









Overlooked and widespread pennate diatom-diazotroph symbioses in the sea

Christopher R. Schvarcz ¹✉, Samuel T. Wilson ¹, Mathieu Caffin¹, Rosalina Stancheva ², Qian Li ^{1,4}, Kendra A. Turk-Kubo ³, Angelicque E. White ¹, David M. Karl ¹, Jonathan P. Zehr³ & Grieg F. Steward ¹

Persistent nitrogen depletion in sunlit open ocean waters provides a favorable ecological niche for nitrogen-fixing (diazotrophic) cyanobacteria, some of which associate symbiotically with eukaryotic algae. All known marine examples of these symbioses have involved either centric diatom or haptophyte hosts. We report here the discovery and characterization of two distinct marine pennate diatom-diazotroph symbioses, which until now had only been observed in freshwater environments. Rhopalodiaceae diatoms *Epithemia pelagica* sp. nov. and *Epithemia catenata* sp. nov. were isolated repeatedly from the subtropical North Pacific Ocean, and analysis of sequence libraries reveals a global distribution. These symbioses likely escaped attention because the endosymbionts lack fluorescent photopigments, have *nifH* gene sequences similar to those of free-living unicellular cyanobacteria, and are lost in nitrogen-replete medium. Marine Rhopalodiaceae-diazotroph symbioses are a previously overlooked but widespread source of bioavailable nitrogen in marine habitats and provide new, easily cultured model organisms for the study of organelle evolution.

¹Department of Oceanography, Daniel K. Inouye Center for Microbial Oceanography: Research and Education (C-MORE), University of Hawai'i at Mānoa, Honolulu, HI 96822, USA. ²Department of Biological Sciences, California State University San Marcos, San Marcos, CA 92096, USA. ³Department of Ocean Sciences, University of California, Santa Cruz, CA 95064, USA. ⁴Present address: School of Oceanography, Shanghai Jiao Tong University, Shanghai, China. ✉email: schvarcz@hawaii.edu

Biological dinitrogen (N_2) fixation is a globally important process supporting primary production in terrestrial and aquatic ecosystems¹. In the nitrogen (N)-depleted surface waters of open ocean biomes, up to ~50% of new production can be supported by N_2 fixation². Identifying the spectrum of N_2 -fixing plankton and their diverse physiologies is essential for predicting the effects of future ocean warming on biological productivity³. Although the intrinsic biochemical capacity to fix N_2 is restricted to bacteria and archaea, some eukaryotes have adapted to low-nitrogen concentrations in the oligotrophic ocean by establishing mutualistic symbioses with diazotrophic bacteria⁴. The two most prominent types of marine diazotroph symbioses described thus far are the facultative associations between heterocyst-forming cyanobacteria and centric diatoms^{5,6} and the association between unicellular cyanobacteria and haptophytes^{7,8}.

An endosymbiotic relationship between unicellular diazotrophic cyanobacteria and pennate diatoms in the family Rhopalodiaceae has been identified in freshwater and brackish environments but has never been reported in the ocean⁹. Freshwater species in the genera *Epithemia* Kützing and *Rhopalodia* O. Müller host obligate, N_2 -fixing endosymbionts of cyanobacterial origin, also referred to as 'spheroid bodies'^{10–12}. *Epithemia*-spheroid body symbiosis is a valuable model system for studying the transformation of endosymbionts into organelles^{13,14}. For example, endosymbionts in different species of freshwater Rhopalodiaceae differ in the genes and pathways retained after reductive evolution, reflecting differences in the trajectories of their organellogenesis¹⁵.

In this work, we report the discovery of two species of oceanic, endosymbiont-bearing, rhopalodiacean diatoms, *Epithemia pelagica* Schvarcz, Stancheva & Steward sp. nov. and *Epithemia catenata* Schvarcz, Stancheva & Steward sp. nov. We show that *Epithemia* symbionts are globally distributed in the marine environment and that *E. pelagica* and *E. catenata* symbioses exhibit unique daily patterns of N_2 fixation.

Results and discussion

The *Epithemia* strains were isolated from seawater samples collected from the subtropical North Pacific Ocean at Station ALOHA (22°45' N, 158°00' W)¹⁶. Samples were collected throughout the year and from multiple depths extending through the euphotic zone (Supplementary Table 1). Overall, seven strains were isolated that represent two morphologically and genetically distinct species. *E. pelagica* is characterized by small solitary cells, asymmetrical along the apical axis (Fig. 1a–c, n), while *E. catenata* cells are larger, nearly symmetric along the apical axis, hyaline, and chain-forming (Fig. 1d–g, o). Both species typically foster 1–2 unicellular endosymbionts per cell, which have average cell dimensions of $2.9 \times 2.4 \mu\text{m}$ for *E. pelagica* (Fig. 1c, h, n) and $4.0 \times 2.5 \mu\text{m}$ for *E. catenata* (Fig. 1f, k, o) and tend to be centrally located next to the host cell's nucleus (Fig. 1n, o). Similar to freshwater Rhopalodiaceae¹⁷, the endosymbionts of *E. pelagica* (*EpSB*) and *E. catenata* (*EcSB*) lack detectable chlorophyll and phycoerythrin (Fig. 1h–m), implying that these endosymbionts may have lost their ability to photosynthesize and therefore have an obligate requirement for fixed carbon from their host. The endosymbionts were never observed growing outside of host cells and the host cultures lost their endosymbionts after being propagated for extended periods on N-replete medium (K medium¹⁸, with $5 \times 10^{-5} \text{ M NH}_4^+$, $8.82 \times 10^{-4} \text{ M NO}_3^-$; Supplementary Fig. 20). Furthermore, the hosts showed little to no growth when subsequently returned to low-N medium ($5 \times 10^{-8} \text{ M NH}_4^+$).

Phylogenetic analyses support the hypothesis that the endosymbionts are obligate and are coevolving with their *Epithemia* host. The endosymbiont SSU (encoding 16S rRNA) and *nifH*

(encoding nitrogenase iron protein) gene sequences cluster by host species and are highly conserved among different strains of the same host (99.5–100% nucleotide sequence identity). Such partner fidelity is expected for obligate symbionts and is similarly observed in marine prymnesiophyte-UCYN-A symbioses^{19,20}. The endosymbionts *EpSB* and *EcSB* reside on distinct phylogenetic branches, forming a monophyletic clade with the spheroid bodies of freshwater species *E. turgida* (Ehrenberg) Kützing (*EtSB*) and *Rhopalodia gibberula* (Ehrenberg) O. Müller (*RgSB*) (Fig. 2b). The diatom host phylogeny places *E. pelagica* and *E. catenata* on separate branches of the revised *Epithemia* lineage²¹, with their nearest neighbors being *Rhopalodia* sp. 13vi08.2B (GCCT21) and *Rhopalodia* sp. 21IV14-4D (voucher HK433), respectively (Fig. 2a). Both of these *Rhopalodia* spp. were isolated from coastal marine environments but were not noted to contain endosymbionts^{21,22}. *E. catenata*'s cell morphology differs significantly from the rest of Rhopalodiales²¹ but shares many characteristics with the tentatively classified *Nitzschia nienhuisii* F.A.S. Sterrenburg & F.J.G. Sterrenburg²³, for which no molecular data exist. Spheroid body-like structures are visible in a previously captured photomicrograph of *N. nienhuisii* (Figure 89 in Lobban²⁴), suggesting that this diatom species, which has been observed in the Pacific²⁴ and Atlantic²³ Oceans and Caribbean Sea^{25,26}, may also harbor endosymbionts.

E. pelagica-like symbioses have a global distribution throughout tropical and subtropical oceans (Fig. 3). Nucleotide sequences 100% identical to the *nifH* gene of *EpSB* were observed in samples from the North Pacific Ocean, North and South Atlantic Oceans, Indian Ocean, and the seas surrounding China, Philippines, and Japan. These include amplified *nifH* sequences from NCBI's non-redundant nucleotide (nt) and Sequence Read Archive (SRA) databases ($n = 81,023$; ave. length = 272 bp), as well as unamplified sequences from Tara Oceans metagenomes and metatranscriptomes corresponding to an assembled transcript (unigene MATOU-v1_93255274) that covers the entire length of our *EpSB nifH* sequences (760 bp). MATOU-v1_93255274 was most frequently found in metagenomes and metatranscriptomes generated from eukaryote-associated size fractions (2, 10, 8, and 2 samples corresponding to the 0.8–5 μm , 5–20 μm , 20–180 μm , and 180–2000 μm size fractions, respectively). *Epithemia* symbioses may be even more abundant and widespread, because lowering the threshold of nucleotide identity from 100% to 98% expands the global distribution of *EpSB*-like *nifH* sequences to include the Gulf of Mexico, Coral Sea, and the Arctic Ocean (Supplementary Fig. 21a). *EcSB*-like *nifH* sequences were rarer and only found in samples from the North Pacific Ocean and Coral Sea (Supplementary Fig. 21b). At Station ALOHA, where the isolates were obtained, quantitative PCR measurements of *E. pelagica* symbioses indicated host LSU and *EpSB nifH* gene copies as high as $18 \pm 8 \times 10^3 \text{ L}^{-1}$ and $0.7 \pm 0.2 \times 10^3 \text{ L}^{-1}$, respectively (Supplementary Fig. 22a). Identical *EpSB nifH* sequences were also detected in metagenomes constructed from sinking particles collected at 4000 m depth at Station ALOHA over a 3-year period (Supplementary Fig. 22b)^{27,28}. The recurring presence of *EpSB* sequences in sinking particles collected in the bathypelagic zone of the water column is strong evidence for a sustained population in the surface waters of the North Pacific Subtropical Gyre.

The daily patterns of N_2 fixation in *E. pelagica* and *E. catenata* endosymbionts are distinct from other pelagic diazotrophs. In general, N_2 fixation by marine cyanobacteria occurs during either the day or the night²⁹. For both *EpSB* and *EcSB*, N_2 fixation occurred during the day and night (Fig. 4). During the day, N_2 fixation ceases either a few hours prior to the end of the light period (*E. pelagica*; Fig. 4a) or in conjunction with the lights being switched off (*E. catenata*; Fig. 4b). For both strains, N_2 fixation was subsequently undetected for the first 6 h of the dark

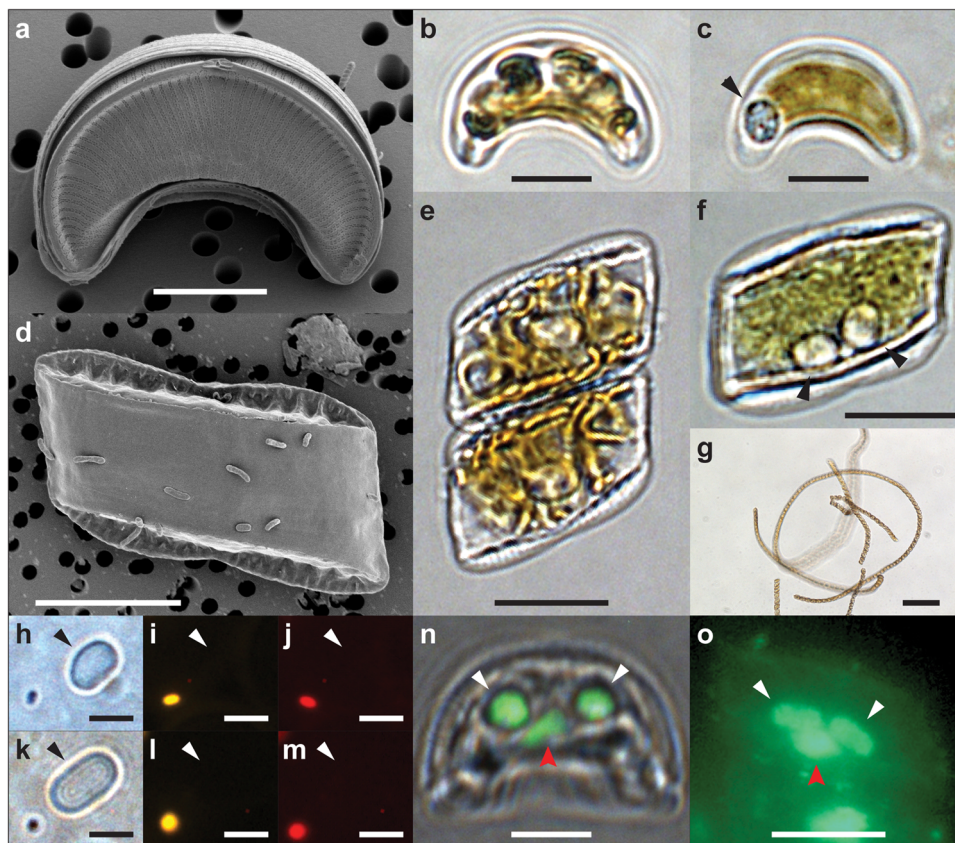


Fig. 1 Characterization of marine *Epithemia* and their endosymbionts. **a, b, c** Micrographs of *E. pelagica* UHM3200 in valve view, including visualization by **a** scanning electron microscopy, and light microscopy for **b** live cells and **c** cells osmotically shocked to display the endosymbionts, as indicated by the black arrow. The osmotic shock treatment (**c, f**) disrupts host cell contents and displaces the endosymbionts. The natural intracellular location of endosymbionts is illustrated in **n** and **o**. **d, e, f, g** Micrographs of *E. catenata* UHM3210 in girdle view, including visualization by **d** scanning electron microscopy, light microscopy for **e** live cells and **f** osmotically shocked cells with two endosymbionts highlighted by black arrowheads, and **g** undisturbed live cells growing in long chains. Micrographs of *EpSB* (**h, i, j**) and *EcSB* (**k, l, m**) released from crushed UHM3200 and UHM3210 host cells, respectively, as seen under brightfield (**h, k**), phycoerythrin fluorescence (**i, l**), and chlorophyll fluorescence (**j, m**). Arrowheads point to the endosymbiont cells, and a single *Synechococcus* WH7803 cell is present in each field to serve as a positive control for fluorescence. **n, o** Fluorescence micrographs of fixed *E. pelagica* and *E. catenata* cells, respectively, where endosymbionts (white arrows) and nuclei (red arrows) have been stained with nucleic acid-binding SYBR Gold dye (green). In **n**, the fluorescence channel has been overlaid on a brightfield micrograph. Micrograph scale bars are 5 μm (**a-c, n**), 10 μm (**d-f, o**), 100 μm (**g**), and 3 μm (**h-m**). The experiment assessing the autofluorescence of released endosymbionts (**h-m**) was performed once, while all other micrographs represent results that were consistently reproduced in multiple experiments.

period and resumed just after midnight. Overall, *EpSB* and *EcSB* are able to fix N_2 for a much longer period of time during a day-night cycle than other marine diazotrophs, especially the unicellular cyanobacterium *Crocospaera subtropica* Mareš & J.R. Johansen^{30,31}, which is the closest free-living relative of rhopalodiacean endosymbionts (Fig. 2b). *C. subtropica* synthesizes its carbohydrates during the day and respire them at night to fuel N_2 fixation, while the evolutionary transition to an endosymbiont has enabled the *EpSB* and *EcSB* spheroid bodies to perform N_2 fixation during the day which is most likely fueled by metabolism of the host cell¹⁴. A similar evolutionary transition is hypothesized to have occurred for UCYN-A which fixes N_2 during the daytime³².

Analysis of marine *Epithemia* endosymbiont *nifH* sequences suggests that these endosymbionts have been commonly misidentified in oceanic samples as free-living unicellular cyanobacteria within the UCYN-C group (Supplementary Fig. 23). UCYN-C is considered to represent free-living unicellular cyanobacteria closely related to *Crocospaera* Zehr, Rachel A. Foster, Waterbury & E. Webb (including strains formerly identified as *Cyanothece* Komárek³³), but published reports of UCYN-C have

employed *nifH* primers and probes with an identical match to *E. pelagica* (Supplementary Fig. 24)^{34–36}. Other studies have detected the presence of spheroid body-like sequences in marine metatranscriptomes³⁷ and *Tara* Oceans samples³⁸, but until now there were no characterized marine rhopalodiacean symbioses to support these observations. With the observed distribution of *EpSB*- and *EcSB*-like sequences throughout tropical and subtropical oceans, *Epithemia* symbioses may be the most geographically widespread eukaryote–diazotroph association on the planet, occurring in freshwater, brackish, and marine environments. Establishing the spatial-temporal patterns of marine *Epithemia* will be essential for determining their contribution to the marine N cycle and the biological carbon pump². Furthermore, the repeated isolation of endosymbiont-bearing *Epithemia* species demonstrates their ease of cultivation compared to other marine diazotroph–eukaryote symbioses, which are either uncultivated or difficult to maintain long-term in culture⁴. Thus, marine *Epithemia* are valuable new model systems for genomic and experimental investigations, because they can be used to unravel the interplay of host–symbiont physiologies and the genetic adaptations that maintain these relationships³⁹.

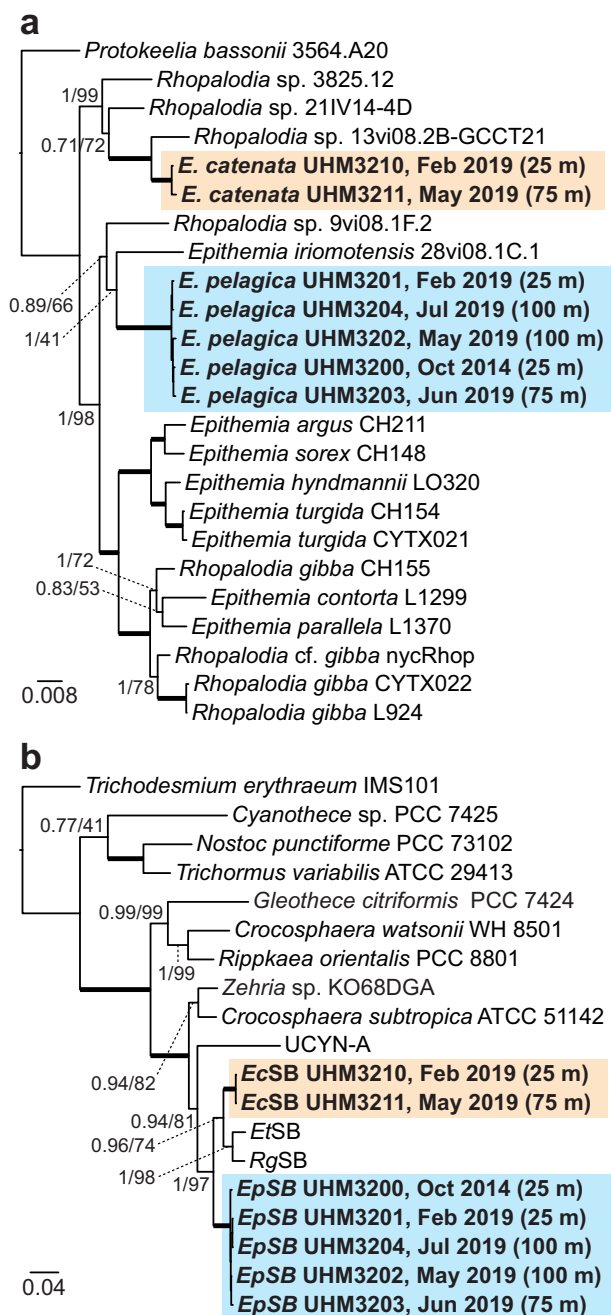


Fig. 2 Phylogenetic analyses. **a** Multigene phylogeny of the *E. pelagica* and *E. catenata* diatom hosts, based on the SSU (18S rRNA; 1322 nt), *psbC* (986 nt), and *rbcL* (1341 nt) genes. **b** Multigene phylogeny of the *Epithemia* endosymbionts, based on the SSU (16S rRNA; 420 nt) and *nifH* (760 nt) genes. The phylogenies are Bayesian majority consensus trees, and support values are provided for Bayesian and maximum likelihood methods (Bayesian posterior probabilities/ML bootstrap percent). Bold branches indicate complete support (posterior probability of 1 and bootstrap percent of 100), and the phylogeny scales are in units of nt substitutions per site. Labels for the new isolates (highlighted and bolded) include the dates and depths of collection. Accession numbers for all sequences are provided in the Source Data file.

Systematic biology. Phylum Bacillariophyta Karsten
 Class Bacillariophyceae Haeckel
 Order Rhopalodiales D. G. Mann
 Family Rhopalodiaceae (Karsten) Topachevs'kyj & Oksiyuk
 Genus *Epithemia* Kützing

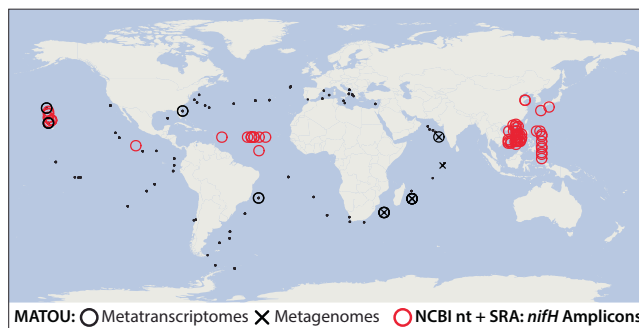


Fig. 3 Global map of environmental sequences that share 100% nucleotide identity with an *E. pelagica* *nifH* phylogroup. Datasets screened include the Marine Atlas of Tara Oceans Unigenes (MATOU) and the representation of unigene MATOU-v1_93255274 in Tara Oceans metagenomes and metatranscriptomes, as well as *nifH* amplicons published in NCBI's non-redundant nucleotide (nt) and Sequence Read Archive (SRA) databases. Black dots indicate the location of Tara Oceans stations analyzed in MATOU. Source data are provided in the Source Data file.

Epithemia pelagica Schvarcz, Stancheva & Steward sp. nov.
 Figure 1a–c, n; Supplementary Figs. 1a–n, 2a–h, 6–10.

Holotype. Slide UC2085162 from *Epithemia pelagica* UHM3201 deposited in the University and Jepson Herbaria at the University of California, Berkeley. Holotype specimen is illustrated in Supplementary Fig. 1b.

Etymology. The epithet refers to the pelagic habitat of this species, collected from the open Pacific Ocean.

Type locality. This species was isolated from seawater collected from a depth of 25 m in the North Pacific Subtropical Gyre (22°45' N, 158°00' W), Station ALOHA (water depth ca. 4800 m) on October 14, 2014 (*Epithemia pelagica* UHM3200) and February 21, 2019 (*Epithemia pelagica* UHM3201).

Diagnosis. This species is characterized by solitary, strongly dorsiventral, small cells, 6.7–17.8 μm long, 5–9.8 μm wide. Valves are lunate with rounded apices, convex dorsal margin, and concave ventral margin. Raphe-bearing keel is eccentric, positioned on the dorsal margin, slightly bent down towards the dorsal margin at the center of the valve. Transapical costae are fine and resolved only near the raphe keel, where they are internally thickened and function as fibulae beneath the raphe. Striae are not resolvable with light microscopy (LM). *E. pelagica* possesses all structural features of genus *Rhopalodia* (now *Epithemia*), but differs from other species by its minute size, weakly silicified frustules with delicate costae and very fine striae not resolvable with LM.

Note. This species also shares some morphological similarities with *Protokeelia* C.W. Reimer & J.J. Lee, such as minute size, lunate valve with undulate valve face, and protuberant raphe sternum.

Description. See Supplementary Note 1.

Epithemia catenata Schvarcz, Stancheva & Steward sp. nov.
 Figure 1d–g, o; Supplementary Figs. 3a–c, 4a–k, 5a–f, 11, 12.

Holotype. Slide UC2085161 from *Epithemia catenata* UHM3210 deposited in the University and Jepson Herbaria at University of

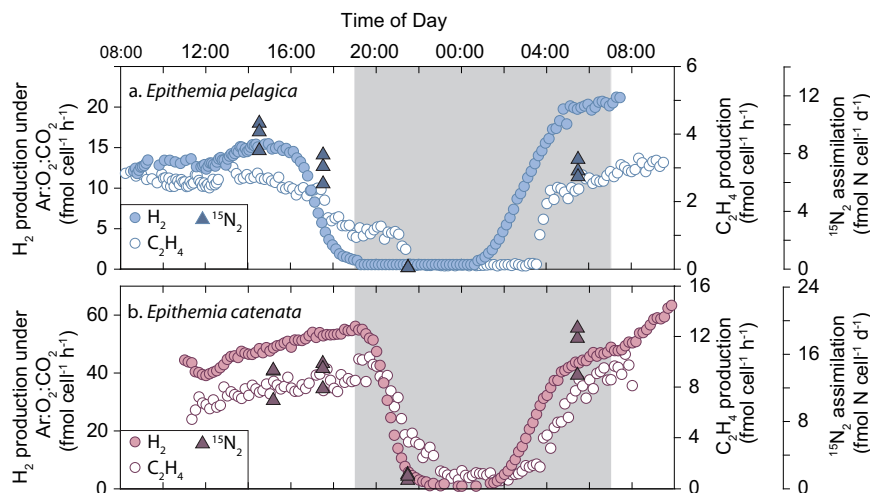


Fig. 4 Daily patterns of N_2 fixation for *Epithemia*. **a** Dihydrogen (H_2) production, ethylene (C_2H_4) production from acetylene (C_2H_2), and $^{15}\text{N}_2$ assimilation for *E. pelagica* UHM3200. **b** H_2 production, C_2H_4 production from C_2H_2 , and $^{15}\text{N}_2$ assimilation for *E. catenata* UHM3210. Sampling resolution was 10 min for H_2 and C_2H_4 production and 2 h for the $^{15}\text{N}_2$ assimilation with the values represented by average \pm standard deviation ($n = 3$). The $^{15}\text{N}_2$ assimilation beginning at 20:30 had rates of $0.1 \text{ fmol N cell}^{-1} \text{ d}^{-1}$ for *E. pelagica*. The night period is represented by gray shading in both plots. Source data are provided in the Source Data file.

California, Berkeley. Holotype specimen is illustrated in Supplementary Fig. 3a.

Etymology. The epithet refers to the colonial habit of this diatom, which forms long chains.

Type locality. This species was isolated from seawater collected from a depth of 25 m in the North Pacific Subtropical Gyre ($22^\circ 45' \text{ N}$, $158^\circ 00' \text{ W}$), Station ALOHA (water depth ca. 4,800 m) on February 21, 2019 (*Epithemia catenata* UHM3210) and from a depth of 75 m on May 4th, 2019 (*Epithemia catenata* UHM3211).

Diagnosis. This species is characterized by cells joined together in chains. Frustules are delicate, translucent without visible striation, with wide, slightly rhomboidal girdles and narrower elliptical valves. The raphe keel is central or nearly so on valve face, continuing from pole to pole, slightly sigmoid, and fibulate, involved in cell-cell interlock. Frustules are $17.2\text{--}28.8 \mu\text{m}$ along apical axis, $8.2\text{--}13.7 \mu\text{m}$ along transapical axis, $11.7\text{--}16.9 \mu\text{m}$ along perivalvar axis, fibulae are 8–12 in $10 \mu\text{m}$, 2–3 μm long. This species differs morphologically from all known *Epithemia/Rhopalodia* taxa by its colonial habit, raphe position, and frustular ultrastructure.

Note. The generic placement of *E. catenata* is based on current molecular phylogenetics data (Fig. 2a, Supplementary Figs. 13–19) and the presence of cyanobacterial endosymbionts. SH and AU tests of constrained phylogenetic topologies (Supplementary Table 3) show greater support for the inclusion of *E. catenata* within the genus *Epithemia*, while showing weaker support for the exclusion of *E. catenata* from either the genus *Epithemia* or family Rhopalodiaceae. *E. catenata* shares more morphological characteristics with *Nitzschia nienhuisii* than with other *Epithemia/Rhopalodia* species, such as gross frustule symmetry, keel structure and position, chain-formation, and hyaline frustules (LM Figs. 2–4 in Sterrenburg & Sterrenburg²³; LM Figs. 89–92, SEM Figs. 93, 94 in Lobban²⁴; SEM Fig. 2a, b, LM Fig. 2c, d in Lópes-Fuerte et al.²⁶), suggesting that these two species may be congeneric. The initial description of *N. nienhuisii* acknowledges there is doubt with respect to *N. nienhuisii*'s generic ranking²³, and a later study employing SEM noted the

baffling structure of *N. nienhuisii*, which does not appear to possess fibulae on the keel typical of *Nitzschia* Hassall²⁴.

Description. See Supplementary Note 2.

Methods

Epithemia isolation and culture. The *Epithemia* cells were isolated from 0.5 L of seawater collected from depths of 25, 75, and 100 m in the North Pacific Subtropical Gyre ($22^\circ 45' \text{ N}$, $158^\circ 00' \text{ W}$). Seawater was collected during the near-monthly Hawaii Ocean Time-series (HOT) expeditions to the long-term monitoring site Station ALOHA (water depth ca. 4800 m) in October 2014 (HOT cruise #266) and February–July 2019 (HOT cruises #310–313). Serial dilution (single algal strains UHM3202, UHM3203, UHM3204) or micropipette isolation of single cells (clonal strains UHM3200, UHM3201, UHM3210, UHM3211) were used to establish the *Epithemia* cultures, which were grown in a seawater-based, low-nitrogen medium. Filtered ($0.2 \mu\text{m}$) and autoclaved, undiluted Station ALOHA seawater was amended with $2 \mu\text{M}$ EDTA, 50 nM ferric ammonium citrate, $7.5 \mu\text{M}$ phosphoric acid, trace metals (100 nM MnSO_4 , 10 nM ZnCl_2 , 10 nM Na_2MoO_4 , 1 nM CoCl_2 , 1 nM NiCl_2 , 1 nM Na_2SeO_3), vitamins ($50 \mu\text{g/L}$ inositol, $10 \mu\text{g/L}$ calcium pantothenate, $10 \mu\text{g/L}$ thiamin, $5 \mu\text{g/L}$ pyridoxine HCl, $5 \mu\text{g/L}$ nicotinic acid, $0.5 \mu\text{g/L}$ para-aminobenzoic acid, $0.1 \mu\text{g/L}$ folic acid, $0.05 \mu\text{g/L}$ biotin, $0.05 \mu\text{g/L}$ vitamin B_{12}), and $106 \mu\text{M}$ Na_2SiO_3 . Although not tested here, simpler formulations of diazotroph media such as PMP⁴⁰ or RMP⁴¹ may also be suitable for growing *Epithemia*, when made with 100% seawater and adding Na_2SiO_3 . The cultures were subsequently incubated at 24°C on a 12:12 h light:dark cycle with $50\text{--}100 \mu\text{mol quanta m}^{-2} \text{ s}^{-1}$ using cool white fluorescent bulbs. All *E. pelagica* and *E. catenata* symbioses were stable under these medium and incubation conditions. *E. pelagica* was successfully isolated from at least one of the three depths that were targeted during each sampling occasion.

Morphological observations. *Epithemia* living and fixed cells were imaged by light and epifluorescence microscopy using a Nikon Eclipse 90i microscope at $40\times\text{--}60\times$ magnification. Diatom cell sizes were determined using >60 live, exponentially growing cells, imaged in either valve view (*E. pelagica*) or girdle view (*E. catenata*). Endosymbiont (spheroid body) cell sizes were averaged from DNA-stained cells for *E. pelagica* UHM3200 ($n = 78$) and *E. catenata* UHM3210 ($n = 91$), imaged by epifluorescence microscopy after preparing samples as follows: *Epithemia* cells were fixed in 4% glutaraldehyde for 30 min, pelleted at $1000 \times g$ for 1 min, the supernatant was exchanged with 0.5% Triton X-100 (in autoclaved filtered seawater), samples were incubated for 10 min with gentle agitation, cells were then pelleted at $4000 \times g$ for 1 min, supernatant was exchanged with autoclaved filtered seawater and fixed in 4% glutaraldehyde, and samples were stained with $1\times$ final concentration of SYBR Gold nucleic acid stain (Invitrogen, cat. # S11494) for 2 h. For routine observations of endosymbionts (e.g., determining presence/absence and number per host cell), osmotic shock was used to disrupt the cell contents of diatom host cells and improve visualization of the endosymbionts. This was achieved by gently pelleting cells and exchanging the medium with either ultrapure water or 2–3 M NaCl solution, followed by immediate observation. While this is a simple technique for detecting and visualizing endosymbionts (Fig. 1c, f), it

does not accurately represent the natural location of endosymbionts within the host cells, as seen when compared to fixed cell preparations for epifluorescence microscopy (Fig. 1n, o). To assess the presence of fluorescent photopigments in endosymbiont cells, live host cells were pelleted at $4000 \times g$ for 5 min and crushed using a microcentrifuge tube pestle (SP Bel-Art, cat. # F19923-0000) to release the endosymbionts. The crushed pellet was resuspended in 75% glycerol containing live *Synechococcus* WH7803 cells (positive control for fluorescence), and samples were observed by epifluorescence microscopy using filter cubes appropriate for observing phycoerythrin (EX: 551/10, BS: 560, EM: 595/30) and chlorophyll (EX: 480/30, BS: 505, EM: 600LP).

The loss of endosymbionts from *Epithemia* cultures (UHM3200 and UHM3210) was observed after propagating cells for four months in nitrogen-replete medium (K)¹⁸, where approximately 5–10% of the culture was transferred to fresh medium about every two weeks. Observations were only made at the end of the four-month period. Endosymbionts were not observed growing freely in these cultures, and the absence of endosymbionts within host cells was confirmed by the failure to observe spheroid bodies by light microscopy after osmotic shock of the diatoms, as well as a failure to amplify the endosymbiont SSU (16S rRNA) and *nifH* genes from cellular DNA extracts. PCR reactions were performed in parallel with DNA extracts from control cultures (grown in low-nitrogen medium), using the same template DNA amount (10 ng) and PCR conditions (see methods for *Marker gene sequencing and phylogenetics*).

Ultrastructural observations by electron microscopy (EM) were conducted for *E. pelagica* UHM3200 and *E. catenata* UHM3210. EM preparations of diatoms typically involve the oxidative removal of organic matter to uncover the fine details of frustule ultrastructure. However, in the case of *E. catenata*, oxidatively cleaned cells lacked structural integrity, leading to collapsed frustules when dried and viewed by scanning EM (SEM). For this reason, both species were prepared for SEM with and without (Fig. 1a, d) the oxidative removal of organic matter, and cleaned *E. catenata* frustules were further analyzed by transmission EM (TEM). To remove organic matter, 100 mL of exponentially growing culture was pelleted by centrifugation at $1000 \times g$ for 10 min and resuspended in 30% H₂O₂. Cells were boiled in H₂O₂ for 1–2 h, followed by rinsing cells six times in ultrapure water by sequential centrifugation at $1000 \times g$ for 10 min and resuspension of cell pellets. Suspensions of the cleaned cells were dried on aluminum foil and mounted on aluminum stubs with double-sided copper tape. For some *E. catenata* SEM preparations, the cleaned frustules were dehydrated in an ethanol dilution series and exchanged into hexamethyldisilazane (HMDS) prior to drying on aluminum foil; this was to minimize the collapse of frustules resulting from drying. To prepare cells with organic matter intact, 25 mL of exponentially growing culture was mixed with an equal volume of fixative solution (5% glutaraldehyde, 0.2 M sodium cacodylate pH 7.2, 0.35 M sucrose, 10 mM CaCl₂) and incubated overnight at 4 °C. Cells were gently filtered onto a 13 mm diameter 1.2 µm pore size polycarbonate membrane filter (Isopore, Millipore Sigma), washed with 0.1 M sodium cacodylate buffer (pH 7.4, 0.35 M sucrose), fixed with 1% osmium tetroxide in 0.1 M sodium cacodylate (pH 7.4), dehydrated in a graded ethanol series, and critical point dried. Filters were mounted on aluminum stubs with double-sided conductive carbon tape. All SEM stubs were sputter coated with Au/Pd, prior to observing on a Hitachi S-4800 field emission scanning electron microscope at the University of Hawai'i at Mānoa (UHM) Biological Electron Microscope Facility (BEMF). Cleaned *E. catenata* cells were prepared for TEM by drying a drop of sample on a formvar/carbon-coated grid and observing on a Hitachi HT7700 transmission electron microscope at UHM BEMF.

Additional light microscopy of hydrogen-peroxide cleaned frustules was conducted for *E. pelagica* UHM3201 and *E. catenata* UHM3210. Samples were mounted in Naphrax (PhycoTech, Inc., cat. # P-Naphrax200) and observed at 100× using an Olympus BX41 Photomicroscope (Olympus America Inc., Center Valley, Pennsylvania) with differential interference contrast optics and an Olympus SC30 Digital Camera at California State University San Marcos.

A key to the strains used in each micrograph is provided in Supplementary Table 2.

Marker gene sequencing and phylogenetics. For each *Epithemia* strain, 25–50 mL of culture was pelleted at $4000 \times g$ for 10 min, and DNA was extracted from the pellet using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research, cat. # D4300). Marker genes were amplified with the Expand High Fidelity PCR System (Roche, cat. # 4743733001), using conditions previously described for genes SSU encoding 18S rRNA (Euk328f/Euk329r)⁴², LSU encoding 28S rRNA (DIR/D2C)⁴³, *rbcL* (*rbcL*.66+/dp7–)^{44,45}, *psbC* (*psbC*+/*psbC*–)⁴⁴, and *cob* (*Cob1f*/*Cob2r*)²¹. For the endosymbionts, a partial sequence for the SSU (16S rRNA) gene was amplified using a primer set targeting unicellular cyanobacterial diazotrophs, CYA359F/Nitro821R^{46,47}, and the *nifH* gene was amplified using new primers specific to the *nifH* of *Cyanothece*-like organisms, ESB-nifH-F (5'-TACGGAAAAGCGGTA TCGG-3') and ESB-nifH-R (5'-CACCAACAGRATACCGAAGTC-3'), with a 55 °C annealing temperature and 75 s extension time. All primers were synthesized by Integrated DNA Technologies (IDT). Amplified products were cloned and transformed into *E. coli* using the TOPO TA Cloning Kit for Sequencing (Invitrogen, cat. # K457501), and plated colonies were picked and grown in Circlegrow medium (MP Biomedicals, cat. # 113000132). Plasmids were extracted with the Zippy Plasmid Miniprep kit (Zymo Research, cat. # D4019) and sequenced from

the M13 vector primers using Sanger technology at GENEWIZ (South Plainfield, NJ). For the diatom SSU (18S rRNA) gene, sequencing reactions were also performed using the 502f and 1174r primers⁴⁸.

Phylogenetic trees (Fig. 2) were inferred using concatenated alignments for both diatom host genes (SSU encoding 18S rRNA, *psbC*, *rbcL*) and endosymbiont genes (SSU encoding 16S rRNA, *nifH*). For each gene, nucleotide sequences were aligned using MAFFT v7.453⁴⁹ (L-INS-i method), and sites with gaps or missing data were removed. An appropriate nucleotide substitution model was selected for each gene alignment using jModelTest v2.1.10⁵⁰. Bayesian majority consensus trees were inferred from the concatenated alignments using MrBayes v3.2.7⁵¹ with two runs of 4–8 chains, until the average standard deviation of split frequencies dropped below 0.01. Maximum likelihood bootstrap values were generated for the Bayesian tree using RAxML v8.2.12⁵², implemented with 1000 iterations of rapid bootstrapping. To further analyze the phylogenetic position of the new *Epithemia* species in the broader context of Surirellales and Rhopalodiales diatoms, individual gene trees (SSU encoding 18S rRNA, LSU, *rbcL*, *psbC*, and *cob*; Supplementary Figs. 13–19) were constructed from sequences aligned using MAFFT (automatic detection method) and trimmed using trimAl v1.2⁵³ (gappycout method). rRNA gene phylogenies were also inferred using sequences aligned according to the global SILVA alignment for SSU and LSU genes using SINA⁵⁴, which were either left untrimmed in the case of the LSU gene or trimmed to remove highly variable positions (SINA's "012345" positional variability filter) and gappy positions (trimAl v1.2, gappycout method) in the case of the SSU gene. These trimming strategies were selected based on their ability to maximize the monophyly of the previously described Rhopalodiales clade and minimize the separation of known conspecific strains, such as the strains of *E. pelagica* described here. All gene phylogenies were inferred using the Bayesian methods described above. To investigate the level of support for constrained tree topologies placing *E. catenata* within or outside of the genus *Epithemia* and family Rhopalodiaceae, SH⁵⁵ and AU⁵⁶ statistical tests were performed in IQ-TREE⁵⁷ (implementing ModelFinder⁵⁸) using all alignments from the individual gene trees (Supplementary Table 3).

Given *E. catenata*'s unusual morphology, test trees were inferred with the inclusion of diatom sequences from orders Bacillariales (*Nitzschia*, *Pseudonitzschia*), Cymbellales (*Didymosphenia*), Naviculales (*Amphiprora*, *Navicula*, *Pinnularia*), and Thalassiophysales (*Amphora*, *Halamphora*, *Thalassiphysa*); however, *E. catenata* was consistently placed within Rhopalodiales, and these trees were not pursued further.

An additional *nifH* phylogeny was constructed using all environmental sequences from NCBI's non-redundant nucleotide (nt) database >300 bp and sharing >95% nucleotide sequence identity with *EpSB* and *EcSB nifH* sequences (Supplementary Fig. 23), including 51 environmental sequences from prior studies investigating marine diazotrophs^{34,59–66}. Environmental *nifH* sequences were aligned to the previously generated *nifH* sequence alignment using MAFFT (automatic method detection and addfragments options), and the best-scoring maximum likelihood phylogeny was inferred using RAxML with 1000 iterations of rapid bootstrapping. NCBI accession numbers for all tree sequences are in the Source Data file.

Analysis of *Epithemia* endosymbiont *nifH* sequences in environmental datasets. Nucleotide sequences for *EpSB* and *EcSB nifH* were queried against NCBI's non-redundant nucleotide (nt) database using webBLAST⁶⁷ (megablast; <https://blast.ncbi.nlm.nih.gov/>) and SRA databases for *nifH* amplicon sequencing projects from the marine environment using the SRA Toolkit⁶⁸ (dc-megablast, with database validation using vdb-validate; <https://github.com/ncbi/sra-tools>). Database hits with 98–100% nucleotide identity over an alignment of the entire subject sequence (BLAST alignment length = subject sequence length) were identified, and the associated sample's latitude and longitude coordinates (where available) were mapped. Coordinates were also mapped for metagenome and metatranscriptome samples containing matches to unigene MATOU-v1_93255274 from the Marine Atlas of Tara Oceans Unigenes⁶⁹, a unigene that shares 100% identity over the entire length of the *EpSB* UHM3202 *nifH* sequence and >99.4% identity with all other *EpSB nifH* sequences.

The presence of *EpSB* and *EcSB nifH* sequences was examined in metagenomes prepared from sinking particles collected at 4000 m depth at Station ALOHA^{27,28}. The sinking particles were collected during intervals of 12, 10, and 8 days during 2014, 2015, and 2016, respectively, using a McLane sediment trap equipped with a 21-sample bottle carousel. The presence of *EpSB* and *EcSB nifH* sequences in the metagenomes was assessed by blastn⁷⁰, after first removing low quality bases from metagenomic reads using Trimmomatic v0.39⁷¹ (parameters: LEADING:20 TRAILING:20 MINLEN:100). For each sediment trap metagenome, the total number of reads matching *EpSB* or *EcSB nifH* nucleotide sequences with 100% identity were tallied and normalized to the total number of reads in the database. Only *EpSB*-matching reads were detected in this analysis.

Quantitative PCR. Specific PCR primers were designed targeting a 102 bp region of *E. pelagica*'s LSU gene (Epel-LSU-F, 5'-GAAACAGTGAAGCCAAC-3'; Epel-LSU-R, 5'-AGGCCATTATCATCCCTTGTC-3') and an 85 bp region *EpSB*'s *nifH* gene (EpSB-nifH-F, 5'-CACACTAAAGCACAACTACC-3'; EpSB-nifH-R, 5'-CAAGTAGTACTTCGCTAGCTC-3') and were synthesized by IDT. Gene copy concentrations were quantified for Station ALOHA water samples (~2 L) collected

by Niskin bottles at 5, 25, 45, 75, 100, 125, 150, and 175 m on January 16 and July 1 (except 5 m), 2014, during HOT cruises #259 and #264. Samples were filtered onto 25 mm diameter, 0.02 µm pore size aluminum oxide filters (Anotop; Whatman, cat. # WHA68092102) and stored at -80 °C until extracting DNA using the MasterPure Complete DNA and RNA Purification Kit (Epicentre, cat. # MC85200) according to Mueller et al.⁷². Briefly, a 3-mL syringe filled with 1 mL of tissue and cell lysis solution (MasterPure) containing 100 µg mL⁻¹ proteinase K was attached to the outlet of the filter, and the filter inlet was sealed with a second 3-mL syringe. The lysis solution was pulled halfway through to saturate the filter membrane, and the entire assembly was incubated at 65 °C for 15 min while attached to a rotisserie in a hybridization oven rotating at ca. 16 rpm. The lysis buffer was then drawn fully into the inlet syringe, transferred to a microcentrifuge tube, and placed on ice. The remaining steps for protein precipitation and removal and nucleic acid precipitation were carried out following the manufacturer's instructions. For each sample, DNA was resuspended in a final volume of 100 µL. Quantitative PCR (qPCR) was performed using the PowerTrack SYBR Green Master Mix system (Applied Biosystems, cat. # A46109) and run on an Eppendorf Mastercycler epgradient S realplex² real-time PCR machine. Reactions (20 µL total volume) were prepared according to the manufacturer's protocol, containing 500 nM of each primer. Sample reactions (four replicates) contained 2 µL of environmental DNA extract (24–76 ng DNA), while standards (three replicates) contained 2 µL of gBlocks Gene Fragments (IDT) that were prepared at 1, 2, 3, 4, 5, and 6 log gene copies/µL. The gBlocks Gene Fragments were 500 bp in length and encompassed the entire *E. pelagica* UHM3201 LSU sequence and positions 1–500 of the *EpSB* UHM3201 *nifH* sequence, respectively. The main cycling conditions consisted of an initial denaturation and enzyme activation step of 95 °C for 2 min, followed by 40 cycles of 95 °C for 5 s and 57 °C for 30 s for the LSU and *nifH* genes, respectively. Melting curves were analyzed to verify the specificity of the amplifications, and reactions containing *Epithemia catenata* DNA extract were included as negative controls. Reaction efficiencies were 104.23% and 95.15% for the LSU and *nifH* genes, respectively. The limit of detection for these assays was not empirically determined. gBlocks sequences, qPCR threshold cycle values, and conversion equations are provided in the Source Data file.

Physiology experiments. The daily patterns of N₂ fixation were quantified for *E. pelagica* UHM3200 and *E. catenata* UHM3210 using two techniques: acetylene (C₂H₂) reduction to ethylene (C₂H₄) and argon induced dihydrogen (H₂) production (AIHP). Both analyses were conducted using a gaseous flow-through system that quantified the relevant trace gas on the sample outlet line with a temporal resolution of 10 min⁷³. To conduct the measurements, a 10-mL subsample of each *Epithemia* culture was placed in a 20-mL borosilicate vial and closed using gas-tight rubber stoppers and crimp seals. Separate bottles were used for H₂ production and C₂H₂ reduction. During the experimental period, the temperature was maintained at 25 ± 0.2 °C using a benchtop incubator (Incu-Shaker; Benchmark Scientific) and light exposure was 200 µmol photons m⁻² s⁻¹ at wavelengths of 380–780 nm with a 12:12 h square light:dark cycle (Prime HD+; Aqua Illumination). To conduct the AIHP method, the sample vial containing the culture was flushed with a high purity gas mixture consisting of argon (makeup gas; 80%), oxygen (20%), and carbon dioxide (0.04%). In the absence of N₂, all of the electrons that would have been used to reduce N₂ to NH₃ are diverted to H₂ production, thereby providing a measure of Total Nitrogenase Activity (TNA). The C₂H₂ reduction assay also represents a measure of TNA. Our analytical set-up introduced C₂H₂ at a 1% addition (vol/vol) to the high purity air with a total flow rate (13 mL min⁻¹) identical to the AIHP method. The gas emissions were analyzed using separate reductive trace gas analyzers that were optimized for the quantification of H₂ and C₂H₄. To verify the observed daily patterns in N₂ fixation, ¹⁵N₂ assimilation measurements were conducted on triplicate samples of *Epithemia* cultures at targeted time points. Five milliliters of ¹⁵N-enriched seawater was added to the subsamples, which were subsequently crimp sealed and incubated for a 2 h period with the same light and temperature conditions as the daily gas measurements. At the end of the incubation, the contents of each vial were filtered onto a pre-combusted glass fiber filter. The concentration and isotopic composition (δ¹⁵N) of particulate nitrogen for incubated and non-incubated (i.e., natural abundance) samples was measured using an elemental analyzer/isotope ratio mass spectrometer (Carlo-Erba EA NC2500 coupled with a ThermoFinnigan Delta Plus XP). For each of the described analyses, cell-specific rates were calculated based on the average of triplicate cell concentration measurements, obtained from cell samples preserved at 4 °C with Lugol's iodine solution and quantified within a week using a Sedgwick-Rafter counting chamber (Electron Microscopy Sciences, cat. # 68050-52). All rate measurement data is provided in the Source Data file.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Sequences produced for this study have been deposited in GenBank under accession numbers MW562846–MW562894. Analyses were also conducted using data from NCBI's non-redundant nucleotide (nt) and protein (nr) databases and the following SRA run database accessions: DRR075654–DRR075675, DRR090493–DRR090512,

DRR119299–DRR119318, SRR11748760–SRR11748769, SRR11784070–SRR11784101, SRR1994968–SRR1994982, SRR2846720, SRR2846725, SRR2848263–SRR2848264, SRR2848267, SRR2849323, SRR2849339, SRR2849358, SRR2849373, SRR2849384, SRR2849398, SRR2976582–SRR2976583, SRR2988260, SRR3225470–SRR3225471, SRR3275263–SRR3275264, SRR3502230, SRR3502520–SRR3502528, SRR3898627–SRR3898675, SRR3924383–SRR3924409, SRR5083564–SRR5083575, SRR5693565–SRR5693584, SRR5693645–SRR5693657, SRR576444, SRR576446, SRR576451, SRR576453–SRR576469, SRR5814033–SRR5814187, SRR6057892–SRR6057916, SRR6299285–SRR6299287, SRR6334371–SRR6334373, SRR7142301–SRR7142368, SRR7239923–SRR7239946, SRR7527146–SRR7527159, SRR7632639–SRR7632648, SRR7632653, SRR7632671–SRR7632680, SRR7648270, SRR7648273–SRR7648274, SRR7648284–SRR7648299, SRR7648310, SRR7648320–SRR7648321, SRR7648326–SRR7648327, SRR7648331–SRR7648339, SRR7648341, SRR7648343, SRR7648345–SRR7648350, SRR7668191, SRR7699187–SRR7699216, SRR8104593–SRR8104721, SRR8247196–SRR8247211, SRR8468235–SRR8468237, SRR8468246–SRR8468261, SRR8844064–SRR8844199, SRR9675236. Source data for Figs. 2–4 and Supplementary Figs. 13–17, 18e, 20, and 21 are provided with this paper.

Received: 23 February 2021; Accepted: 6 January 2022;

Published online: 10 February 2022

References

- Zehr, J. P. & Capone, D. G. Changing perspectives in marine nitrogen fixation. *Science* **368**, eaay9514 (2020).
- Karl, D. et al. The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature* **388**, 533–538 (1997).
- Wrightson, L. & Tagliabue, A. Quantifying the impact of climate change on marine diazotrophy: Insights from earth system models. *Front. Mar. Sci.* **7**, 952 (2020).
- Foster, R. A. & Zehr, J. P. Diversity, genomics, and distribution of phytoplankton-cyanobacterium single-cell symbiotic associations. *Annu. Rev. Microbiol.* **73**, 435–456 (2019).
- Villareal, T. A. Marine nitrogen-fixing diatom-cyanobacteria symbioses. In *Marine pelagic cyanobacteria: Trichodesmium and other diazotrophs* (eds Carpenter, E. J., Capone, D. G., & Rueter, J. G.) 163–175 (Springer, 1992).
- Harke, M. J. et al. Periodic and coordinated gene expression between a diazotroph and its diatom host. *ISME J.* **13**, 118–131 (2019).
- Zehr, J. P. et al. Unicellular cyanobacteria fix N₂ in the subtropical North Pacific Ocean. *Nature* **412**, 635–638 (2001).
- Thompson, A. W. et al. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**, 1546–1550 (2012).
- Stancheva, R. & Lowe, R. Diatom symbioses with other photoautotroph. In *Diatoms: Fundamentals and Applications* (eds Seckbach, J. & Gordon, R.) 225–244 (Wiley-Scrivener, 2019).
- Precht, J., Kneip, C., Lockhart, P., Wenderoth, K. & Maier, U.-G. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. *Mol. Biol. Evol.* **21**, 1477–1481 (2004).
- Kneip, C., Voss, C., Lockhart, P. J. & Maier, U.-G. The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. *BMC Evol. Biol.* **8**, 30 (2008).
- Nakayama, T. et al. Spheroid bodies in rhopalodiacean diatoms were derived from a single endosymbiotic cyanobacterium. *J. Plant Res.* **124**, 93–97 (2011).
- Kneip, C., Lockhart, P., Voss, C. & Maier, U.-G. Nitrogen fixation in eukaryotes—new models for symbiosis. *BMC Evol. Biol.* **7**, 55 (2007).
- Nakayama, T. et al. Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. *Proc. Natl. Acad. Sci. USA* **111**, 11407–11412 (2014).
- Nakayama, T. & Inagaki, Y. Genomic divergence within non-photosynthetic cyanobacterial endosymbionts in rhopalodiacean diatoms. *Sci. Rep.* **7**, 13075 (2017).
- Karl, D. M. & Lukas, R. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep-Sea Res. Part 2 Top. Stud. Oceanogr.* **43**, 129–156 (1996).
- Adler, S., Trapp, E. M., Dede, C., Maier, U.-G. & Zauner, S. *Rhopalodia gibba*: The first steps in the birth of a novel organelle? in *Endosymbiosis* (ed. Löffelhardt, W.) 167–179 (Springer Vienna, 2013).
- Keller, M. D., Selvin, R. C., Claus, W. & Guillard, R. R. L. Media for the culture of oceanic ultraphytoplankton. *J. Phycol.* **23**, 633–638 (2007).
- Thompson, A. et al. Genetic diversity of the unicellular nitrogen-fixing cyanobacteria UCYN-A and its prymnesiophyte host. *Environ. Microbiol.* **16**, 3238–3249 (2014).
- Cornejo-Castillo, F. M. et al. Cyanobacterial symbionts diverged in the late Cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat. Commun.* **7**, 11071 (2016).

21. Ruck, E. C., Nakov, T., Alverson, A. J. & Theriot, E. C. Phylogeny, ecology, morphological evolution, and reclassification of the diatom orders Surirellales and Rhopalodiales. *Mol. Phylogenet. Evol.* **103**, 155–171 (2016).
22. Ashworth, M. P. et al. Molecular and morphological investigations of the stauros-bearing, raphid pennate diatoms (Bacillariophyceae): *Craspedostauros* E.J. Cox, and *Stautotropis* T.B.B. Paddock, and their relationship to the rest of the Mastogloiales. *Protist* **168**, 48–70 (2017).
23. Sterrenburg, F. A. S. & Sterrenburg, F. J. G. An outline of the marine littoral diatom biocoenosis of the Banc d'Arguin, Mauritania, West Africa. *Botanica Marina* **33**, 459–465 (1990).
24. Lobban, C. S. Benthic marine diatom flora of Guam: new records, redescription of *Psammodyctyon pustulatum* n. comb., n. stat., and three new species (*Colliculoamphora gabgabensis*, *Lauderia excentrica*, and *Rhoicones pagoensis*). *Micronesica* **2015-02**, 1–49 (2015).
25. Hein, M. K., Winsborough, B. M. & Sullivan, M. J. *Iconographia Diatomologica* Vol. 19 (Gantner Verlag, 2008).
26. López-Fuerte, F. O. et al. Primeros registros de *Nanofrustulum shiloi* (Lee, Reimer & McEnery) Round, Hallsteinsen & Paasche y *Nitzschia nienhuisii* Sterrenburg & Sterrenburg (Bacillariophyceae; Ochrophyta) en aguas Mexicanas. *CICIMAR Océánides* **31**, 35–41 (2016).
27. Boeuf, D. et al. Biological composition and microbial dynamics of sinking particulate organic matter at abyssal depths in the oligotrophic open ocean. *Proc. Natl. Acad. Sci. USA* **116**, 11824–11832 (2019).
28. Poff, K. E., Leu, A. O., Eppley, J. M., Karl, D. M. & DeLong, E. F. Microbial dynamics of elevated carbon flux in the open ocean's abyss. *Proc. Natl. Acad. Sci. USA* **118**, e2018269118 (2021).
29. Berman-Frank, I., Quigg, A., Finkel, Z. V., Irwin, A. J. & Haramaty, L. Nitrogen-fixation strategies and Fe requirements in cyanobacteria. *Limnol. Oceanogr.* **52**, 2260–2269 (2007).
30. Mohr, W., Intermaggio, M. P. & LaRoche, J. Diel rhythm of nitrogen and carbon metabolism in the unicellular, diazotrophic cyanobacterium *Crocospaera watsonii* WH8501. *Environ. Microbiol.* **12**, 412–421 (2010).
31. Wilson, S. T. et al. Coordinated regulation of growth, activity and transcription in natural populations of the unicellular nitrogen-fixing cyanobacterium *Crocospaera*. *Nat. Microbiol.* **2**, 17118 (2017).
32. Muñoz-Marín, M. D. C. et al. The transcriptional cycle is suited to daytime N₂ fixation in the unicellular cyanobacterium 'Candidatus Atelocyanobacterium thalassa' (UCYN-A). *mBio* **10**, 1027 (2019).
33. Mareš, J. et al. Taxonomic resolution of the genus *Cyanothece* (Chroococcales, Cyanobacteria), with a treatment on *Gloeothece* and three new genera, *Crocospaera*, *Rippkaea*, and *Zehria*. *J. Phycol.* **55**, 578–610 (2019).
34. Foster, R. A. et al. Influence of the Amazon River plume on distributions of free-living and symbiotic cyanobacteria in the western tropical North Atlantic Ocean. *Limnol. Oceanogr.* **52**, 517–532 (2007).
35. Langlois, R. J., Hümmer, D. & LaRoche, J. Abundances and distributions of the dominant nifH phylotypes in the Northern Atlantic Ocean. *Appl. Environ. Microb.* **74**, 1922–1931 (2008).
36. Hashimoto, R., Watai, H., Miyahara, K., Sako, Y. & Yoshida, T. Spatial and temporal variability of unicellular diazotrophic cyanobacteria in the eastern Seto Inland Sea. *Fish. Sci.* **82**, 459–471 (2016).
37. Pfreundt, U., Spungin, D., Bonnet, S., Berman-Frank, I. & Hess, W. R. Global analysis of gene expression dynamics within the marine microbial community during the VAHINE mesocosm experiment in the southwest Pacific. *Biogeosciences* **13**, 4135–4149 (2016).
38. Pierella Karlusich, J. J. et al. Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *Nat. Commun.* **12**, 4160 (2021).
39. Nowack, E. C. M. & Weber, A. P. M. Genomics-informed insights into endosymbiotic organelle evolution in photosynthetic eukaryotes. *Annu. Rev. Plant Biol.* **69**, 51–84 (2018).
40. Webb, E. A., Moffett, J. W. & Waterbury, J. B. Iron stress in open-ocean cyanobacteria (*Synechococcus*, *Trichodesmium*, and *Crocospaera* spp.): Identification of the IdiA protein. *Appl. Environ. Microb.* **67**, 5444–5452 (2001).
41. Hynes, A. M., Webb, E. A., Doney, S. C. & Waterbury, J. B. Comparison of cultured *Trichodesmium* (Cyanophyceae) with species characterized from the field. *J. Phycol.* **48**, 196–210 (2012).
42. Moon-van der Staay, S. Y., De Wachter, R. & Vault, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001).
43. Scholin, C. A., Herzog, M., Sogin, M. & Anderson, D. M. Identification of group- and strain-specific genetic markers for globally distributed *Alexandrium* (Dinophyceae). II. Sequence analysis of a fragment of the LSU rRNA gene. *J. Phycol.* **30**, 999–1011 (1994).
44. Alverson, A. J., Jansen, R. K. & Theriot, E. C. Bridging the rubicon: phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. *Mol. Phylogenet. Evol.* **45**, 193–210 (2007).
45. Daugbjerg, N. & Andersen, R. A. Phylogenetic analyses of the *rbcL* sequences from haptophytes and heterokont algae suggest their chloroplasts are unrelated. *Mol. Biol. Evol.* **14**, 1242–1251 (1997).
46. Nübel, U., Garcia-Pichel, F. & Muyzer, G. PCR primers to amplify 16S rRNA genes from cyanobacteria. *Appl. Environ. Microb.* **63**, 3327–3332 (1997).
47. Mazard, S. L., Fuller, N. J., Orcutt, K. M., Bridle, O. & Scanlan, D. J. PCR analysis of the distribution of unicellular cyanobacterial diazotrophs in the Arabian Sea. *Appl. Environ. Microb.* **70**, 7355–7364 (2004).
48. Worden, A. Z. Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat. Microb. Ecol.* **43**, 165–175 (2006).
49. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
50. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Meth.* **9**, 772–772 (2012).
51. Ronquist, F. et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).
52. Stamatakis, A. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
53. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
54. Pruesse, E., Peplies, J. & Glöckner, F. O. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* **28**, 1823–1829 (2012).
55. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1114 (1999).
56. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
57. Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
58. Kalyanamoorthy, S. et al. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Meth.* **14**, 587–589 (2017).
59. Langlois, R. J., LaRoche, J. & Raab, P. A. Diazotrophic diversity and distribution in the tropical and subtropical Atlantic Ocean. *Appl. Environ. Microb.* **71**, 7910–7919 (2005).
60. Moisaner, P. H., Beinart, R. A., Voss, M. & Zehr, J. P. Diversity and abundance of diazotrophic microorganisms in the South China Sea during intermonsoon. *ISME J.* **2**, 954–967 (2008).
61. Kong, L., Jing, H., Kataoka, T., Sun, J. & Liu, H. Phylogenetic diversity and spatio-temporal distribution of nitrogenase genes (nifH) in the northern South China Sea. *Aquat. Microb. Ecol.* **65**, 15–27 (2011).
62. Bird, C. & Wyman, M. Transcriptionally active heterotrophic diazotrophs are widespread in the upper water column of the Arabian Sea. *FEMS Microbiol. Ecol.* **84**, 189–200 (2013).
63. Bombar, D., Kubo, K. A. T., Robidart, J., Carter, B. J. & Zehr, J. P. Non-cyanobacterial nifH phylotypes in the North Pacific Subtropical Gyre detected by flow-cytometry cell sorting. *Environ. Microbiol. Rep.* **5**, 705–715 (2013).
64. Martínez-Pérez, C. et al. The small unicellular diazotrophic symbiont, UCYN-A, is a key player in the marine nitrogen cycle. *Nat. Microbiol.* **1**, 1–7 (2016).
65. Chen, M. et al. Biogeographic drivers of diazotrophs in the western Pacific Ocean. *Limnol. Oceanogr.* **64**, 1403–1421 (2019).
66. Liu, J. et al. Effect of mesoscale eddies on diazotroph community structure and nitrogen fixation rates in the South China Sea. *Reg. Stud. Mar. Sci.* **35**, 101106 (2020).
67. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
68. Leinonen, R., Sugawara, H. & Shumway, M., International Nucleotide Sequence Database Collaboration. The sequence read archive. *Nucleic Acids Res.* **39**, D19–D21 (2011).
69. Villar, E. et al. The Ocean Gene Atlas: exploring the biogeography of plankton genes online. *Nucleic Acids Res.* **46**, W289–W295 (2018).
70. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
71. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
72. Mueller, J. A., Culley, A. I. & Steward, G. F. Variables influencing extraction of nucleic acids from microbial plankton (viruses, bacteria, and protists) collected on nanoporous aluminum oxide filters. *Appl. Environ. Microb.* **80**, 3930–3942 (2014).
73. Wilson, S. T., Caffin, M., White, A. E. & Karl, D. M. Evaluation of argon induced hydrogen production as a method to measure nitrogen fixation by cyanobacteria. *J. Phycol.* **57**, 863–873 (2021).

Acknowledgements

This work was supported by National Science Foundation awards (OIA 1736030 and OCE 15-59356 to G.F.S.; OCE 1756524 to S.T.W) and the Simons Foundation (#329108 to D.M.K. and A.E.W. and #811977 to S.T.W.). We thank Tina M. Weatherby and Marilyn F. Dunlap at the University of Hawai'i at Mānoa Biological Electron Microscopy

Facility for their assistance with transmission and scanning electron microscopy, Natalie Wallsgrove at the Stable Isotope Facility for her support, and the Hawaii Ocean Time-series program for their assistance with sample collection at Station ALOHA (NSF OCE 1756517 to D.M.K. and A.E.W.). The technical support and advanced computing resources from the University of Hawai'i Information Technology Services – Cyberinfrastructure, funded in part by NSF MRI 1920304, are gratefully acknowledged.

Author contributions

C.R.S., S.T.W., M.C., Q.L., and G.F.S. contributed to experimental design. C.R.S. isolated and maintained both strains of *Epithemia* and conducted the light and electron microscopy analysis and phylogenetic analysis. C.R.S. and Q.L. conducted gene sequencing and qPCR. C.R.S. with input from K.A.T.-K. conducted the metagenomic analysis. S.T.W. and M.C. conducted the N₂ fixation measurements. R.S. contributed species descriptions and light microscopy imaging of cleaned diatom frustules. All authors, including C.R.S., S.T.W., M.C., R.S., Q.L., K.A.T.-K., A.E.W., D.M.K., J.P.Z., and G.F.S., contributed to the interpretation of results. C.R.S. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-28065-6>.

Correspondence and requests for materials should be addressed to Christopher R. Schvarcz.

Peer review information *Nature Communications* thanks Lucas Stal, Matt Ashworth, Rosalina Stancheva Hristova and the other anonymous reviewer(s) for their contribution to the peer review this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022