A1 and UCYN-A2), sparse BLAST hits to UCYN-A1 or *Cyanotoce* sp., (even at a very relaxed e-value threshold), and any detected hits to UCYN-A1- or *Cyanotoce* sp.-like sequences were redundant with genes already present in the UCYN-A2 draft genome; this ascertains that no UCYN-A2 genes were missed. 4) Searching the sequence reads by TBLASTN for all 114 *Cyanotoce* sp. 51142 genes that appeared to be missing in UCYN-A1 returned only thirteen of the query genes, of which most had highest similarity values to different organisms. 5) Recently obtained field data show peaks in *nifH* expression of UCYN-A2 during daytime, closely matching the temporal patterns of *nifH* expression determined for UCYN-A1 in the open oligotrophic ocean around Hawaii (Church et al., 2005; Thompson et al., 2014). This may be viewed as further confirmation for the absence of oxygen-evolving PSII in UCYN-A2, given the oxygen sensitivity of the nitrogenase enzyme.

Each UCYN-A strain has only a handful of genes that are either absent or disrupted in the other genome (Table 2). The loss of genes in symbiont genomes is a gradual process, and highly reduced genomes characteristically exhibit slow gene loss in the form of erosion of individual genes or operons, rather than larger deletions via chromosomal rearrangements (Moran and Mira, 2001; Wernegreen et al., 2002; Moran, 2003). The pattern of lost, disrupted, or truncated genes seen in the UCYN-A strains examined here appears consistent with such slow gene decay.

Gene inactivation and loss in symbionts mainly occurs because genes become functionally redundant and therefore non-essential, e.g. due to metabolite exchange with the host. Many of the functions encoded by pseudogenes in UCYN-A1/A2 indeed appear
dispensable when considered in the context of the symbiont-host relationship, such as restriction endonucleases, pyrimidine synthesis, or cell motility (Table 2). However, the intact versions of those genes in the other genome, and the unique genes in UCYN-A2, raise the question whether they have been retained because their function is still important, or whether they are also non-essential/redundant but have so far escaped inactivation and elimination. Noteworthy examples are the genes involved in cell wall biogenesis and cell shape determination in UCYN-A2. The latter genes occur in rod-shaped cells and also in *Cyanothece* sp. 51142. These genes could indicate that UCYN-A2 has a different morphology than UCYN-A1, and could point to differences in how it is structurally associated with its host, which might also influence the fragility of the association. Interestingly, genes involved in cell wall biogenesis, which have become pseudogenes in UCYN-A1, are also among disrupted genes in the obligate cyanobacterial endosymbiont of the diatom *Rhopalodia gibba*, (Kneip et al. 2008). Another interesting case is the UCYN-A2 peroxidase gene 2528848519. Peroxidases act in detoxifying active oxygen species such as H$_2$O$_2$, e.g. the thioredoxin peroxidase in *Synechocystis* PCC6803 (68% nucleotide identity to UCYN-A2 gene) (Yamamoto et al., 1999). Active oxygen species are formed during respiration and photosynthesis, but also many other processes (Miyake and Yokota, 2000). The presence of a peroxidase could indicate that UCYN-A2 experiences higher intracellular oxygen concentrations than UCYN-A1. UCYN-A2 would then have to respire more oxygen in order to fix N$_2$, and in the process would generate more reactive oxygen species, thus potentially relying on this peroxidase gene. Based on searches in metagenomic and metatranscriptomic datasets, the UCYN-A1 genome was initially assumed to represent a global population with very similar
genome sequences (≥97% nucleotide sequence identity, Tripp et al. 2010), analogous to
the low sequence diversity seen in Crocosphaera watsonii (Zehr et al., 2007; Bench et al.,
2011). While the phylogenomic analysis strongly supports the two UCYN-A strains to be
sister species (Fig. 5), one of the striking results from our genome comparison is the
relatively large range of sequence similarity seen among shared genes in UCYN-A1 and
UCYN-A2 (Fig. 2). The combination of this sequence divergence with the extremely
high similarity in basic genome features, gene content, and synteny suggests that the
genome reduction occurred prior to the speciation event and genetic divergence. It is
therefore likely that the common ancestor of UCYN-A1 and UCYN-A2 was already a
symbiont. Vicariance might have triggered the genetic divergence in the course of
speciation of the prymnesiophyte host into strains that possibly are slightly better adapted
to different oceanic realms. This would have allowed the cyanobacterial genomes to
accumulate gene sequence mutations after driving forces causing large genome
rearrangements were no longer significant, which appears typical for symbiont genomes
that have already been highly reduced (Tamas et al., 2002; Moran, 2003; Silva et al.,
2003). Interestingly, genes involved in N₂ fixation were among the most conserved
orthologs, likely reflecting the importance of this process in maintaining the symbiosis,
since it arguably represents the function most beneficial to the host and which must have
been vital in the initial formation of the symbiotic relationship.

Small, conserved and highly syntenic genomes exhibiting high amino acid
divergence can also be found in the free-living heterotrophic SAR11 clade (Wilhelm et
al., 2007; Grote et al., 2012). SAR 11 is an example for genome reduction due to
“streamlining”, while the genome reduction seen in UCYN-A appears typical for
The amino acid divergence between the UCYN-A strains lies within the range seen in the SAR11 Ia cluster (which have 2% 16S rRNA divergence, Grote et al., 2012). However, UCYN-A1 and UCYN-A2 have even more conserved genome content than SAR11 Ia and are considerably more conserved than members of the cyanobacterial Prochlorococcus group (Kettler et al., 2007), which appears typical for obligate intracellular organisms (Grote et al., 2012). This evolutionary pattern is unusual and suggests that the genomes of these UCYN-A strains are under strong selection, since they are highly specialized symbionts of eukaryote algae.

Although nifH sequences of UCYN-A1 and UCYN-A2 can co-occur in some samples from around the world, the question has been raised whether these two different strains could be adapted to different nutrient regimes, and could therefore have overlapping but different distributions in the ocean (Thompson et al., 2014). However, we find no evidence in the genomes of UCYN-A1 and UCYN-A2 that would resemble genetic differentiation analogous to that in e.g. the high-light or low-light ecotypes of Prochlorococcus sp. (Moore et al., 1998; Kettler et al., 2007), or the ‘coastal’ ecotypes of Synechococcus sp. (Ahlgren and Rocap, 2006; Palenik et al., 2006). This lack of genetic differentiation, and the overall level of genome reduction, is characteristic for genomes of obligate symbionts with high dependency on their host (Moran, 2003; Hilton et al., 2013), and suggests that UCYN-A may not be directly exposed to, or affected by the external environment. Analyzing the genomes of the host algae and other UCYN-A strains will be necessary to identify genes that might represent adaptation to different environmental conditions.
While the two strains show no immediately apparent gene adaptations to cope with horizontal nutrient gradients or light quality, it is interesting that UCYN-A1 appears to be smaller than UCYN-A2 (Thompson et al., 2014), has fully excised genes compared to UCYN-A2 (Fig. 3) and greater truncation of genes (Fig. 4). The genomic signatures in UCYN-A point to typical genome reduction in a symbiont via genetic drift, a mechanism which is particularly enhanced under small effective population sizes (van Ham et al., 2003; Giovannoni et al., 2014). However, the further reduced genome of UCYN-A1 could also reflect an adaptation to the open ocean environment with very low levels of nutrients. Comparative genomics and ecological studies (Scanlan et al., 2009), as well as trait evolution analyses (Larsson et al., 2011), have shown a trend in genome reduction among cyanobacteria adapted to oligotrophic environments. For the host of UCYN-A, the ecological advantage of hosting a “diazoplast” would come at the cost of having to sustain it with carbon energy, nutrients, and a range of metabolites. Thus, it appears possible that more severe nutrient deprivation (especially for phosphorus, Scanlan et al. 2009) experienced by an open ocean ecotype of the host would also induce more extensive genome compaction (i.e. streamlining) in the symbiont. Further studies are necessary to fully understand these observations.

**Conclusions**

The genomes of the two UCYN-A strains show considerable divergence at the amino acid and nucleotide levels along with high conservation of genome structure, gene content, and basic genome features, suggesting that they had a common symbiotic ancestor and then were separated spatially in the course of speciation. While there is
some evidence for unequal distribution and possibly habitat-specific genomic
streamlining in these two strains, it remains unclear whether they occupy different or
overlapping niches. The genome size and the number of pseudogenes not yet fully
excised from the genome of both strains might suggest that UCYN-A is still in a
relatively early stage of symbiotic association with the eukaryotic host, analog to e.g. the
diazotrophic spheroid bodies found in rhopalodiacean diatoms (Kneip et al., 2008).
Genome sequencing of additional UCYN-A strains and of host genomes will show
whether the small differences in genetic potential reflect environmental adaptation in
these organisms, and whether genetic material from UCYN-A has migrated into the host
genome, as found in organelle-like stages of symbiosis (Nakayama and Ishida, 2009).
The existence of different UCYN-A strains associated with different prymnesiophytes
has implications for the trophic transfer and vertical export of N and C, and for the
distribution and regulation of N\textsubscript{2} fixation in the ocean. Further studies are needed for a
better understanding symbiotic N\textsubscript{2} fixation, and the genomic basis for UCYN-As role as a
globally important N\textsubscript{2} fixer.

Acknowledgements

We thank F. Malfatti and F. Azam at SIO for assistance with sampling and for
providing lab facilities, Kendra Turk-Kubo for carrying out PCR reactions to confirm
pseudogene sequences, S. Biller at MIT for assistance with sample handling, S. Bench for
bioinformatics assistance, and A. Thompson, J. Tripp and J. Hilton for valuable
discussions. Comments from three anonymous reviewers greatly improved the paper.
This work was supported by a Gordon and Betty Moore Foundation Marine Investigator
Award (JZ), the MEGAMER facility (supported by GBMF) and the Center for Microbial Oceanography: Research and Education (NSF grant 0424599).

**Conflict of Interest statement**

The authors declare no conflict of interest.

Supplementary information is available at ISMEJ's website.

**References**


Figure legends

Figure 1: Work flow diagram describing the cell-sorting, genome sequencing, and assembly approach used in this study. The chosen FCM sort gate was determined in earlier experiments by screening different sorted populations for the presence of UCYN-A2 nifH by QPCR, as described previously. The PRICE assembly was carried out as described in Ruby et al. (2013).

Figure 2: Circular map showing all 52 scaffolds of the UCYN-A2 draft genome aligned to the UCYN-A1 chromosome. Each concentric ring represents a scaffold, with the color code representing percent nucleotide identity. The scaffolds are sorted by length, with the longest scaffold (249,164 nt) on the outermost ring, and decreasing in length towards the center ring (shortest contig of 675 nt). The inlet graph is a histogram of % amino acid identity for all 1159 ortholog genes.

Figure 3: Examples of missing genes in UCYN-A1, demonstrating the resulting genome compaction. A total of 31 genes was found to be unique in UCYN-A2. The alignment was done using the Artemis Comparison Tool and shows closely matching gene neighborhoods apart from the missing genes (percent nucleotide identity given for aligned genes).
Figure 4: (A) Comparison of amino acid lengths of ortholog genes in UCYN-A1, UCYN-A2, and *Cyanothece* sp. 51142. (B) The range of percent gene length of the UCYN-A1 and UCYN-A2 orthologs compared to the *Cyanothece* sp. 51142 orthologs.

Figure 5: Phylogeny of 57 cyanobacteria based on a concatenated alignment of 135 highly conserved protein sequences. A detailed list and description of the genes can be found in Blank and Sánchez-Baracaldo (2010). Maximum likelihood analyses were performed using RAxML 7.4.2 (Stamatakis 2006). Bootstrap values are indicated above branches. The vertical bar marks sequences belonging to a strongly supported clade of marine unicellular N₂ fixers previously described as the SPM group (*Synechocystis*, *Pleurocapsas*, and *Microcystis*).
Table 1: Genome statistics of UCYN-A1 and UCYN-A2.

<table>
<thead>
<tr>
<th></th>
<th>UCYN-A1</th>
<th>UCYN-A2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Genome Size</strong></td>
<td>1443806</td>
<td>1485499</td>
</tr>
<tr>
<td><strong>Number of scaffolds</strong></td>
<td>1</td>
<td>52</td>
</tr>
<tr>
<td><strong>GC %</strong></td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td><strong>Coding Base Count %</strong></td>
<td>81.41</td>
<td>79.32</td>
</tr>
<tr>
<td><strong>Protein coding genes</strong></td>
<td>1200</td>
<td>1246</td>
</tr>
<tr>
<td><strong>RNA genes</strong></td>
<td>42</td>
<td>49</td>
</tr>
<tr>
<td><strong>rRNA genes</strong></td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>5S rRNA genes</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>16S rRNA genes</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>23S rRNA genes</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>tRNA genes</td>
<td>36</td>
<td>37</td>
</tr>
<tr>
<td>other RNA genes</td>
<td>6</td>
<td></td>
</tr>
</tbody>
</table>
Table 2: Annotated genes that are absent or possibly pseudogenes in the other genome. Also shown are 3 annotated genes in UCYN-A2 that match un-annotated regions in UCYN-A1. This table does not list hypothetical proteins, which account for another 25 UCYN-A1 genes that match pseudogenes in UCYN-A2, 15 genes unique in UCYN-A2, 13 genes that match pseudogenes in UCYN-A1, and 2 genes that match un-annotated ORFs in UCYN-A1 (supplemental table 1). Where given, the numbers in brackets next to the gene IDs depict the number of consecutive annotated partial genes in the other genome aligned to this particular gene sequence.

<table>
<thead>
<tr>
<th>Category</th>
<th>IMG gene ID</th>
<th>gene length (AA)</th>
<th>annotation</th>
<th>Function description</th>
</tr>
</thead>
<tbody>
<tr>
<td>646530577</td>
<td>159</td>
<td>Peroxiredoxin</td>
<td></td>
<td>protein related to alkyl hydroperoxide reductase (AhpC)</td>
</tr>
<tr>
<td>646529831</td>
<td>167</td>
<td>restriction endonuclease</td>
<td>defense</td>
<td></td>
</tr>
<tr>
<td>646530256</td>
<td>207</td>
<td>HAS barrel domain protein</td>
<td>domain in ATP synthases</td>
<td></td>
</tr>
<tr>
<td>646530363</td>
<td>398</td>
<td>NurA domain-containing protein</td>
<td>NurA domain, endo- and exonucleases</td>
<td></td>
</tr>
<tr>
<td>646530393</td>
<td>103</td>
<td>NifZ domain-containing protein</td>
<td>N₂ fixation, nif operon</td>
<td></td>
</tr>
<tr>
<td>646530716</td>
<td>318</td>
<td>transcriptional regulator, GntR family</td>
<td>transcription factors, possibly regulation of primary metabolism</td>
<td></td>
</tr>
<tr>
<td>646530983</td>
<td>554</td>
<td>predicted ATPase</td>
<td>function unknown</td>
<td></td>
</tr>
<tr>
<td>646530866</td>
<td>462</td>
<td>NAD-dependent aldehyde dehydrogenase</td>
<td>17 Kegg pathways, aldehyde substrates, various functions</td>
<td></td>
</tr>
</tbody>
</table>

UCYN-A1 genes that are possible pseudogenes in UCYN-A2
<table>
<thead>
<tr>
<th>Accession</th>
<th>Start</th>
<th>Description</th>
<th>Function/Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>646530177 (3)</td>
<td>369</td>
<td>glycerol dehydrogenase-like oxidoreductase</td>
<td>Glycerolipid metabolism, possibly involved in fermentation</td>
</tr>
<tr>
<td>646530270 (2)</td>
<td>236</td>
<td>phosphopantetheinytransferase</td>
<td>Pantothenate and CoA biosynthesis</td>
</tr>
<tr>
<td>646530303 (2)</td>
<td>812</td>
<td>uncharacterized domain HDIG-containing protein</td>
<td>Predicted membrane-associated HD superfamily hydrolase</td>
</tr>
<tr>
<td>646530304 (2)</td>
<td>1081</td>
<td>carboxamoyl-phosphate synthase large subunit</td>
<td>Pyrimidine synthesis</td>
</tr>
<tr>
<td>646530471 (5)</td>
<td>884</td>
<td>Fe-S oxidoreductase</td>
<td>Diverse reactions, energy production/conversion</td>
</tr>
<tr>
<td>646530981 (2)</td>
<td>457</td>
<td>predicted membrane protein</td>
<td>Function unknown</td>
</tr>
<tr>
<td>646530499 (3)</td>
<td>749</td>
<td>copper/silver-translocating P-type ATPase</td>
<td>Transmembrane protein, inorganic ion transport and metabolism</td>
</tr>
<tr>
<td>646530912 (2)</td>
<td>514</td>
<td>lysyl-tRNA synthetase (class II)</td>
<td>Translation, ribosomal structure and biogenesis</td>
</tr>
<tr>
<td>2528847256</td>
<td>371</td>
<td>Predicted membrane protein</td>
<td>Function unknown</td>
</tr>
<tr>
<td>2528847449</td>
<td>430</td>
<td>glucosylglycerol phosphatase (EC 3.1.3.69)</td>
<td>Osmoprotectant synthesis</td>
</tr>
<tr>
<td>2528847463</td>
<td>236</td>
<td>Tellurite resistance protein</td>
<td>Contains C-terminal domain of Mo-dependent nitrogenase</td>
</tr>
<tr>
<td>2528848101</td>
<td>208</td>
<td>thymidylate kinase</td>
<td>Pyrimidine metabolism, DNA synthesis</td>
</tr>
<tr>
<td>2528848157</td>
<td>347</td>
<td>cell shape determining protein, MreB/Mrl family</td>
<td>Cytoskeleton synthesis, cell shape determination</td>
</tr>
<tr>
<td>2528848158</td>
<td>248</td>
<td>rod shape-determining protein MreC</td>
<td>Cytoskeleton synthesis, cell shape determination</td>
</tr>
<tr>
<td>Accession</td>
<td>Start</td>
<td>Description</td>
<td>Function</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>--------------------------------------------------</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>2528848159</td>
<td>186</td>
<td>rod shape-determining protein MreD</td>
<td>cytoskeleton synthesis, cell shape determination</td>
</tr>
<tr>
<td>2528848382</td>
<td>427</td>
<td>folate/biopterin transporter</td>
<td>membrane transport</td>
</tr>
<tr>
<td>2528848428</td>
<td>165</td>
<td>2TM domain</td>
<td>function unclear, transmembrane alpha helices</td>
</tr>
<tr>
<td>2528848397</td>
<td>56</td>
<td>Sigma-70, region 4</td>
<td>DNA directed RNA polymerase</td>
</tr>
<tr>
<td>2528847785</td>
<td>344</td>
<td>folate-binding protein YgfZ</td>
<td>Predicted aminomethyltransferase, possibly glycine synthesis</td>
</tr>
<tr>
<td>2528848398</td>
<td>63</td>
<td>Sigma-70 region 3</td>
<td>DNA directed RNA polymerase</td>
</tr>
<tr>
<td>2528848519</td>
<td>215</td>
<td>Peroxiredoxin</td>
<td>detoxification of active oxygen species such as H2O2</td>
</tr>
<tr>
<td>2528847715</td>
<td>231</td>
<td>Zn-dependent hydrolases, including glyoxylases</td>
<td>pyruvate metabolism</td>
</tr>
<tr>
<td>2528848513</td>
<td>277</td>
<td>Tetratricopeptide repeat/TPR repeat</td>
<td>unclear function- involved in chaperone, cell-cycle, transcription, and protein transport complexes</td>
</tr>
<tr>
<td>2528848259</td>
<td>94</td>
<td>RNA-binding proteins (RRM domain)</td>
<td>function unclear</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accession</th>
<th>Start</th>
<th>Description</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>2528847640</td>
<td>38</td>
<td>Cytochrome B6-F complex subunit 5</td>
<td>photosynthesis, connects PSI and PSII in e' transport chain</td>
</tr>
<tr>
<td>2528848301</td>
<td>64</td>
<td>LSU ribosomal protein L33P</td>
<td>structural constituent of ribosome</td>
</tr>
<tr>
<td>2528848058</td>
<td>470</td>
<td>Hemolysins and related proteins containing CBS domains</td>
<td>membrane protein, regulate activity of associated enzymatic transporters</td>
</tr>
</tbody>
</table>

UCYN-A2 genes absent in UCYN-A1

UCYN-A2 genes that match un-annotated ORFs in UCYN-A1
<table>
<thead>
<tr>
<th>Accession</th>
<th>Length</th>
<th>Description</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>2528848162</td>
<td>211</td>
<td>Uncharacterized protein, similar to the N-terminal domain of Lon protease</td>
<td>proteolysis</td>
</tr>
<tr>
<td>2528848190</td>
<td>165</td>
<td>Predicted RNA-binding protein</td>
<td>general function prediction only</td>
</tr>
<tr>
<td>2528848352</td>
<td>86</td>
<td>Glutaredoxin-like domain (DUF836)</td>
<td>domain of unknown function</td>
</tr>
<tr>
<td>2528848421</td>
<td>267</td>
<td>Helix-turn-helix domain</td>
<td>DNA binding, gene expression regulation</td>
</tr>
<tr>
<td>2528848427</td>
<td>461</td>
<td>Domain of unknown function (DUF697)</td>
<td>function unknown</td>
</tr>
<tr>
<td>2528847887</td>
<td>301</td>
<td>CAAX protease self-immunity</td>
<td>probably protease, transmembrane protein</td>
</tr>
<tr>
<td>2528848219</td>
<td>396</td>
<td>Glycosyltransferases involved in cell wall biogenesis</td>
<td>Cell wall/membrane/envelope biogenesis</td>
</tr>
<tr>
<td>2528847369</td>
<td>350</td>
<td>UDP-N-acetylglicosamine-N-acetylmuramylpentapeptide N-acetylglicosamine transferase</td>
<td>Cell wall/membrane/envelope biogenesis</td>
</tr>
<tr>
<td>2528848143</td>
<td>294</td>
<td>competence/damage-inducible protein CinA C-terminal domain</td>
<td>transformation</td>
</tr>
<tr>
<td>2528847937</td>
<td>196</td>
<td>Putative translation factor (SUA5)</td>
<td>Translation, ribosomal structure and biogenesis</td>
</tr>
<tr>
<td>2528847508</td>
<td>140</td>
<td>Predicted endonuclease involved in recombination (possible Holliday junction resolvase in Mycoplasmas and B. subtilis)</td>
<td>Replication, recombination and repair</td>
</tr>
<tr>
<td>2528848115</td>
<td>600</td>
<td>Subtilisin-like serine proteases</td>
<td>proteolysis or cell motility</td>
</tr>
<tr>
<td>2528847345</td>
<td>385</td>
<td>phosphate ABC transporter substrate-binding protein, PhoT family (TC 3.A.1.7.1)</td>
<td>inorganic ion transport and metabolism</td>
</tr>
</tbody>
</table>

**UCYN-A2 genes that are possible pseudogenes in UCYN-A1**
Cell sorting by flow cytometry

Scripps Pier 5/31/2011 surface water

35k cells sorted;
Lysis & amplification of genomic DNA with Repli-G midi kit

Fragment library preparation at MIT (shearing, adapters), Illumina sequencing (MiSeq SP) at MIT, total of 4.39*10^6 paired end reads of 150 bp

150-350 bp library

250-550 bp library

De Novo assembly (Newbler)
167 contigs (1.47 Mb total) with high similarity to the UCYN-A1 reference genome (1.44 Mb)

Paired read iterative contig extension (PRICE) of Newbler contigs gives 52 contigs (1.485 Mb total length), mapping to nearly the entire length of UCYN-A reference genome
UCYN-A2 draft genome
52 scaffolds
1,485,499 bp
1,246 predicted proteins
31% GC content

Conserved ribosomal protein operon

rRNA operon, (-) strand

UCYN-A1 reference genome, Coding regions (+/- strand)

N₂ fixation cluster

rRNA operon, (+) strand

Frequency

% amino acid identity