

## The Bay of Bengal exposes abundant photosynthetic picoplankton and newfound diversity along salinity-driven gradients

Jan Strauss<sup>1\*</sup>, Chang Jae Choi<sup>1\*</sup>, Jonathan Grone<sup>1,2</sup>, Fabian Wittmers<sup>1,2</sup>, Valeria Jimenez<sup>1</sup>, Kriste Makareviciute-Fichtner<sup>1</sup>, Charles Bachy<sup>1,3</sup>, Gualtiero Spiro Jaeger<sup>4</sup>, Camille Poirier<sup>1</sup>, Charlotte Eckmann<sup>1,5</sup>, Carolin R. Löscher<sup>6</sup>, VVSS Sarma<sup>7</sup>, Amala Mahadevan<sup>4</sup> & Alexandra Z. Worden<sup>1,2,5</sup>

\*Co-first authors

Corresponding author: Alexandra Z. Worden, azworden@geomar.de

<sup>1</sup>Ocean EcoSystems Biology Unit, RD3, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

<sup>2</sup>Faculty of Mathematics and Natural Sciences, Christian-Albrecht University of Kiel, Kiel, Germany

<sup>3</sup>Sorbonne Université, CNRS, FR2424, Station biologique de Roscoff, Roscoff, 29680, France

<sup>4</sup>Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

<sup>5</sup>University of California Santa Cruz, Santa Cruz, CA 95064, USA

<sup>6</sup>Nordcee, DIAS, Department of Biology, University of Southern Denmark, Odense, Denmark

<sup>7</sup>CSIR-National Institute of Oceanography, Regional Centre, Visakhapatnam, India

### Summary

The Bay of Bengal is a 2,600,000 km<sup>2</sup> expanse in the Indian Ocean upon which many humans rely for food. However, the phytoplankton underpinning food chains here are poorly characterized. We examined phytoplankton abundance and diversity along a transect exhibiting strong salinity gradients (fresher northwards), a salinity-stratified photic zone, and  $28.4 \pm 1.1^\circ\text{C}$  temperatures. In surface waters, *Prochlorococcus*, predominantly HLII, averaged  $117,005 \pm 44,014$  cells ml<sup>-1</sup>. Cyanobacterial amplicons in the subsurface chlorophyll maximum (SCM) were dominated by LLII and ‘rare’ *Prochlorococcus* ecotypes, HLVI and LLVII. *Synechococcus* averaged  $83,622 \pm 23,373$  cells ml<sup>-1</sup>, declined markedly with depth, and population structure of the dominant clade, Clade II, differed between surface and SCM, with Clade X being notable at both. Photosynthetic eukaryote surface communities differentiated between southern (prasinophytes, especially *Ostreococcus*) and northern (stramenopiles, specifically, diatoms, pelagophytes dictyochophytes, plus the prasinophyte *Micromonas*) stations. Across the transect, the SCM was almost exclusively *Ostreococcus* Clade OII, and eukaryotic abundances ranged south-to-north from 19,447 to 4,162 cells ml<sup>-1</sup>. Additionally, at the southernmost stations, OII and novel *Ostreococcus bengalensis* rivaled each other through the photic zone. Our results highlight unique characteristics of Bay of Bengal phytoplankton communities, with dominance of a single eukaryotic taxon and hitherto ‘rare’ cyanobacteria at depth, unlike characterized temperature-stratified subtropical gyres.

**Keywords:** Bay of Bengal; plankton diversity; flow cytometry; 16S rRNA sequence variants; *Prochlorococcus*; *Synechococcus*; *Ostreococcus*

## Introduction

The Bay of Bengal is a unique ocean region that is strongly influenced by freshwater influx from the major rivers that drain the Indian subcontinent and the southern slopes of the Himalaya, as well as heavy rainfall during the summer monsoon (Subramanian, 1993; Mahadevan, 2016). Unlike other large tropical ocean regions, it is also a semi-enclosed basin and thus exceptionally fresh due to the freshwater input, which amounts to  $\sim 1.5 \times 10^{12} \text{ m}^3 \text{ year}^{-1}$  (Gomes *et al.*, 2000). Thus, the Bay of Bengal is low in salinity and density at the surface, factors that inhibit the upward transport of nutrients (like nitrate) for phytoplankton and downward flux of oxygen for zooplankton and other organisms, including microbes (Sarma *et al.*, 2016). The nutrient-depleted surface and oxygen-deprived depths contribute to a mystery about how the levels of phytoplankton and zooplankton production are adequately sustained to support the active fisheries in the Bay of Bengal (Paulmier and Ruiz-Pino, 2009; McCreary *et al.*, 2013). Indeed, human reliance on a particular ocean region is probably nowhere greater than around the Bay of Bengal, where some of the densest populations in India, Bangladesh, Myanmar and Sri Lanka rely on it for food and on bay-influenced monsoonal dynamics for water.

The Bay of Bengal is not only one of the freshest oceanic regions in the world, it also has complex physicochemical properties and circulation (Spiro Jaeger and Mahadevan, 2018). Strong lateral salinity gradients form a multitude of density fronts within the surface layer, leading to instabilities over a range of spatial and temporal scales (Sarkar *et al.*, 2016). Additionally, the fresh, less dense water at the surface results in an intense salinity-driven vertical stratification in the near-surface layer that inhibits vertical mixing (Shroyer *et al.*, 2020). The southeastern portions are influenced by the Arabian Sea, which is considered more biologically productive than the Bay of Bengal (Prasanna Kumar *et al.*, 2002; Madhupratap *et al.*, 2003; Thushara *et al.*, 2019). With respect to nutrients, although the riverine flux brings in nutrients, a substantial portion is consumed within a few kilometers from the coast (Sarma *et al.*, 2013; Krishna *et al.*, 2016). Nutrient limitation is exacerbated by the strong salinity gradients (Madhupratap *et al.*, 2003), which unlike vertical stratification in other tropical and subtropical oceans where seasonal nutrient limitation is seen (Giovannoni and Vergin, 2012) are substantially independent of thermal and other ecologically important gradients, i.e. although the surface waters are warm ( $\sim 26\text{-}29^\circ\text{C}$ ). Furthermore, strong stratification also contributes to a quiescent environment, which may further limit the overall nutrient supply to the system, resulting in lower productivity (Prasanna Kumar *et al.*, 2002; Sarker *et al.*, 2020). Thus, in many respects, salinity is the leading control on the dynamics of the photic zone, and shapes its unique, shallow, spatially and temporary variable mixed layer of oligotrophic

conditions (Mahadevan *et al.*, 2016) for which there is still much to learn with respect to the phytoplankton taxa that underpin productivity in the system.

The major differences between conditions in the Bay of Bengal and those at the Bermuda Atlantic Time-series Study (BATS) or the Hawaii Ocean Time Series (HOTS, Station ALOHA) (Giovannoni and Vergin, 2012; Karl and Church, 2014) make understanding through direct analogy to primary producers in these subtropical sites unwieldy. Hence, it is essential that the photosynthetic organisms that dictate primary production in the Bay of Bengal and form the base of the marine food chain there be elucidated within the context of the complex physicochemical dynamics in this ecosystem. Furthermore, climate change will have multiple effects in the Bay of Bengal that are difficult to quantify, especially in absence of comprehensive characterization of the primary producer communities within (Roxy *et al.*, 2016; Gregg and Rousseaux, 2019; Rixen *et al.*, 2020; Löscher, 2021). Plankton communities at the southern margin of the Bay of Bengal, an area strongly influenced by the high salinity waters of the Arabian Sea, have been reported to be dominated by the cyanobacteria *Synechococcus* and *Prochlorococcus* at the surface based on V4-V5 rRNA amplicon data (Angelova *et al.*, 2019). Based on assignment by broad placement tools, eukaryotic plankton there reportedly had high relative abundances of the picoprasinophyte *Ostreococcus tauri*, followed by the diatom *Ditylum brightwellii*, pelagophytes and prymnesiophytes within the Braarudosphaeraceae (Angelova *et al.*, 2019). However, while these studies shed light on particular phytoplankton community members, it is critical to understand the phytoplankton community as a whole because of its the importance to the Bay of Bengal ecosystem, as well as the ecosystem services provided to human communities in the coastal part of the delta, especially given that this ecosystem will be sensitive to changes in the global water cycle.

We examined planktonic microbial communities in the Bay of Bengal during the peak southwest summer monsoon in 2015 (Wijesekera *et al.*, 2016). The Bay of Bengal receives the highest rainfall in the world during the summer monsoon and, as a result, it is one of the freshest subtropical oceanic regions, making this system unique compared to the other similar latitude oceans. Samples collected from 18 profiles were analyzed by flow cytometry at multiple depths up to 200 m depth, and by V1-V2 16S rRNA gene amplicon sequencing at two depths (surface and subsurface chlorophyll maximum) using phylogenetic approaches for fine-scale phytoplankton distribution analysis. We further developed high-resolution full-length 16S rRNA gene phylogenetic reconstructions for two major planktonic microbes in this system, the cyanobacterium *Prochlorococcus* and the picoeukaryotic *Ostreococcus* to provide insights into new or rarely

detected members and their spatiotemporal patterns in connection with environmental conditions. Collectively, our results reveal truly unique community structure as compared to well-characterized thermally-stratified photic zones in the subtropics, and indicate that there is a wealth of newfound diversity in the Bay of Bengal that aligns with its unusual properties.

## Results

### *Oceanographic conditions along the Bay of Bengal transect*

The Bay of Bengal is a complex ecosystem where phytoplankton have been characterized by microscopy, HPLC, and more recently amplicon sequencing (Jyothibabu *et al.*, 2015, 2018; Angelova *et al.*, 2019; Sarma *et al.*, 2020). However, often these studies have focused on a specific group or did not have the power to resolve the overall community, hence data for primary producer communities as a whole in this region are sparse (Fig. 1A). Our cruise extended from the southwestern to northeastern Bay of Bengal, and sampled various physical and chemical regimes (highly stratified, monsoon influenced, oxygen minimum zone boundaries, riverine nutrient input regions). Within the top 10 m, temperatures were similar across the transect ( $29.1 \pm 0.2^\circ\text{C}$ ; Fig. S1, Table S1), while the salinity declined from 34.2 to 29.4 moving northwards (Fig. 1B). Stations S5 and S7-18 had a clear lens of low salinity waters in the upper 10 m overlying more saline waters. This large differential lessened in the southward direction, such that surface salinity was significantly higher at Stations S1-S4 ( $33.3 \pm 0.8$ ) than Stations S5-S18 ( $30.6 \pm 0.7$ ;  $p < 0.001$ , Wilcoxon test). The vertical extent of the photic zone itself varied substantially over the transect, with the 1% of surface photosynthetic active radiation (PAR) level being just above or below 50 m at stations S1-S2, S9, and S13-S16, while at S3-S8, S10 and S12 it varied around 75 m depth and was deepest at S17-S18 (close to 100 m). Thus, temperature varied considerably ( $20.2 - 29.5^\circ\text{C}$ ) over photic zone depths across the transect, but it was less variable ( $>27^\circ\text{C}$ ) at the highest *in vivo* chlorophyll *a* fluorescence layer and shallower depths. The overall range in salinity over photic zone depths (34.9 - 29.4) was akin to that along surface waters (Fig. 2A, Fig. S1). Vertical distribution of temperature and salinity across stations indicated that S1-S4 had distinct water mass characteristics from other stations, being saltier and warmer at a given density and also having significantly lower dissolved oxygen at a given depth as compared to S5-S18 (Fig. S1). Additionally, densities at S1-S4 had commonalities to each other that differed from the other stations (Fig. S1).

*In vivo* chlorophyll *a* fluorescence showed that a subsurface chlorophyll maximum (SCM) was present at all stations, ranging from between 30 to 60 m (Fig. 2B). Chlorophyll fluorescence

followed patterns in nutrients which varied with depth in the photic zone and along the transect (Fig. S2, Table S1). Dissolved inorganic nitrogen (DIN, sum of nitrate, nitrite and ammonium) concentrations in surface waters varied from 0.7  $\mu\text{M}$  (Station 1) to 4.6  $\mu\text{M}$  (Station 15). With depth, nutrient concentrations generally increased such that in the SCM DIN averaged  $8.6 \pm 4.0 \mu\text{M}$ . Similarly, silicate concentrations were  $2.1 \pm 0.9 \mu\text{M}$  and  $4.0 \pm 2.4 \mu\text{M}$  in surface waters and the SCM, respectively. Silicon to nitrogen (Si:N) ratios were on average lower in SCM than the surface, with the exception of Stations S18-S19 in the surface (Table S1). Phosphate concentrations averaged  $0.4 \pm 0.2 \mu\text{M}$  in the surface waters and  $0.6 \pm 0.2 \mu\text{M}$  in the SCM and were generally lower in the northernmost stations. Finally, in general, oxygen was significantly higher and salinity lower in the surface compared to the underlying waters, again apart from southern stations.

#### *Abundances of major phytoplankton groups and variations cell size*

In total, four to six depths were analyzed by flow cytometry, typically ranging from 1 to 100 m and corresponding to (1) surface, (2) mixed layer, (3) SCM, (4) below SCM, (5) base of oxycline and sometimes an anoxic layer, if reached (Table S1). Eleven surface samples collected using a bucket (when the CTD was lost overboard; Fig. S3) were also analyzed.

Surface waters were numerically dominated by cyanobacteria, with abundances of *Prochlorococcus* and *Synechococcus* averaging  $104,341 \pm 45,640 \text{ cells ml}^{-1}$  and  $83,622 \pm 23,373 \text{ cells ml}^{-1}$  (n=18), respectively, across sites (Fig. 2C). *Synechococcus* dropped dramatically in cell abundance, typically by depths of 15-20 m. Thus, by the SCM, *Prochlorococcus* averaged  $60,140 \pm 21,277 \text{ cells ml}^{-1}$  (n=18), while *Synechococcus* averaged  $8298 \pm 15,562 \text{ cells ml}^{-1}$  (n=18), with the large standard deviation in the latter arising from the fact that 7 stations showed  $<1000 \text{ Synechococcus ml}^{-1}$  (i.e.,  $551 \pm 143 \text{ cells ml}^{-1}$ ) at the SCM. Three factors stood out from these analyses; first, the peak abundance of *Prochlorococcus* was typically between 14 to 30 m, not at the surface or SCM. At station S1 the highest abundances of *Prochlorococcus* were seen at 63 m ( $50,256 \text{ cells ml}^{-1}$ ) and 72 m ( $40,734 \text{ cells ml}^{-1}$ ), not higher in the photic zone. Finally, at stations S7 to S18 *Prochlorococcus* maintained high cellular abundances down to 100 m (the deepest sample collected), although the SCM was at 50 m and both *Synechococcus* as well as photosynthetic eukaryotes behaved more classically and had dropped in abundance below the SCM (Fig. 2C, rightmost panel).

Eukaryotic phytoplankton averaged  $2075 \pm 796 \text{ cells ml}^{-1}$  (n=17) in the surface with Station S1 excluded due to it having much higher numbers ( $12,027 \text{ eukaryotic phytoplankton cells ml}^{-1}$ ). At all stations across the transect, eukaryotic phytoplankton cell abundances manifested a peak

coincident with the depth of the SCM, in which they averaged  $10,655 \pm 4589$  cells  $\text{ml}^{-1}$  ( $n=18$ ; Fig. 2C). Station S1 exhibited a SCM but had high eukaryotic cell abundances at both the surface and SCM.

We also examined mean Forward Angle Light Scatter (FALS), a rough indicator of cell size, for each of these broad phytoplankton groups. As expected, eukaryotic phytoplankton exhibited bead normalized FALS values of  $276.1 \pm 122.6$  that were significantly greater than either *Prochlorococcus* ( $1.7 \pm 1.0$ ) or *Synechococcus* ( $7.4 \pm 2.8$ ;  $p < 0.001$ , Wilcoxon test) at the surface (Table S1). At other depths these trends generally persisted; however, the magnitude of differences was dampened at depth, for example at the SCM, eukaryotic phytoplankton, *Prochlorococcus*, and *Synechococcus* exhibited normalized mean FALS values of  $101.2 \pm 25.8$ ,  $18.3 \pm 5.4$ , and  $4.6 \pm 2.1$ , respectively (Table S1). Furthermore, eukaryotic FALS in the surface waters increased within the transect from south to north.

Our results this far indicated remarkable differences in phytoplankton cell characteristics. However, the extent to which these were attributable to acclimation to light availability, differences in speciation, or both, was unclear. Thus, we generated 16S V1-V2 amplicons and leveraged preexisting reference alignments for analysis of *Synechococcus* and photosynthetic eukaryotes (Ahlgren and Rocop, 2012; Vergin *et al.*, 2013; Sudek *et al.*, 2015; Choi *et al.*, 2017, 2020). We also developed a preliminary reference alignment for diatoms and a refined alignment for *Prochlorococcus* that incorporated the full-length 16S rRNA gene sequences generated herein (Fig. 3A, Fig. S5).

#### *Distribution of Cyanobacterial Ecotypes and Oligotypes*

Cyanobacteria contributed  $39 \pm 7\%$  of all 16S amplicons (including heterotrophic bacteria and photosynthetic eukaryotes) at the surface and  $15 \pm 5\%$  at the SCM. *Prochlorococcus* and *Synechococcus* contributed similarly to the surface community, forming  $47 \pm 16\%$  and  $49 \pm 15\%$  of the total cyanobacteria amplicons, respectively ( $n=18$ ) except for the station S1 where *Synechococcus* alone contributed 83% of cyanobacterial amplicons (Fig. S4). *Prochlorococcus* dominance was observed at the SCM where it represented  $93 \pm 6\%$  of cyanobacteria ( $n=15$ , S1-S2 excluded; Fig. S4). In the southernmost stations S1 and S2, *Synechococcus* SCM amplicon contributions to cyanobacteria ( $57 \pm 10\%$ ) were higher than *Prochlorococcus* ( $42 \pm 10\%$ ; Fig. S4). These patterns were further dissected by examining the specific ecotypes and clades present. Within *Prochlorococcus*, the HLII ecotype dominated the surface water community, comprising  $99 \pm 1\%$  ( $n=18$ ) of *Prochlorococcus* amplicons. HLII contributed much less at the SCM ( $6 \pm 6\%$ ,

n=15), except at stations S2 and S4 where it contributed  $62 \pm 4\%$  (n=2) at the SCM (Fig. 3A). At the SCM the HLVI ecotype formed  $36 \pm 12\%$  of *Prochlorococcus* amplicons, with up to 66% at station S1, and across all stations was followed by LLII ecotype ( $27 \pm 12\%$ ). LLVII, one of the most diverged *Prochlorococcus* ecotypes (Fig. 3A), contributed  $15 \pm 10\%$  of SCM *Prochlorococcus* amplicons across the stations, with up to 31% at the northernmost station S18 and comprised 13 ASVs. Systematic or significant changes in amplicon relative abundances over the salinity gradient were not observed, likely because that gradient was across surface waters where LLVII was very low in relative abundance - and rather HLII was the dominant ecotype.

Within *Synechococcus*, Clade II dominated the surface waters by representing  $88 \pm 5\%$  of the total *Synechococcus* amplicons followed by *Synechococcus* Clade X by  $6 \pm 5\%$  (Fig. 3B). Increased contributions from Clade X were observed from the SCM forming  $46 \pm 19\%$  (n=15 excluding Stations S2 and S4,  $42 \pm 21\%$  based on all stations) of the total *Synechococcus* amplicons except for the stations S2 and S4 where it contributed  $10 \pm 8\%$ . Contributions from Clade II were still high, averaging  $43 \pm 19\%$  (n=15 excluding stations S2 and S4 where it represented  $84 \pm 11\%$ ).

*Prochlorococcus* HLII and *Synechococcus* Clade II population biology was then dissected using oligotyping to determine whether there might be more subtle distributional patterns over the transect (Fig. 3). Multiple oligotypes were observed in the HLII amplicons, with two dominants, oligotypes A and B, representing  $56 \pm 4\%$  and  $23 \pm 1\%$  of the total HLII amplicons, respectively. At the SCM oligotype B dominated HLII, representing  $80 \pm 5\%$  (n=15) of HLII amplicons followed by oligotype A ( $11 \pm 3\%$ , n=15). Again stations S2 and S4 differed, with oligotype A and oligotype B having roughly equivalent contributions. Within *Synechococcus* Clade II, oligotype A dominated the surface, comprising  $95 \pm 1\%$  of the Clade II amplicons. However, increased contributions from oligotype B were observed at the SCM, averaging  $50 \pm 13\%$  except for stations S1-S4 where oligotype A represented  $90 \pm 9\%$  of Clade II amplicons.

#### *Distribution and diversity of eukaryotic phytoplankton*

Overall distributions of eukaryotic phytoplankton were markedly different at the surface and SCM, with higher diversity in the former. Two major groups, stramenopiles and green algae (primarily prasinophytes), dominated total plastid-derived amplicon relative abundances at the surface and the SCM, but differently (Fig. 4A,B). The surface community was mostly comprised of stramenopiles forming  $70 \pm 7\%$  (n=14, stations S5-S18) of plastid amplicons, except for the four southernmost stations S1-S4 where prasinophytes contributed  $69 \pm 19\%$  (n=4) while stramenopiles accounted for  $26 \pm 15\%$ . In particular, station S1 was massively dominated by prasinophytes (91%

of plastid amplicons; Fig. 4A). Otherwise, contributions from all green algae at the surface were lower for stations S5-S18 ( $17 \pm 9\%$  of plastid amplicons). Clear shifts between the stations S1-S4 and the remaining stations S5-S18 were not observed from eukaryotic groups with lower relative abundances ( $8 \pm 4\%$ , prymnesiophytes, which were closest to *Chrysochromulina* sp., but still distant at 95% nt identity;  $<1\%$ , cryptophytes;  $2 \pm 2\%$ , other eukaryotic phytoplankton;  $n=18$ ). At the SCM, prasinophytes dominated eukaryotic phytoplankton amplicons ( $83 \pm 5\%$ ), while stramenopiles dropped to  $15 \pm 5\%$  and amplicons from other eukaryotic groups (prymnesiophytes, cryptophytes, and others) were near detection levels (Fig. 4B).

Among photosynthetic stramenopiles, diatoms contributed  $58 \pm 11\%$  of the stramenopile amplicons in the surface, followed by dictyochophytes and pelagophytes forming  $19 \pm 6\%$  and  $16 \pm 8\%$ , respectively (Table S1). At the SCM, pelagophytes had higher relative contributions, forming  $57 \pm 11\%$  of stramenopile amplicons whereas diatoms and dictyochophytes formed  $33 \pm 11\%$  and  $6 \pm 2\%$ , respectively, with most of the latter being from a single genus, *Florenciella*, regardless of depth. Other stramenopiles, including chrysophytes, represented  $<5\%$  of stramenopile amplicons (Table S1). We further analyzed diatoms and pelagophytes using reference trees for high-resolution analysis of these taxa (Choi *et al.*, 2020). The diatoms with the highest relative abundance belonged to clade 32, which includes the pennate diatom *Tryblionella apiculata* ('unaccepted' synonym *Nitzschia apiculata*) (Fig. S5). The dominant pelagophytes belonged to the *Pelagomonas calceolata* Clade (PcC) and Pelagophyte Environmental Clade VIII, PEC-VIII, both of which had distinct spatial distributions vertically, and along the transect. PcC contributed the most at stations S1-S6 in the surface, forming  $77 \pm 10\%$  ( $n=6$ ) of pelagophyte amplicons, whereas at S7-S18 PEC-VIII contributed  $76 \pm 8\%$  of pelagophyte amplicons. Pelagophyte amplicons at the SCM were largely assigned to PcC, forming  $97 \pm 2\%$  of pelagophyte amplicons, and the rest were mostly PEC-VIII ( $2 \pm 2\%$ ) (Fig. S6).

Class II prasinophytes dominated the SCM and the southernmost stations throughout the water column, and were generally picoplanktonic species. In fact, *Ostreococcus* Clade OII dominated the SCM, comprising  $90 \pm 5\%$  of Viridiplantae (green algal) amplicons and  $75 \pm 8\%$  of eukaryotic phytoplankton. *Bathycoccus* and *Micromonas* clades A/B (represented by genome sequenced species *M. commoda*) were also detected, albeit at lower relative abundances. In stations S5-18 stramenopiles dominated plastid-derived amplicons in surface waters, whereas at S1-S4 prasinophytes dominated (Fig. 4A). Stations S1-S2 again exhibited patterns differing from elsewhere at the surface. However, at other stations (S3-18), contributions from Clade *Ostreococcus* OII were low in the surface ( $9 \pm 12\%$  of Viridiplantae,  $n=16$ ), and instead



*Micromonas* clades A/B exhibited the highest relative abundances among green algal amplicons, together forming  $56 \pm 18\%$  (S3-S18, n=16) of surface green algal amplicons.

The near complete dominance of *Ostreococcus* Clade OII at the SCM raised questions about possible microdiversity within this clade. We found Clade OII comprised various oligotypes with three dominant oligotypes (A to C, Fig. 4C). OII oligotype A dominated SCM OII amplicons, forming  $94 \pm 3\%$  (n=14, Stations S5-S18), but less at stations S1-S4 ( $56 \pm 17\%$ ). In the latter, increased contributions from oligotypes B and C were observed, forming  $37 \pm 19\%$  and  $6 \pm 4\%$  of OII amplicons, respectively. At the surface, oligotype A generally contributed the most, comprising  $55 \pm 6\%$  (S4-S6, n=3) and  $92 \pm 13\%$  (S7-S18, n=13) of Clade OII amplicons. Oligotype A contributions were less in the three southernmost stations ( $20 \pm 6\%$  S1-S3, n=3), and instead oligotypes B ( $65 \pm 5\%$  at S1-S3, n=3) dominated and C was also notable.

#### *Identification of a novel Ostreococcus clade*

As shown by broad phytoplankton abundances and distributions, phytoplankton diversity patterns and physical parameter characteristics of the southern stations were markedly different from those closer to freshwater inputs in more northern parts of the transect. Amplicon placement in the prasinophyte subtree showed a group of LCA (Last Common Ancestor) sequences at an *Ostreococcus* node. These amplicons ranged from 91-95% nt identity to known *Ostreococcus* species over the 289 nt of the V1-V2 16S region, 14 nt difference from Clade A, 20 nt from Clade B, 25 nt from Clade C, and 26 nt from Clade D. Because the LCA group represented up to 29% of green algal amplicons, we searched GenBank nr and recovered six full-length 16S rRNA gene sequences with 100% identity to the LCA V1-V2 16S amplicons. These came from the coast of Georgia, USA (e.g., FJ744977) (Poretsky *et al.*, 2010), waters near a dolphin in San Diego, CA (e.g., JQ196239), and a bay around the Liaodong Peninsula in Northeast China (EU259787). Phylogenetic reconstruction using the newly recovered 16S rRNA genes showed that they comprised a bootstrap supported clade (ML bootstrap support = 100% out of 1,000 replicates, Bayesian Posterior Probability = 1) distinct from other *Ostreococcus* clades/species (Fig. S7-8). To validate that this newfound diversity is distinct from previously described microbial diversity within eukaryotic algae, and particularly prasinophytes, described in environmental 16S rRNA gene clone libraries in the Arabian Sea (Fuller *et al.*, 2006) and V4 16S amplicon sequencing in the Bay of Bengal (Angelova *et al.*, 2019), we performed additional phylogenetic analysis using the recovered full-length sequences from our study. Here, we also incorporated our representative V1-V2 16S rRNA amplicon sequence from the newfound LCA *Ostreococcus* clade as well as sequences from the prior studies (Fig. S8). Sequences from the Arabian sea and the prior Bay of

Bengal study clustered within *Ostreococcus* Clade B/OII, which included the cultured species *Ostreococcus* sp. RCC809 and *Ostreococcus* RCC393, whereas the V1-V2 ASVs clustered within statistically supported Clade E (ML bootstrap support = 100%, out of 1,000 replicates, Bayesian Posterior Probability = 1; Fig. S8). Within the V4 region of the 16S rRNA gene there were 1 to 4 nt differences between members of *Ostreococcus* Clade E.

Due to recovery of a clearly distinct clade in the *Ostreococcus* 16S rRNA gene phylogeny, we undertook sequencing of the complete 18S-ITS1-5S-ITS2 through partial 28S, which has been used to delineate Class II prasinophyte species previously (Marin and Melkonian, 2010). To accomplish this, we developed a forward primer specific to *Ostreococcus* to enrich for sequences from this genus, and applied it to Bay of Bengal samples where the novel LCA group was abundant. Maximum likelihood phylogenetic analysis of the resulting full-length 18S rRNA gene sequences again established a statistically supported clade (Maximum likelihood bootstrap support = 98% out of 1,000 replicates, Bayesian Posterior Probability = 1) that was distinct from the other *Ostreococcus* and from environmental full-length 18S rRNA gene sequences (Fig. 5A). Phylogenetic analysis of the ITS supported these results (Fig. S7). We next constructed the ITS2 secondary structure, which contained the extra helix so far only seen in *Ostreococcus* and *Bathycoccus* (Fig. 5B). Comparison of this helix across the known *Ostreococcus* species and the novel LCA group showed that it contained differences that were not present in the others (Fig. 5C) pointing to designation as a new species, which we propose be informally named *Ostreococcus bengalensis* nom. nud.

To this point the only other *O. bengalensis* full-length sequences recovered from GenBank had little contextual information apart from location. For Georgia and San Diego sequences, temperature records from other sources for the period of collection indicated the water in May would be ~24°C, reaching 29°C by July (Georgia), and averaging 20°C in July (San Diego, although the sequence may be from an enclosed tank or the natural environment based on the nr record). To further examine the environmental distribution of the novel species, we turned to TARA Oceans V9 data, but found this variable region unsuitable for resolving *Ostreococcus* species, as reported previously (Monier *et al.*, 2016). We then turned to V4 18S rRNA amplicon data from Ocean Sampling Day, TARA Oceans and Malaspina to examine the distribution of *O. bengalensis*, alongside the other two species of importance in oceanic environments. A caveat in these analyses is that the relative abundances will be subject to changes in any and all other eukaryotic taxa, whether phytoplanktonic or not, in each sample. Moreover, oddly and despite this caveat, the relative abundances we observed using MetaPR2 were sometimes extraordinarily high

in some Ocean Sampling Day samples (e.g., OSD111, surface water off Portugal), especially in contrast to a recent V9 analysis of TARA ASVs (Bachy *et al.*, 2022). Still from a distributional perspective, valuable insights were gained in that aside from its high abundance in the Bay of Bengal, *O. bengalensis* appears in V4 data in North Atlantic and continental shelf seas such as the Bay of Biscay, the Adriatic Sea and the Strait of Gibraltar. Additionally, it is seen at lower relative abundances in the Rio de la Plata plume of the South Atlantic and in the Singapore Strait (Fig. 6A). These results contrasted with the distributions of *O. lucimarinus*, which exhibited highest relative abundances in some continental shelf regions and in upwelling zones in the eastern Pacific (Fig. 6C). Lastly, OII (the dominant *Ostreococcus* in the Bay of Bengal 16S amplicons) was observed at open ocean sites of the Atlantic and Pacific (Fig. 6B).

#### *Statistical analysis of communities and potential environmental drivers*

Statistical analyses showed, not surprisingly, that the taxonomic composition between the surface and SCM significantly differed for both cyanobacteria and eukaryotic phytoplankton (ANOSIM test statistic  $R=0.937$ ,  $p=0.001$ ). Ordination analyses and data projections using non-metric multidimensional scaling and compositional data analysis supported that microbial communities were most different at the southernmost stations S1 - S4, especially at the surface (Fig. 4A,C). Additionally, compositional data analysis confirmed that the cyanobacterial ecotypes *Synechococcus* Clade II and *Prochlorococcus* HLII most strongly influenced taxonomic compositions at the surface (Fig. 3A,B), whereas the *Prochlorococcus* ecotypes HLVI and LLVII as well as the eukaryotic phytoplankton *Ostreococcus* Clade OII and *P. calceolata* Clade (PcC) most strongly affected taxonomic compositions at the SCM (Fig. 3, Fig. 4, Fig. S6). As shown by a permutation test (999 permutations) for canonical correspondence analysis (CCA), the ordination was significant and the environmental variables salinity (test statistic = 3.82,  $p=0.001$ ), nitrite (test statistic = 4.84,  $p=0.001$ ) and nitrate (test statistic = 1.54,  $p=0.048$ ) were significant, explaining most variances in the observed ASV matrix (Fig. S9). As indicated by the environmental vectors and ASV pattern, distinct communities at the surface and the SCM became apparent again that were most strongly driven by higher temperatures, oxygen and light and higher salinities and nutrients, respectively (Fig. S9). *Prochlorococcus* HLII is most strongly driven by increasing PAR. Further, Spearman's rank correlation analyses between relative abundances of cyanobacterial ecotypes and oligotypes of *Prochlorococcus* HLII and environmental variables indicated that the most abundant HLII oligotype, oligotype B, exhibited significantly positive correlations with salinity ( $r(34)=0.70$ ), nitrite ( $r(36)=0.69$ ) and nitrate ( $r(35)=0.59$ ) at a significance level of  $p<0.001$ . In contrast, HLII oligotypes A and C showed negative Spearman's rank correlation patterns with

salinity (oligotype A:  $r(34)=-0.69$ ; oligotype C:  $r(34)=-0.60$ ), nitrite (oligotype A:  $r(36)=-0.68$ ; oligotype C:  $r(36)=-0.69$ ) and nitrate (oligotype A:  $r(35)=-0.69$ ; oligotype C:  $r(35)=-0.62$ ) at a significance level of  $p<0.001$ .

## Discussion

The Bay of Bengal nurtures the South Asian summer monsoon, a tremendous ocean-atmosphere-land feedback that delivers freshwater to more than a third of the human population with impacts on agriculture, health, economics and ecosystems throughout South Asia (Mahadevan *et al.*, 2016). Monsoonal driven freshwater inputs impact the ocean's structure, circulation and biogeochemistry in numerous ways, but are predicted to change with global temperature increases (Ashfaq *et al.*, 2009). Substantial efforts have been dedicated to a better characterization of water mass dynamics and their impact on primary production in the Bay of Bengal (Mahadevan *et al.*, 2016). These dynamical attributes have substantial influence on the biota that reside within the marine ecosystem, including fisheries and of course the primary producers that form the base of the food network.

Many studies have tackled the primary producer communities in the Bay of Bengal using satellite data (which reflects the surface layer), chlorophyll and microscopy-based observations (Jyothibabu *et al.*, 2015, 2018; Roxy *et al.*, 2016; Gregg and Rousseaux, 2019). While important, the methods employed have generally lacked resolution for the smaller size fractions (such as *Prochlorococcus* or eukaryotic picoeukaryotes (Ramaiah *et al.*, 2010)), and even those that have employed amplicon approaches have focused on the uppermore surface layer (Larkin *et al.*, 2016, 2020, 2022). The particular importance of *Prochlorococcus* has been highlighted in a satellite imaging-based study concluding a shift in primary producer community from eukaryotic phytoplankton to cyanobacteria in response to decreasing nitrate concentrations (Gregg and Rousseaux, 2019). This community shift has further been projected to critically impact bulk primary production, export production and the expansion and intensity of the oxygen minimum zone in the Bay of Bengal, and with that one of the key biogeochemical features of those waters (Löscher, 2021). The predicted community shift in primary producers might even be amplified with changing oxygen gradients that appear to influence picoeukaryotic communities in the Bay of Bengal based on results from an 18S rRNA gene amplicon study (Chen *et al.*, 2022). However, a comprehensive view of the phytoplankton residing in this region and potential connections to its unique physicochemical characteristics has been lacking to date.

Given the importance of this ecosystem, and the prediction of substantial changes due to temperature rise and changes to the water cycle, it is essential that better knowledge is developed. Indeed, a recent chlorophyll *a* based study suggest picophytoplankton dominate, as compared to micro- and nanoplankton size fractions, likely due to low concentrations of nitrate and phosphate, and potentially low salinity of the northern region of the bay (Sarma *et al.*, 2016, 2020). Here, we examined the entire phytoplankton community composition using the plastid 16S rRNA marker gene, which is considered to have less variable copy numbers than the 18S rRNA gene (Fuller *et al.*, 2006; Decelle *et al.*, 2015; de Vargas *et al.*, 2015; Needham and Fuhrman, 2016; Gong *et al.*, 2020), and flow cytometry. Combining this approach with targeted sequencing of novel taxa (i.e., full-length nuclear 18S-ITS1-5S-ITS2-partial28S sequence and full-length plastid 16S rRNA gene) allowed us to examine diversity of the phytoplankton within the context of cellular abundance, nutrients and other physical features of the water column to characterize the primary producer community during the southwest summer monsoon season.

A dominant feature during our study period was the strong salinity-driven stratification that has previously been reported to inhibit nutrient supplies from subsurface to surface waters (Sarma *et al.*, 2016). We also observed relatively low near-surface concentrations of nitrate and phosphate, particularly in the more northern region of the transect, where the distinction between the surface layer and the SCM was largest in terms of salinity and density and riverine freshwater input the strongest (Fig. 2, Fig. S1, Fig. S2, Table S1). Phytoplankton community composition, particularly for eukaryotes, was also most distinct between these depth layers in the northern stations (Fig. 4A,B). For the southern stations, there were less sharp differences in salinity, but inorganic nutrient concentrations still differed between the surface and SCM. At other stations along the transect, nutrient concentrations varied but were on average higher in the SCM than in the surface.

*Prochlorococcus* diversity has been studied extensively with attention to the high-light (HL)- and low-light (LL)-adapted ecotypes that can be delineated with the 16S rRNA gene or ITS sequences as well as genomically (Rocap *et al.*, 2003; Johnson *et al.*, 2006; Biller *et al.*, 2015). At least 13 *Prochlorococcus* lineages (HLI-VI and LLI-VII) have been identified and their distributional patterns connected to oceanic regions or environmental conditions, such as light availability and temperature (Biller *et al.*, 2015) as well as their differential contributions to primary production (Thompson and Kouba, 2019). Additionally, while early culturing studies indicated *Prochlorococcus* could not use nitrate, and instead relied on ammonium, these results arose from selection occurring in the initial culturing steps, where ammonium was used as a sole nitrogen source in order to select for *Prochlorococcus* over eukaryotic taxa and *Synechococcus*.

Nitrate assimilation genes were later identified in metagenomic assemblies from uncultivated *Prochlorococcus* in subtropical regions (Martiny, Kathuria, *et al.*, 2009), and has been suggested for some members of the HLII ecotype (Berube *et al.*, 2015), which harbors high diversity (Kashtan *et al.*, 2014). Thus a complex set of factors underpin the niche differentiation of *Prochlorococcus* ecotypes (Johnson *et al.*, 2006), many of which remain to be fully characterized.

Surface relative amplicon dominance by HLII has been shown during the spring intermonsoon season in the Bay of Bengal (Pujari *et al.*, 2019; Larkin *et al.*, 2020; Wei *et al.*, 2020), as seen here during the peak southwest summer monsoon season, primarily in the surface but also marginally at the SCM in some stations. Our surface observations have similarities with subtropical gyres, such as the Sargasso Sea (BATS), which exhibits strong seasonal thermal stratification in the photic zone (Steinberg *et al.*, 2001), during which HLII dominates in the surface, followed by HLI, which typically has higher relative abundances during the cooler seasons (Zinser *et al.*, 2006, 2007; Malmstrom *et al.*, 2010; Kashtan *et al.*, 2014) and in higher latitudes (Berube *et al.*, 2015; Larkin *et al.*, 2020; Thompson *et al.*, 2021). At Station ALOHA, which typically has warmer surface waters than BATS, and undergoes less dynamic winter mixing, (Malmstrom *et al.*, 2010), HLII appears to always dominate (Thompson *et al.*, 2018). Single-nucleotide level differences can reflect population biology that connects to ecology (Cordero and Polz, 2014), and we noted different HLII oligotype ratios between the surface and SCM (Fig. 3A) likely reflect niche adaptations that have yet to be elucidated.

While similarities were seen at the surface, we found that at depth *Prochlorococcus* ecotypes differed substantially from reports on the typically much deeper DCM at BATS and ALOHA. Our results capture its diversity for the first time below the upper 7 meters. At the SCM, we observed high relative abundances of two *Prochlorococcus* ecotypes, HLVI (the SCM dominant) and LLVII, which have typically been considered ‘rare’ (Fig. 3A). LLII was also present at high relative abundances, an ecotype that has been reported at BATS or ALOHA alongside LLIV, although typically deeper than LLI, which at BATS can exhibit cell numbers nearly as abundant as the HL ecotypes (Malmstrom *et al.*, 2010). HLVI has not been reported yet in these ecosystems. Unfortunately, little else is known about their ecology and distributions. HLVI was discovered in the middle-to-lower photic zone (75-150 m) of the South China Sea and seems to be adapted to lower lower light levels and potentially nitrate assimilation based on metagenomic analyses. The 16S rRNA gene phylogeny constructed herein indicates that it is in a basal position among high light ecotypes (Fig. 3A), corresponding with the possibility that it might be functionally closer to low light-adapted clades (Huang *et al.*, 2012; Berube *et al.*, 2019). However, HLVI has also been

found to contribute low relative abundances to the surface *Prochlorococcus* community in the North Pacific, in addition to the more prominent HLI and HLII ecotypes (Larkin *et al.*, 2016). Moreover, a transcriptionally active HL-like *Prochlorococcus* population has been reported in the DCM (125 m) at station ALOHA (Shi *et al.*, 2011) and may represent HLVI. In the context of the limited knowledge of this ecotype, our findings raise questions about the aspects of genomic and metabolic potential in HLVI that lead to a rivaling of LLII in the Bay of Bengal (Fig. 3A). Overall, it appears that the ecotype distributions at depth in the Bay of Bengal differ, potentially arising from the different mechanisms of stratification, where deeper photic zone waters are warmer than at the two above time series stations (Bingham and Lukas, 1996; Steinberg *et al.*, 2001) (Fig. 1D).

Higher contributions from another largely unknown *Prochlorococcus* ecotype, LLVII, were also observed at the SCM in the Bay of Bengal where it formed up to 31% of *Prochlorococcus* amplicons (Fig. 3A). *Prochlorococcus* LLVII (formerly clade NC1, Non-Cultured 1) was noted based on ITS sequences in the lower euphotic zone of subtropical waters (Martiny, Tai, *et al.*, 2009). LLVII appears to be highly diverged from other ecotypes and formed a monophyletic group based on 16S rRNA analysis comprising seven LLVII reference sequences, although with weak statistical support (Fig. 3A). Three reference sequences were generated herein; two came from 125 m depth at Hawaii Ocean Time-series, and two came from an Arabian Sea OMZ study (see Experimental procedures). In our samples, one ASV assigned to LLVII clearly dominated across all stations and depths sampled across the transect ( $69 \pm 42\%$  of LLVII). From an environmental perspective, statistical analyses of the Bay of Bengal data indicate that LLVII relative amplicon abundances correlated positively with salinity and dissolved nitrogen concentrations but negatively with temperature and oxygen. Metagenomic analysis of HLVI has shown the capacity to utilize nitrate (Berube *et al.*, 2019) and may drive vertical distributions observed here, although clearly other factors or co-associated environmental factors could be responsible.

While *Prochlorococcus* and *Synechococcus* coexist in the open ocean, *Prochlorococcus* is generally more abundant than *Synechococcus*, often by 10-fold or more (Flombaum *et al.*, 2013). We observed similar average abundances for *Prochlorococcus* and *Synechococcus* at the surface (Fig. 2C). With larger average size (based on FALS), and depending on growth rates, *Synechococcus* may therefore rival *Prochlorococcus* in terms of carbon fixation (see also, Jardillier *et al.*, 2010), although likely only in the upper most layer since it drops off dramatically in abundance with depth. Although reported from open ocean to coastal waters, and tropical to polar regions, and encompassing a wide range of physiological diversity (Sohm *et al.*, 2016), the factors dictating *Synechococcus* clade distributions are not well understood, perhaps in part

because >20 clades have been identified, with some lacking cultured representatives. As seen in the Pacific subtropical gyre (Kent *et al.*, 2019; Xia *et al.*, 2019), we found *Synechococcus* Clade II to be the dominant clade in the Bay of Bengal. Two distinct Clade II oligotypes were clearly partitioned in the surface waters and the SCM (Fig. 3B). Notable contributions were also seen for Clade X at the SCM (Fig. 3B), an ecotype that has been reported to share a similar environmental niche with Clade II, inhabiting warm, oligotrophic tropical and subtropical open oceans (Ahlgren and Rocap, 2012; Sohm *et al.*, 2016), and statistical analysis indicates that Clade X is more widely present at greater photic zone depth with increasing inorganic nitrogen and salinity ( $p < 0.001$ , Spearman's rank correlation). Prior studies have shown that temperature is a major factor in *Synechococcus* clade distributions (Scanlan D. J. *et al.*, 2009), although other factors contribute as well (Pittera *et al.*, 2014; Sohm *et al.*, 2016; Carlson *et al.*, 2022), and our results indicate more complex population biology and environmental variables drive *Synechococcus* fine scale diversity.

Although a greater number of studies have investigated eukaryotic phytoplankton in the Bay of Bengal than cyanobacteria, little is known about the molecular diversity of eukaryotic phytoplankton. In fact, none of the dominant taxa identified herein have been reported previously in the Bay of Bengal, be they prasinophyte, diatom, pelagophyte or dictyochophyte. For example, the diatom amplicon we detected in highest relative abundances (Fig. S5) had >99% nt identity to *Tryblionella apiculata* (unofficial synonym *Nitzschia apiculata*). Prior microscopy-based studies of diatoms in the bay have reported an abundance of *Thalassiontrix*, *Chaetoceros*, *Thalassionema*, *Navicula*, *Thalassiosira*, *Coscinodiscus*, *Skeletonema* and *Nitzschia angularis* (Paul *et al.*, 2007, 2008; Ramaiah *et al.*, 2010). *T. apiculata* has been reported in coastal rivers (West coast of North America) and in an isolated oasis south of the Sahara desert (Gastineau *et al.*, 2021). To our knowledge, this diatom has not been commonly reported in the Bay of Bengal, or open marine water column, however, additional blast searches of 16S rRNA gene sequences from the cultured representative had 98% identity to multiple marine environmental clones, indicating that it may well be important in undersampled warm, low salinity waters such as northern parts of our transect, although physical transport from even more northerly regions of the bay may also have influenced our results.

Among stramenopiles other than diatoms, both dictyochophytes and pelagophytes have recently been reported in the Bay of Bengal based on the RuBisCO gene (*rbcL*), and pelagophytes by V4-V5 18S rRNA amplicons but with no further insights into which genera, much less species (Angelova *et al.*, 2019; Pujari *et al.*, 2019). The dictyochophyte with highest relative amplicon abundances that we observed was *Florenciella*, a mixotrophic flagellate, that feeds on



cyanobacteria and heterotrophic bacteria in culture (Li *et al.*, 2021) and one of the few phagotrophic phytoplankton we observed. Likewise, *P. calceolata* and environmental Clade VIII (PEC-VIII) have not been reported in the Bay of Bengal. *P. calceolata* has been increasingly recognized as a widely distributed oceanic photosynthetic picoeukaryote (Worden *et al.*, 2012; Duerschlag *et al.*, 2022; Guérin *et al.*, 2022) that is particularly important at the SCM and DCM in the Sargasso sea, mesotrophic and oligotrophic stations in the Eastern North Pacific (Dupont *et al.*, 2015; Choi *et al.*, 2020), and the South Pacific Subtropical Ocean (Duerschlag *et al.*, 2022). However, unlike these prior studies *P. calceolata* was not among the dominant eukaryotic phytoplankton at the SCM. Rather, the prasinophyte *Ostreococcus* Clade OII was the clear dominant (representing  $75 \pm 8\%$  of plastid amplicons) at the SCM and at some stations *Micromonas commoda* was relatively abundant in the surface (Fig. 4A). Chlorophytes (often used to refer to all green algae or prasinophytes) and even Mamiellales (Angelova *et al.*, 2019) have been reported in the Bay of Bengal, however most information on which of the diverse taxa within these groups is lacking. In the latter study, because *Ostreococcus* had the most abundant amplicons among eukaryotic plastid sequences, PhytoRef (Decelle *et al.*, 2015) was used for classification and assigned the amplicons to *O. tauri*. Upon checking all the sequences attributed to *O. tauri* in Angelova *et al.* (2019), we found they were different from *O. bengalensis* based on the V4 16S rRNA amplicons by 10 to 53 nt and clustered instead with V4-V5 16S rRNA spanning environmental sequences of Fuller *et al.* (2006), *Ostreococcus* sp. RCC393 and *Ostreococcus* sp. RCC809 (Fig. S8), thereby likely reflecting unrecognized diversity within *Ostreococcus* Clade B/OII. It should be noted that PhytoRef has an *O. bengalensis* reference sequence (FJ744977), but in our phylogenetic analysis of partial-length sequences (Fig. S8), it was placed in a statistically supported clade of *O. bengalensis* sequences, and thus has been misnamed in that database as *O. tauri*. Our analysis further showed 1 to 4 nt differences for the previously published 16S rRNA V4 region from those within the fully supported Clade E, which also includes V1-V2 16S rRNA amplicons representing *O. bengalensis* generated herein (Fig. S8). To our knowledge, these results, and those of Angelova *et al.* (2019) with the reanalysis performed herein, are the first to show such a massive dominance of a single eukaryotic species, i.e. *Ostreococcus* Clade OII, in the SCM/DCM, at least as reflected by amplicon relative abundance data, in combination with a low FALS signal for the cells enumerated by flow cytometry which was also indicative of *Ostreococcus*-sized organisms.

*Ostreococcus* has previously been shown to form four phylogenetically distinct clades based on ITS (Rodríguez *et al.*, 2005) and full-length 18S rRNA gene (Worden, 2006). The two major

marine clades, termed Clade OI (Clade A), which is represented by *O. lucimarinus*, and Clade OII (Clade B) rarely co-occur. QPCR studies demonstrated that the former is present in cooler mesotrophic and coastal waters versus the latter being a more oceanic ecotype present throughout the water column during periods of seasonal mixing at both BATS and ALOHA before being confined to the DCM during stratified periods (Demir-Hilton *et al.*, 2011; Limardo *et al.*, 2017).

While we found OII and not *O. lucimarinus* in the Bay of Bengal samples, consistent with prior work showing these two species rarely co-occur, multiple lines of evidence supported the presence of a fifth *Ostreococcus* Clade (Fig. 5, Fig. S7-8). After acquiring the full length 16S rRNA gene sequence for the LCA taxon seen in Bay of Bengal amplicon data, phylogenetic analyses showed that it formed a statistically supported independent *Ostreococcus* clade. The other lines of evidence for determining the new Bay of Bengal clade was its own species came from the ITS2 secondary structure. *Bathycoccus* and *Ostreococcus* have unique characteristics relative to other prasinophytes and phytoplankton, specifically an additional helix located between universal helices 3 and 4 (Fig. 5B) (Marin and Melkonian, 2010). This additional helix was present in the Clade E sequences and had a unique nucleotide composition compared to homologous helices from other *Ostreococcus* clades, validating designation as a new *Ostreococcus* species (Fig. 5C). Phylogenetic reconstruction based on the full-length 18S rRNA gene also placed the new *Ostreococcus* in a statistically supported clade distinct from known *Ostreococcus* species (Fig. 5A), as did phylogenetic analysis of the more comprehensive set of positions spanning the 18S-ITS1-5S-ITS2-to partial 28S, for which fewer sequences are available in public databases. Collectively, these analyses validated the informal designation of the novel group as being a fifth species, *O. bengalensis*.

The geographic distribution of *O. bengalensis* amplicons indicate that it is predominantly in mesotrophic, potentially lower salinity, continental shelf regions of the Atlantic Ocean and in the Bay of Bengal (Fig. 6A). In addition, our analysis of available V4 18S rRNA amplicons indicated that the clade we describe here corresponds to a unique set of amplicons in coastal temperate stations on both sides of the Atlantic Ocean and the Mediterranean Sea in OSD data which was proposed to potentially represent a different *Ostreococcus* based on the V4 data and termed Clade E (Tragin and Vaultot, 2019). It has also been reported in South China Sea amplicon data along the western coast of the Philippines (dela Peña *et al.*, 2021). The distribution of *Ostreococcus bengalensis* in the Bay of Bengal samples contrasted with the broader dominance of *Ostreococcus* Clade OII, which can have high abundances in the Sargasso Sea, North Pacific and western North

Atlantic (Treusch *et al.*, 2012; Limardo *et al.*, 2017; Bolaños *et al.*, 2020), and in published V4 18S amplicon data from open ocean sites in the Pacific and North Atlantic (Fig. 6A,B).

Taking the physical differences across our transect into account, eukaryotic phytoplankton appeared to be much more strongly differentiated at the southernmost stations than the cyanobacteria. In this sense it appears that the low density surface layer in the northern and central stations had a stronger influence on coastal eukaryotic species, adapted to brackish conditions and fluctuating environments, than is apparent for the cyanobacteria. In the southern stations, weaker stratification likely underpinned the similarities between surface and SCM eukaryotic community composition and the higher overall salinity levels selected for more typical marine taxa. For example, southern stations had considerable differences in dominant surface groups, most noticeable the higher relative abundances of *Ostreococcus* Clade OII and *O. bengalensis* (Fig. 4A). The ratios of *O.* Clade OII oligotypes matched relatively well between the surface and SCM at these stations, in a much more pronounced manner than for *Synechococcus* Clade II or *Prochlorococcus* HLII oligotypes, as well as differing from the ratios observed at other stations across the transect (Fig. 4C).

Unexpectedly, Si:N ratios were lower than the Redfield-Brzezinski ratio (Brzezinski, 1985) in the majority of sampling stations (Table S1), a traditional indicator for silicon limitation for diatoms. In our study, however, dissolved silicon concentrations and Si:N ratios were the lowest in two of the stations with highest relative abundances of diatoms close to land, perhaps reflecting enhanced drawdown (S17 and S18, Table S1). While silicon concentrations and Si:N ratios fluctuated along the transect, they were on average higher in the surface, coinciding with the higher relative abundance of diatom amplicons at the surface, as opposed to their low contributions at the SCM (maximum 6% of plastid amplicons; Fig. 4A, Table S1). Northwards of S6, diatoms generally formed high proportions of the total eukaryotic phytoplankton amplicons (Fig. 4A), and in particular the *T. apiculata*-like was the key member of the diatom community. In contrast, within pelagophytes sharp transitions were seen at the surface, with PcC contributing the most at stations S1-S6 whereas PEC-VIII dominated S7-S18 (Fig. S6). Again, at the surface *Ostreococcus* Clade OII was notable in S1-S2, where *O. bengalensis* was also detected, but at S3-S6 smaller proportions of OII were present at the surface than other eukaryotic phytoplankton and *O. bengalensis* was not observed. Across the entire transect it completely dominated plastid-derived amplicons in the SCM. The complexity of eukaryotic phytoplankton dynamics during these peak summer monsoon samples was not reconciled by statistical analyses based on the parameters we measured and clearly requires additional study.

Collectively, our present study combining high-resolution microbiome profiling with flow cytometry analysis revealed new insights into the community structure and distribution patterns of plankton across a salinity gradient in the Bay of Bengal. Driven by the unique meteorological, oceanographical and physicochemical properties of its waters, we observed a clear distinction of phytoplankton in the bay compared to well-studied subtropical gyre systems within the same temperature ranges, including high relative abundances of two *Prochlorococcus* ecotypes, HLVI and LLVII, generally considered rare. Additionally, next to the abundant *Synechococcus* Clade II, we observed notable abundances of Clade X at similar abundance ratios supporting that both clades share similar environmental niches. Among plastid-derived ASVs, we found numerous sequences assigned to a diatom clade that includes the pennate diatom *Tryblionella apiculata* reported from coastal rivers of e.g. North America but not previously in the Bay of Bengal. Additionally, we observed relative abundance dominance within dictyochophytes by *Florenciella* and within pelagophytes by clades PcC (*P. calceolata*) and PEC-VIII (uncultured pelagophyte). All of these had distinct spatial distributions, vertically and along the salinity gradient of the transect. In the south, *Ostreococcus* Clade OII and a new species, *O. bengalensis*, that we delineated based on classical metrics for species definition, dominated relative abundances of eukaryotic phytoplankton. Generally, we observed different patterns of cyanobacterial and eukaryotic phytoplankton communities between surface and SCM, with a broader array of eukaryotic phytoplankton at the surface of most stations, except the southernmost stations. Our results are critically important in the context of the projected increase of picophytoplankton and the concurrent decrease of larger eukaryotic phytoplankton in the Bay of Bengal and consequences for primary production. Knowledge of this formerly hidden diversity will help to inform projections on primary production and carbon dioxide uptake from the atmosphere.

Together, our results indicate that there are more patterns to be unveiled that incorporate ‘hidden’ diversity of the smallest phototrophic marine organisms that has been missed by survey sampling of the very surface of the photic zone, especially in biologically underexplored oceans like the Bay of Bengal. The insights herein on the composition of Bay of Bengal phytoplanktonic microbial communities, along with their variations to environmental conditions along a longitudinal gradient provide a springboard for more targeted studies and experimentation to understand niche partitioning in this ecosystem, monsoonal-influences, and potential future trajectories. In this region of the world, intense human activity linked to the effects of global warming is already leading to changes in the coastal and marine environment in terms of salinity

and ocean circulation. These biological baselines help for better characterization of the structure and processes in the Bay of Bengal in order to improve the representation of its role in the monsoons and implications on the recent history of the Earth's climate.

## **Experimental procedures**

### *Oceanographic sampling*

Samples were collected from the *RV Roger Revelle* at 18 hydrographic stations using a CTD Niskin Rosette as well as 11 samples collected by bucket after loss of the rosette between 24 August 2015 to 18 September 2015 (Table S1). Vertical profiles of temperature, salinity, fluorescence and nutrients were measured at 0.5 m intervals. Seawater was collected at four to six specific depths ranging from 2-100 m depth corresponding to surface, mixed layer (ML), subsurface chlorophyll maximum (SCM), below SCM, base of the oxycline and anoxic layer. DNA samples were collected by filtering 500 mL seawater onto 0.2  $\mu\text{m}$  pore size polyethersulfone membrane filters (Supor 200, Pall Gelman). Filters were placed into sterile cryovials, flash-frozen in liquid nitrogen and transferred to  $-80\text{ }^{\circ}\text{C}$  until further use. Flow cytometry (FCM) samples were preserved in 0.25% (final concentration) electron microscopy grade glutaraldehyde, fixed at room temperature in the dark for 20 min, flash-frozen in liquid nitrogen and stored at  $-80\text{ }^{\circ}\text{C}$  until further use. Corresponding nutrient samples were taken as previously described and analyzed for concentrations of dissolved inorganic nutrients (nitrate, nitrite, ammonium, phosphate and silicate) using an autoanalyzer (Skalar, San series; Netherlands) (Kremling *et al.*, 1983). The detection limits for nitrate, nitrite, ammonium, phosphate and silicate were  $\pm 0.02$ ,  $\pm 0.02$ ,  $\pm 0.02$ ,  $\pm 0.01$  and  $\pm 0.02\text{ }\mu\text{M}$ , respectively.

### *Flow cytometry*

Samples were analyzed on a BD Influx flow cytometer equipped with a 488 nm 200 mW laser and running on 0.2  $\mu\text{m}$  filtered  $1\times$  PBS as sheath fluid. Scatter and fluorescence signals were collected using the BD FACS Software (software v1.0.0.650) with Forward Angle Light Scatter (FALS, a proxy for cell size) set as the trigger parameter. Prior to measurement, fluorescent polystyrene beads (0.75  $\mu\text{m}$  yellow-green, Polysciences) were added as internal standards. Each sample was run for 8 min at  $25\text{ }\mu\text{L min}^{-1}$ , as measured with an inline flow sensor Sensirion SLG-1430 (Sensirion AG) following a 2 min pre-run, and the corresponding volume was calculated by weighing the sample before and after the run. Cytograms were analyzed using WinList 7.1 (Verity Software House) with populations defined and enumerated based on FALS and chlorophyll autofluorescence (captured by a 692/40 nm bandpass) for *Prochlorococcus* and photosynthetic

eukaryotes, and additionally phycoerythrin autofluorescence (captured by a 572/27 nm bandpass) for *Synechococcus*.

#### *DNA extraction and amplicon sequencing*

DNA was extracted from 47 samples (surface and SCM samples from 18 stations, although there is no data for the Station 15 SCM, and 11 additional surface samples) using the DNeasy Plant Mini Kit (Qiagen) with a slight modification of the manufacturer's protocol including the addition of a bead beating step (Demir-Hilton *et al.*, 2011). Extracts were quantified with the Qubit dsDNA HS (High Sensitivity) Assay Kit (Invitrogen) and diluted to 5 ng  $\mu\text{l}^{-1}$  with TE pH 8.0. PCR reactions (50  $\mu\text{l}$ ) were performed in 5  $\mu\text{l}$  of 10 $\times$  buffer, 1 U of HiFi-Taq, 1.6 mM  $\text{MgSO}_4$  (Invitrogen), 200 nM of each Illumina adapted primers 27FB (5'-AGRGTTYGATYMTGGCTCAG-3') and 338RPL (5'-GCWGCCWCCCGTAGGWGT-3') and cycled at 94 °C for 2 min, 30 times at 94 °C for 15 sec, 55 °C for 30 sec, 68 °C for 1 min and finally elongation at 68 °C for 7 min. Reactions were cleaned with the MinElute PCR Purification Kit (Qiagen) and verified on a 1.2% agarose gel to ensure any primer-dimers had been removed. Sequencing was performed using the Illumina MiSeq platform (2 $\times$ 300 bp paired-end reads, MiSeq Reagent Kit v3).

Sequences were demultiplexed and assigned to corresponding samples using CASAVA (Illumina). Quality of reads was assessed using FastQC v.0.10.1 (Andrews, 2012) and read ends were trimmed at a Phred quality (Q) threshold of 25 using a 10 bp sliding window in Sickle 1.33 (Joshi and Fass). Paired-end reads were merged using USEARCH v10.0.240 (Edgar and Flyvbjerg, 2015) when reads had a  $\geq 50$  bp overlap with maximum 5% mismatch and those with maximum error rate  $> 0.001$  or shorter than 200 bp were discarded. Primer sequences were trimmed using Cutadapt v.1.13 (Martin, 2011). Between 97,572 to 443,556 ( $237,952 \pm 87,596$ ) V1-V2 16S amplicons were recovered per sample, resulting in between 58,416 to 391,913 ( $164,882 \pm 91,062$ ) amplicons per sample after QC. Sequences were deposited in the SRA under BioProject PRJINA905425.

#### *Generation of full-length sequence and additional phylogenetic analyses*

After an initial analysis of amplicons using PhyloAssigner (Vergin *et al.*, 2013) with published reference alignments (Sudek *et al.*, 2015; Choi *et al.*, 2017, 2020) we noted that at several key nodes there were multiple amplicons assigned to LCA (last common ancestor). These indicated presence of taxa within *Ostreococcus* and *Prochlorococcus* in our data that were highly abundant but for which sequences from closest extant taxa were not present in the existing alignments. Therefore, we targeted these groups to gain full-length sequences that could be used to improve

taxonomic representation in the reference alignments. For *Ostreococcus*, DNA from the surface waters at stations S1 and S2 were used to recover full length 18S, ITS and partial 28S, and 16S rRNA gene sequence, where an unusual *Ostreococcus* 16S rRNA gene amplicon had highest relative abundances. Similarly, SCM samples from stations S3, S8, S9 and S18 were used for the recovery of 16S rRNA genes from *Prochlorococcus* HLVI and LLVII ecotypes where they were highly represented based on the amplicon data. For *Ostreococcus*, primers targeted the nuclear 18S rRNA gene and the entire ITS (internal transcribed spacer) region spanning ITS1, ITS2 and the intervening 5.8S rRNA gene regions, with flanking 28S rRNA gene (partial only), so we could characterize the secondary structure of ITS2 as well as longer sequences. The new forward primer designed herein anneals to the beginning of the 18S rRNA gene of *Ostreococcus* (18S-122f: 5'-GTGCGTAAATCCCGACTTCG-3') and for the reverse primer a published primer that anneals to a conserved region of the 28S rRNA gene (ITS055R: 5'-CTCCTTGGTCCGTGTTTCAAGACGGG-3') (Marin *et al.*, 2003; Marin and Melkonian, 2010) was used. For recovery of sequences from *Prochlorococcus* ecotypes HLVI and LLVII, the 16S rRNA gene and the entire ITS region spanning ITS1, ITS2 and the intervening 5S rRNA gene regions was amplified using universal forward primer 27f for the bacterial 16S rRNA gene and a reverse primer designed to recover the ITS/23S fragment of *Prochlorococcus* and *Synechococcus* (23S-1608r: 5'-CYACCTGTGTCGGTTT-3') (Rocap *et al.*, 2002). PCR reactions used HotStar Taq polymerase (Qiagen) and thermal cycling conditions as follows: 95 °C for 15 min, followed by 30 cycles of 95 °C for 30 sec, 60 °C for 30 sec and 72 °C for 2 min; final extension was 72 °C for 7 min. The resulting products were cloned and a total of 7 *Ostreococcus* and 6 *Prochlorococcus* clones were bidirectionally Sanger sequenced; sequences were checked with manual inspection after soft-trimming for potential vector contamination using the NCBI UniVec database and have been deposited under GenBank accessions OP909992-OP909998 and OP910119-OP910124, respectively.

Using the new *Ostreococcus* sequences and data from GenBank, we performed a phylogenetic reconstruction of the 18S rRNA, 18S rRNA+ITS1+5.8S rRNA+ITS2+partial 28S rRNA as well as 16S rRNA genes. Phylogenetic inferences were made by Maximum Likelihood methods implemented in RAxML under gamma corrected GTR model of evolution with 1000 bootstrap replicates based on 1568, 2203 and 1294 homologous positions for 18S rRNA, 18S rRNA+ITS1+5.8S rRNA+ITS2+partial 28S rRNA and 16S rRNA gene trees, respectively, as well as with PhyML. Additional phylogenetic reconstructions were performed in MrBayes with the parameters of lset nst=6 rates=invgamma ncat=6, and ngenvl=10,000,000 samplefreqval=1000

and tempval=0.200, and the final tree was produced with assistance from FigTree 1.4.0 (<http://tree.bio.ed.ac.uk/software/figtree>) and topology from RAxML (Fig. S7).

Since the ITS region is not available for many eukaryotic phytoplankton, we also performed an analysis of the 18S rRNA gene alone. To this end, we collected nearly complete (>1,500 bp) 18S rDNA sequences from Mamiellophyceae and prasinophyte sister clades using NCBI Nucleotide database searches and publicly available genome sequences. Prediction of 18S rDNA sequences from genome sequences was supported using the Basic Rapid Ribosomal RNA Predictor (Seemann, 2018). The collected sequences were aligned using MAFFT with default parameters and gaps were masked using trimAl v1.4.1. Phylogenetic analyses were made using Maximum Likelihood and Bayesian methods as implemented in IQ-TREE v.2.2.0 (Minh *et al.*, 2020) and MrBayes 3.2.7 (Ronquist *et al.*, 2012). The GTR+I+G model of nucleotide substitutions was used for both analyses as informed by ModelFinder (Kalyaanamoorthy *et al.*, 2017) and implemented in IQ-TREE. Maximum Likelihood bootstrap values were calculated using 1000 replicates. Bayesian analyses were performed using parameters as outlined above using two independent runs and 10,000,000 generations per run. A Bayesian consensus tree was constructed after a relative burn in of 25% trees to get posterior probabilities for node support. Both the Maximum Likelihood and Bayesian consensus trees were imported into R using treeio v1.20.2 (Wang *et al.*, 2020) and the final combined tree was produced with ggtree v3.4.2 (Yu *et al.*, 2017).

#### *Taxonomic assignment of amplicons*

Quality trimmed 16S rRNA V1-V2 region amplicon sequences were first subjected to rarefaction analysis of dada2 (Callahan *et al.*, 2016) processed sequences, which indicated all samples reached saturation with final slopes of all rarefaction curves below 0.0001. These were then parsed using phylogenetic methods (Eddy, 1998; Matsen *et al.*, 2010) implemented in PhyloAssigner (Vergin *et al.*, 2013). The 5,935,751 amplicons generated via USEARCH (see above) that passed quality control were run against a reference alignment and tree constructed from near full-length bacterial 16S rRNA sequences as well as a subset of eukaryotic plastid rRNA gene sequences. The 2,132,077 amplicons assigned to the cyanobacterial/plastid region of the tree were retrieved (1,568,988, cyanobacteria; 563,089, plastid) for further analysis. Cyanobacteria and plastid-derived amplicons were then classified using a global plastid and cyanobacterial reference alignment and tree optimized for their placement (Choi *et al.*, 2017). Cyanobacteria amplicons were next analyzed using a cyanobacteria reference alignment (Sudek *et al.*, 2015) and *Prochlorococcus* was further analyzed with reference alignment developed herein after



recognizing presence of atypical clades. Reference alignments used herein are available on github [link to repository will follow]

The *Prochlorococcus* optimized alignment used 52 near full-length *Prochlorococcus* 16S rRNA gene sequences retrieved from SILVA database (release 132, 13 December 2017, <https://www.arb-silva.de/documentation/release-132>), through BLASTN searches against GenBank nr, and the three HLVI and three LLVII full-length 16S rRNA gene sequences generated herein. Note that several LLVII sequences were also recovered from GenBank nr, specifically, JN166198, JN166205, KJ589744, KJ589747 from HOT157 125 m and Arabian Sea OMZ DCM. These sequences (and outgroup sequences) were aligned using MAFFT (Kato and Standley, 2013) with default parameters and gaps were masked using trimAl v1.4 (Capella-Gutiérrez *et al.*, 2009). Phylogenetic inferences were made by Maximum Likelihood methods implemented in RAxML (Stamatakis, 2014) under gamma corrected GTR model of evolution with 1000 bootstrap replicates based on 1283 homologous positions as well as with PhyML 3.0.1 (Ronquist *et al.*, 2012) with the same substitution model and 100 bootstrap replicates. Additional phylogenetic reconstructions were performed in MrBayes 3.2.6 (Ronquist *et al.*, 2012) with the parameters of lset nst=6 rates=invgamma ncat=6, and ngenval=10,000,000 samplefreqval=1000 and tempval=0.200, and the final tree was produced with assistance from FigTree 1.4.0 (<http://tree.bio.ed.ac.uk/software/figtree>) and topology from RAxML. We also constructed a preliminary tree for parsing diatom sequences, although plastid sequences from known species are limiting. 371 diatom sequences were recovered from GenBank nr using Blastn with known query sequences, as well as four rhodophytes for an outgroup. Sequences were aligned using MAFFT (Kato and Standley, 2013) with default parameters and alignment gaps masked using trimAl v1.4 (Capella-Gutiérrez *et al.*, 2009). The phylogenetic reference tree was created using Maximum Likelihood methods implemented in FastTree 2 (Price *et al.*, 2010) under gamma corrected GTR model of evolution with 1000 bootstrap replicates based on 1253 homologous positions.

To characterize distributions of *Ostreococcus* we turned to data from Ocean Sampling Day (Kopf *et al.*, 2015), Malaspina (Logares *et al.*, 2020) and TARA Oceans (Ibarbalz *et al.*, 2019) deposited in metaPR2 (Vaulot *et al.*, 2022). Read counts for ASVs assigned to the genus *Ostreococcus* through metaPR2 were extracted and their taxonomic assignment to species level was validated through alignment to reference sequences, including the novel *O. bengalensis*, currently misidentified as *O. tauri* in metaPR2, due to the lack of *O. bengalensis* reference sequences (now provided herein). The global abundance of *O. lucimarinus*, *O. OII* and *O.*

*bengalensis* was visualized using R and Natural Earth map data from [rnaturalearth](https://github.com/ropensci/rnaturalearth) (<https://github.com/ropensci/rnaturalearth>).

#### *Oligotyping of specific Prochlorococcus, Synechococcus and Ostreococcus clades*

Microdiversity within *Prochlorococcus* ecotype HLII, *Synechococcus* Clade II and *Ostreococcus* Clade OII was resolved using the oligotyping pipeline as described previously (Eren *et al.*, 2013). Briefly, amplicons identified as *Prochlorococcus* HLII by PhyloAssigner were aligned with template-based aligner PyNAST (Caporaso *et al.*, 2010) against a curated alignment template, GreenGenes (McDonald *et al.*, 2012). Uninformative gaps were removed and the entropy of each nucleotide position was calculated with the oligotyping pipeline. Nucleotide positions used to define oligotypes were selected iteratively with positions being included if they contributed to converged entropy within an oligotype, and excluded if they contained redundant information or did not contribute to convergence. The most information-rich three variable positions out of 275 positions within V1-V2 region were used to define oligotypes, and each oligotype was required to have a minimum substantive abundance of 20 amplicons, such that an oligotype was not included if the most common sequence for that type occurred < 20 times. The same procedures were performed for *Synechococcus* Clade II and *Ostreococcus* Clade OII amplicons, with two and three positions out of 276 and 289 positions for defining oligotypes, respectively.

#### *Statistical analysis*

For comparison between specific sets of data points we used non-parametric testing procedures within R (R Core Team, 2022). We also performed multivariate statistical analyses of relative abundances of all taxa as derived from taxonomic assignment of amplicon sequences using non-metric multidimensional scaling (nMDS) analysis to compare communities between different samples. Prior to nMDS analysis, low abundant taxa that were not present in more than 0.0001% relative abundances were removed and the analysis was based on the Bray-Curtis distance measure with no initial data transformation applied to not obscure the distance measure. In addition to nMDS, a compositional data analysis (CoDA) of taxonomically assigned amplicon counts was performed (Gloor and Reid, 2016) to determine taxa driving the differences in microbial compositions at different stations. A canonical correspondence analysis (CCA) of Hellinger-transformed amplicon count data was used to investigate the difference in individual taxa with respect to environmental parameters. Data analysis and visualization were performed using functions of the vegan R package (Oksanen *et al.*, 2022) and ggplot2 (Wickham, 2016) as implemented in the ampvis R package (Andersen *et al.*, 2018).

## Acknowledgements

We thank the captain and crew of the RV Roger Revelle and the ONR departmental research initiative, ASIRI, for support of the cruise, AM and GSJ (N000141310451). We also thank Huiwen Deng for assistance during manuscript preparation. This research was supported by MBARI, the Gordon and Betty Moore Foundation (GBMF3788), and NSF Dimensions 2230811 to AZW. F.W. was supported by BIOS-SCOPE (Simons Foundation International).

## Conflict of interest

The authors declare no competing interests.

## Author contribution (CRediT)

**Jan Strauss:** investigation (equal lead), data curation (equal lead), formal analysis (equal lead), visualization (equal lead), writing - review & editing (equal).

**Chang Jae Choi:** investigation (equal lead), data curation (equal lead), software (equal lead), formal analysis (equal lead), visualization (equal lead), writing - original draft preparation (equal lead), writing - review & editing (supporting). **Jonathan Grone:** investigation (supporting), visualization (supporting), writing - review & editing (supporting). **Fabian Wittmers:** investigation (supporting), data curation (supporting), software (equal lead), visualization (supporting), writing - review & editing (supporting). **Valeria Jimenez:** investigation (supporting), writing - review & editing (supporting). **Kriste Makareviciute-Fichtner:** investigation (supporting), formal analysis (supporting), writing - review & editing (supporting). **Charles Bachy:** investigation (supporting), formal analysis (supporting), writing - review & editing (supporting). **Gualtiero Spiro Jaeger:** investigation (supporting). **Camille Poirier:** investigation (supporting), resources (supporting). **Charlotte Eckmann:** investigation (supporting), formal analysis (supporting), writing - review & editing (supporting). **Carolin R. Löscher:** writing - review & editing (supporting). **VVSS Sarma:** resources (supporting), writing - review & editing (supporting). **Amala Mahadevan:** funding acquisition (equal), supervision (equal), writing - review & editing (supporting). **Alexandra Z. Worden:** conceptualization (lead), funding acquisition (equal), supervision (equal), investigation (supporting), writing - original draft preparation (equal lead), writing - review & editing (lead).

## Data availability statement

Reference alignments used to taxonomically assign amplicon sequences in this study are available on github at [link to repository will follow]. Amplicon sequences have been deposited at the

Sequence Read Archive under BioProject PRJINA905425. The 16S and 18S rRNA gene sequences generated in this study have been deposited to DDBJ/EMBL/GenBank under accession numbers OP909992-OP909998 and OP910119-OP910124, respectively. All other data that support the findings of this study are available from the corresponding author upon reasonable request.

## References

- Ahlgren, N.A. and Rocap, G. (2012) Diversity and distribution of marine *Synechococcus*: multiple gene phylogenies for consensus classification and development of qPCR assays for sensitive measurement of clades in the ocean. *Front Microbiol* **3**: 213.
- Andersen, K.S., Kirkegaard, R.H., Karst, S.M., and Albertsen, M. (2018) ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv* 299537.
- Andrews, S. (2012) FastQC: a quality control tool for high throughput sequence data, Babraham Institute.
- Angelova, A.G., Ellis, G.A., Wijesekera, H.W., and Vora, G.J. (2019) Microbial composition and variability of natural marine planktonic and biofouling communities from the Bay of Bengal. *Front Microbiol* **10**: 2738.
- Ashfaq, M., Shi, Y., Tung, W.-W., Trapp, R.J., Gao, X., Pal, J.S., and Diffenbaugh, N.S. (2009) Suppression of south Asian summer monsoon precipitation in the 21st century. *Geophys Res Lett* **36**.
- Bachy, C., Wittmers, F., Muschiol, J., Hamilton, M., Henrissat, B., and Worden, A.Z. (2022) The land-sea connection: insights into the plant lineage from a green algal perspective. *Annu Rev Plant Biol* **73**: 585–616.
- Bennke, C.M., Pollehne, F., Müller, A., Hansen, R., Kreikemeyer, B., and Labrenz, M. (2018) The distribution of phytoplankton in the Baltic Sea assessed by a prokaryotic 16S rRNA gene primer system. *J Plankton Res* **40**: 244–254.
- Berube, P.M., Biller, S.J., Kent, A.G., Berta-Thompson, J.W., Roggensack, S.E., Roache-Johnson, K.H., et al. (2015) Physiology and evolution of nitrate acquisition in *Prochlorococcus*. *ISME J* **9**: 1195–1207.
- Berube, P.M., Rasmussen, A., Braakman, R., Stepanauskas, R., and Chisholm, S.W. (2019) Emergence of trait variability through the lens of nitrogen assimilation in *Prochlorococcus*. *Elife* **8**: e41043.
- Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015) *Prochlorococcus*: the structure and function of collective diversity. *Nat Rev Microbiol* **13**: 13–27.
- Bingham, F.M. and Lukas, R. (1996) Seasonal cycles of temperature, salinity and dissolved oxygen observed in the Hawaii Ocean Time-series. *Deep Sea Res Part 2 Top Stud Oceanogr* **43**: 199–213.
- Bolaños, L.M., Karp-Boss, L., Choi, C.J., Worden, A.Z., Graff, J.R., Haëntjens, N., et al. (2020) Small phytoplankton dominate western North Atlantic biomass. *ISME J* **14**: 1663–1674.
- Brzezinski, M.A. (1985) The Si:C:N ratio of marine diatoms: interspecific variability and the effect of some environmental variables. *J Phycol* **21**: 347–357.
- Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nat Methods* **13**: 581–583.

- Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.
- Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L., and Knight, R. (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**: 266–267.
- Carlson, M.C.G., Ribalet, F., Maidanik, I., Durham, B.P., Hulata, Y., Ferrón, S., et al. (2022) Viruses affect picocyanobacterial abundance and biogeography in the North Pacific Ocean. *Nat Microbiol* **7**: 570–580.
- Chen, Z., Gu, T., Wang, X., Wu, X., and Sun, J. (2022) Oxygen gradients shape the unique structure of picoeukaryotic communities in the Bay of Bengal. *Sci Total Environ* **814**: 152862.
- Choi, C.J., Bachy, C., Jaeger, G.S., Poirier, C., Sudek, L., Sarma, V.V.S.S., et al. (2017) Newly discovered deep-branching marine plastid lineages are numerically rare but globally distributed. *Curr Biol* **27**: R15–R16.
- Choi, C.J., Jimenez, V., Needham, D.M., Poirier, C., Bachy, C., Alexander, H., et al. (2020) Seasonal and geographical transitions in eukaryotic phytoplankton community structure in the Atlantic and Pacific Oceans. *Front Microbiol* **11**: 542372.
- Cordero, O.X. and Polz, M.F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nat Rev Microbiol* **12**: 263–273.
- Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chrétiennot-Dinet, M.J., Neveux, J., et al. (1994) Smallest eukaryotic organism. *Nature* **370**: 255–255.
- Decelle, J., Romac, S., Stern, R.F., Bendif, E.M., Zingone, A., Audic, S., et al. (2015) PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Mol Ecol Resour* **15**: 1435–1445.
- Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C.L., Zehr, J.P., and Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *ISME J* **5**: 1095–1107.
- Duerschlag, J., Mohr, W., Ferdelman, T.G., LaRoche, J., Desai, D., Croot, P.L., et al. (2022) Niche partitioning by photosynthetic plankton as a driver of CO<sub>2</sub>-fixation across the oligotrophic South Pacific Subtropical Ocean. *ISME J* **16**: 465–476.
- Dupont, C.L., McCrow, J.P., Valas, R., Moustafa, A., Walworth, N., Goodenough, U., et al. (2015) Genomes and gene expression across light and productivity gradients in eastern subtropical Pacific microbial communities. *ISME J* **9**: 1076–1092.
- Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics* **14**: 755–763.
- Edgar, R.C. and Flyvbjerg, H. (2015) Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics* **31**: 3476–3482.
- Eren, A.M., Maignien, L., Sul, W.J., Murphy, L.G., Grim, S.L., Morrison, H.G., and Sogin, M.L. (2013) Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods Ecol Evol* **4**:
- Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., et al. (2013) Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci U S A* **110**: 9824–9829.

- Fuller, N.J., Campbell, C., Allen, D.J., Pitt, F.D., Zwirgmaier, K., Le Gall, F., et al. (2006) Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the Arabian Sea using a PCR biased towards marine algal plastids. *Aquat Microb Ecol* **43**: 79–93.
- Gastineau, R., Hamed, C., Baba Hamed, M.B., Abi-Ayad, S.-M.E.-A., Bāk, M., Lemieux, C., et al. (2021) Morphological and molecular identification reveals that waters from an isolated oasis in Tamanrasset (extreme South of Algerian Sahara) are colonized by opportunistic and pollution-tolerant diatom species. *Ecol Indic* **121**: 107104.
- Giovannoni, S.J. and Vergin, K.L. (2012) Seasonality in ocean microbial communities. *Science* **335**: 671–676.
- Gloor, G.B. and Reid, G. (2016) Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can J Microbiol* **62**: 692–703.
- Gomes, H.R., Goes, J.I.G., and Saino, T. (2000) Influence of physical processes and freshwater discharge on the seasonality of phytoplankton regime in the Bay of Bengal. *Cont Shelf Res* **20**: 313–330.
- Gong, W., Hall, N., Paerl, H., and Marchetti, A. (2020) Phytoplankton composition in a eutrophic estuary: comparison of multiple taxonomic approaches and influence of environmental factors. *Environ Microbiol* **22**: 4718–4731.
- Green, B.R. (2011) Chloroplast genomes of photosynthetic eukaryotes. *Plant J* **66**: 34–44.
- Gregg, W.W. and Rousseaux, C.S. (2019) Global ocean primary production trends in the modern ocean color satellite record (1998–2015). *Environ Res Lett* **14**: 124011.
- Guérin, N., Ciccarella, M., Flamant, E., Frémont, P., Mangenot, S., Istace, B., et al. (2022) Genomic adaptation of the picoeukaryote *Pelagomonas calceolata* to iron-poor oceans revealed by a chromosome-scale genome sequence. *Commun Biol* **5**: 983.
- Huang, S., Wilhelm, S.W., Harvey, H.R., Taylor, K., Jiao, N., and Chen, F. (2012) Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. *ISME J* **6**: 285–297.
- Ibarbalz, F.M., Henry, N., Brandão, M.C., Martini, S., Busseni, G., Byrne, H., et al. (2019) Global trends in marine plankton diversity across kingdoms of life. *Cell* **179**: 1084–1097.e21.
- Jardillier, L., Zubkov, M.V., Pearman, J., and Scanlan, D.J. (2010) Significant CO<sub>2</sub> fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME J* **4**: 1180–1192.
- Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M.S., and Chisholm, S.W. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.
- Joshi, N.A. and Fass, J.N. Sickle: a windowed adaptive trimming tool for FASTQ files using quality.
- Jyothibabu, R., Arunpandi, N., Jagadeesan, L., Karnan, C., Lallu, K.R., and Vinayachandran, P.N. (2018) Response of phytoplankton to heavy cloud cover and turbidity in the northern Bay of Bengal. *Sci Rep* **8**: 11282.
- Jyothibabu, R., Madhu, N.V., Maheswaran, P.A., and Nair, K.K.C. (2003) Dominance of dinoflagellates in micro-zooplankton community in the oceanic regions of the Bay of Bengal and the Andaman Sea.

- Jyothibabu, R., Vinayachandran, P.N., Madhu, N.V., Robin, R.S., Karnan, C., Jagadeesan, L., and Anjusha, A. (2015) Phytoplankton size structure in the southern Bay of Bengal modified by the Summer Monsoon Current and associated eddies: implications on the vertical biogenic flux. *Journal of Marine Systems* **143**: 98–119.
- Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**: 587–589.
- Karl, D.M. and Church, M.J. (2014) Microbial oceanography and the Hawaii Ocean Time-series programme. *Nat Rev Microbiol* **12**: 699–713.
- Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., et al. (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**: 416–420.
- Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.
- Kent, A.G., Baer, S.E., Mouginit, C., Huang, J.S., Larkin, A.A., Lomas, M.W., and Martiny, A.C. (2019) Parallel phylogeography of *Prochlorococcus* and *Synechococcus*. *ISME J* **13**: 430–441.
- Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., et al. (2015) The ocean sampling day consortium. *Gigascience* **4**: 27.
- Kremling, K., Ehrhardt, M., and Grasshoff, K. (1983) Methods of seawater analysis, 2nd Edition. Grasshoff, K., Ehrhardt, M., and Kremling, K. (eds) Verlag Chemie Weinheim, New York.
- Krishna, M.S., Prasad, M.H.K., Rao, D.B., Viswanadham, R., Sarma, V.V.S.S., and Reddy, N.P.C. (2016) Export of dissolved inorganic nutrients to the northern Indian Ocean from the Indian monsoonal rivers during discharge period. *Geochim Cosmochim Acta* **172**: 430–443.
- Larkin, A.A., Blinebry, S.K., Howes, C., Lin, Y., Loftus, S.E., Schmaus, C.A., et al. (2016) Niche partitioning and biogeography of high light adapted *Prochlorococcus* across taxonomic ranks in the North Pacific. *ISME J* **10**: 1555–1567.
- Larkin, A.A., Garcia, C.A., Ingoglia, K.A., Garcia, N.S., Baer, S.E., Twining, B.S., et al. (2020) Subtle biogeochemical regimes in the Indian Ocean revealed by spatial and diel frequency of *Prochlorococcus* haplotypes. *Limnol Oceanogr* **65**:
- Larkin, A.A., Hagstrom, G.I., Brock, M.L., Garcia, N.S., and Martiny, A.C. (2022) Basin-scale biogeography of *Prochlorococcus* and SAR11 ecotype replication. *ISME J*.
- Limardo, A.J., Sudek, S., Choi, C.J., Poirier, C., Rii, Y.M., Blum, M., et al. (2017) Quantitative biogeography of picoprasinophytes establishes ecotype distributions and significant contributions to marine phytoplankton. *Environ Microbiol* **19**: 3219–3234.
- Li, Q., Edwards, K.F., Schvarcz, C.R., Selph, K.E., and Steward, G.F. (2021) Plasticity in the grazing ecophysiology of *Florenciella* (Dichtyochophyceae), a mixotrophic nanoflagellate that consumes *Prochlorococcus* and other bacteria. *Limnol Oceanogr* **66**: 47–60.
- Li, S., Bronner, G., Lepère, C., Kong, F., and Shi, X. (2017) Temporal and spatial variations in the composition of freshwater photosynthetic picoeukaryotes revealed by MiSeq sequencing from flow cytometry sorted samples. *Environ Microbiol* **19**: 2286–2300.

- Logares, R., Deutschmann, I.M., Junger, P.C., Giner, C.R., Krabberød, A.K., Schmidt, T.S.B., et al. (2020) Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* **8**: 55.
- Löscher, C.R. (2021) Reviews and syntheses: trends in primary production in the Bay of Bengal – is it at a tipping point? *Biogeosciences* **18**: 4953–4963.
- Madhupratap, M., Gauns, M., Ramaiah, N., Prasanna Kumar, S., Muraleedharan, P.M., de Sousa, S.N., et al. (2003) Biogeochemistry of the Bay of Bengal: physical, chemical and primary productivity characteristics of the central and western Bay of Bengal during summer monsoon 2001. *Deep Sea Res Part 2 Top Stud Oceanogr* **50**: 881–896.
- Mahadevan, A. (2016) The impact of submesoscale physics on primary productivity of plankton. *Ann Rev Mar Sci* **8**: 161–184.
- Mahadevan, A., Paluszkiwicz, T., Ravichandran, M., Sengupta, D., and Tandon, A. (2016) Introduction to the special issue on the bay of Bengal: from monsoons to mixing. *Oceanography* **29**: 14–17.
- Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010) Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J* **4**: 1252–1264.
- Marin, B. and Melkonian, M. (2010) Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA operons. *Protist* **161**: 304–336.
- Marin, B., Palm, A., Klingberg, M., and Melkonian, M. (2003) Phylogeny and taxonomic revision of plastid-containing euglenophytes based on SSU rDNA sequence comparisons and synapomorphic signatures in the SSU rRNA secondary structure. *Protist* **154**: 99–145.
- Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**: 10–12.
- Martiny, A.C., Kathuria, S., and Berube, P.M. (2009) Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci U S A* **106**: 10787–10792.
- Martiny, A.C., Tai, A.P.K., Veneziano, D., Primeau, F., and Chisholm, S.W. (2009) Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environ Microbiol* **11**: 823–832.
- Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**: 538.
- McCreary, J.P., Yu, Z., Hood, R.R., Vinayachandran, P.N., Furue, R., Ishida, A., and Richards, K.J. (2013) Dynamics of the Indian-Ocean oxygen minimum zones. *Prog Oceanogr* **112-113**: 15–37.
- McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**: 610–618.
- Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* **37**: 1530–1534.



- Monier, A., Worden, A.Z., and Richards, T.A. (2016) Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ Microbiol Rep* **8**: 461–469.
- Naik, R.K., Hegde, S., and Anil, A.C. (2011) Dinoflagellate community structure from the stratified environment of the Bay of Bengal, with special emphasis on harmful algal bloom species. *Environ Monit Assess* **182**: 15–30.
- Needham, D.M. and Fuhrman, J.A. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* **1**: 16005.
- Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., et al. (2022) vegan: community ecology package.
- Paul, J.T., Ramaiah, N., Gauns, M., and Fernandes, V. (2007) Preponderance of a few diatom species among the highly diverse microphytoplankton assemblages in the Bay of Bengal. *Mar Biol* **152**: 63–75.
- Paul, J.T., Ramaiah, N., and Sardesai, S. (2008) Nutrient regimes and their effect on distribution of phytoplankton in the Bay of Bengal. *Mar Environ Res* **66**: 337–344.
- Paulmier, A. and Ruiz-Pino, D. (2009) Oxygen minimum zones (OMZs) in the modern ocean. *Prog Oceanogr* **80**: 113–128.
- dela Peña, L.B.R., Tejada, A.J., Quijano, J.B., Alonzo, K.H., Gernato, E.G., Caril, A., et al. (2021) Diversity of marine eukaryotic picophytoplankton communities with emphasis on Mamiellophyceae in northwestern Philippines. *Philipp J Sci* **150**.
- Pittera, J., Humily, F., Thorel, M., Grulois, D., Garczarek, L., and Six, C. (2014) Connecting thermal physiology and latitudinal niche partitioning in marine *Synechococcus*. *ISME J* **8**: 1221–1236.
- Poretsky, R.S., Sun, S., Mou, X., and Moran, M.A. (2010) Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environ Microbiol* **12**: 616–627.
- Prasanna Kumar, S., Muraleedharan, P.M., Prasad, T.G., Gauns, M., Ramaiah, N., de Souza, S.N., et al. (2002) Why is the Bay of Bengal less productive during summer monsoon compared to the Arabian Sea? *Geophys Res Lett* **29**: 88–1–88–4.
- Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.
- Pujari, L., Wu, C., Kan, J., Li, N., Wang, X., Zhang, G., et al. (2019) Diversity and spatial distribution of chromophytic phytoplankton in the Bay of Bengal revealed by RuBisCO genes (*rbcL*). *Front Microbiol* **10**: 1501.
- Ramaiah, N., Fernandes, V., Paul, J.T., Jyothibabu, R., Gauns, M., and Jayraj, E.A. (2010) Seasonal variability in biological carbon biomass standing stocks and production in the surface layers of the Bay of Bengal.
- R Core Team (2022) R: a language and environment for statistical computing.
- Rixen, T., Cowie, G., Gaye, B., Goes, J., do Rosário Gomes, H., Hood, R.R., et al. (2020) Reviews and syntheses: present, past, and future of the oxygen minimum zone in the northern Indian Ocean. *Biogeosciences* **17**: 6051–6080.

- Rocap, G., Distel, D.L., Waterbury, J.B., and Chisholm, S.W. (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Appl Environ Microbiol* **68**: 1180–1191.
- Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.
- Rodríguez, F., Derelle, E., Guillou, L., Le Gall, F., Vaulot, D., and Moreau, H. (2005) Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). *Environ Microbiol* **7**: 853–859.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., et al. (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol* **61**: 539–542.
- Roxy, M.K., Modi, A., Murtugudde, R., Valsala, V., Panickal, S., Prasanna Kumar, S., et al. (2016) A reduction in marine primary productivity driven by rapid warming over the tropical Indian Ocean. *Geophys Res Lett* **43**: 826–833.
- Sarkar, S., Pham, H., Ramachandran, S., Nash, J., Tandon, A., Buckley, J., et al. (2016) The interplay between submesoscale instabilities and turbulence in the surface layer of the bay of Bengal. *Oceanography* **29**: 146–157.
- Sarker, S., Panassa, E., Shahadat Hossain, M., Chowdhury, S.R., Yadav, A.K., and Sharifuzzaman, S.M. (2020) A bio-physicochemical perspective of the Bay of Bengal. *J Mar Biol Assoc U K* **100**: 517–528.
- Sarma, V.V.S.S., Rajula, G.R., Durgadevi, D.S.L., Kumar, G.S., and Loganathan, J. (2020) Influence of eddies on phytoplankton composition in the Bay of Bengal. *Cont Shelf Res* **208**: 104241.
- Sarma, V.V.S.S., Rao, G.D., Viswanadham, R., Sherin, C.K., Salisbury, J., Omand, M., et al. (2016) Effects of freshwater stratification on nutrients, dissolved oxygen, and phytoplankton in the Bay of Bengal. *Oceanography* **29**: 222–231.
- Sarma, V.V.S.S., Sridevi, B., Maneesha, K., Sridevi, T., Naidu, S.A., Prasad, V.R., et al. (2013) Impact of atmospheric and physical forcings on biogeochemical cycling of dissolved oxygen and nutrients in the coastal Bay of Bengal. *J Oceanogr* **69**: 229–243.
- Scanlan D. J., Ostrowski M., Mazard S., Dufresne A., Garczarek L., Hess W. R., et al. (2009) Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.
- Seemann, T. (2018) BASic Rapid Ribosomal RNA Predictor (barrnap).
- Shi, Y., Tyson, G.W., Eppley, J.M., and DeLong, E.F. (2011) Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J* **5**: 999–1013.
- Shroyer, E.L., Gordon, A.L., Jaeger, G.S., Freilich, M., Waterhouse, A.F., Farrar, J.T., et al. (2020) Upper layer thermohaline structure of the Bay of Bengal during the 2013 northeast monsoon. *Deep Sea Res Part 2 Top Stud Oceanogr* **172**: 104630.
- Sohm, J.A., Ahlgren, N.A., Thomson, Z.J., Williams, C., Moffett, J.W., Saito, M.A., et al. (2016) Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J* **10**: 333–345.

- Spiro Jaeger, G. and Mahadevan, A. (2018) Submesoscale-selective compensation of fronts in a salinity-stratified ocean. *Sci Adv* **4**: e1701504.
- Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.
- Steinberg, D.K., Carlson, C.A., Bates, N.R., Johnson, R.J., Michaels, A.F., and Knap, A.H. (2001) Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Res Part 2 Top Stud Oceanogr* **48**: 1405–1447.
- Subirana, L., Péquin, B., Michely, S., Escande, M.-L., Meilland, J., Derelle, E., et al. (2013) Morphology, genome plasticity, and phylogeny in the genus *Ostreococcus* reveal a cryptic species, *O. mediterraneus* sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **164**: 643–659.
- Subramanian, V. (1993) Sediment load of Indian rivers. *Curr Sci* **64**: 928–930.
- Sudek, S., Everroad, R.C., Gehman, A.-L.M., Smith, J.M., Poirier, C.L., Chavez, F.P., and Worden, A.Z. (2015) Cyanobacterial distributions along a physico-chemical gradient in the Northeastern Pacific Ocean. *Environ Microbiol* **17**: 3692–3707.
- Thompson, A.W., van den Engh, G., Ahlgren, N.A., Kouba, K., Ward, S., Wilson, S.T., and Karl, D.M. (2018) Dynamics of *Prochlorococcus* diversity and photoacclimation during short-term shifts in water column stratification at station ALOHA. *Front Mar Sci* **5**:
- Thompson, A.W. and Kouba, K. (2019) Differential activity of coexisting *Prochlorococcus* ecotypes. *Front Mar Sci* **6**:
- Thompson, A.W., Kouba, K., and Ahlgren, N.A. (2021) Niche partitioning of low-light adapted *Prochlorococcus* subecotypes across oceanographic gradients of the North Pacific Subtropical Front. *Limnol Oceanogr* **66**: 1548–1562.
- Thushara, V., Vinayachandran, P.N.M., Matthews, A.J., Webber, B.G.M., and Queste, B.Y. (2019) Vertical distribution of chlorophyll in dynamically distinct regions of the southern Bay of Bengal. *Biogeosciences* **16**: 1447–1468.
- Tragin, M. and Vaultot, D. (2019) Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Sci Rep* **9**: 5190.
- Treusch, A.H., Demir-Hilton, E., Vergin, K.L., Worden, A.Z., Carlson, C.A., Donatz, M.G., et al. (2012) Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *ISME J* **6**: 481–492.
- de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.
- Vaultot, D., Sim, C.W.H., Ong, D., Teo, B., Biwer, C., Jamy, M., and Lopes Dos Santos, A. (2022) metaPR2 : a database of eukaryotic 18S rRNA metabarcodes with an emphasis on protists. *Mol Ecol Resour* **22**: 3188–3201.
- Vergin, K.L., Beszteri, B., Monier, A., Thrash, J.C., Temperton, B., Treusch, A.H., et al. (2013) High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J* **7**: 1322–1332.

- Wang, L.-G., Lam, T.T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., et al. (2020) Treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol Biol Evol* **37**: 599–603.
- Wei, Y., Huang, D., Zhang, G., Zhao, Y., and Sun, J. (2020) Biogeographic variations of picophytoplankton in three contrasting seas: the Bay of Bengal, South China Sea and Western Pacific Ocean. *Aquat Microb Ecol* **84**: 91–103.
- Wickham, H. (2016) ggplot2: elegant graphics for data analysis.
- Wijesekera, H.W., Shroyer, E., Tandon, A., Ravichandran, M., Sengupta, D., Jinadasa, S.U.P., et al. (2016) ASIRI: an ocean–atmosphere initiative for Bay of Bengal. *Bull Am Meteorol Soc* **97**: 1859–1884.
- Worden, A.Z. (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat Microb Ecol* **43**: 165–175.
- Worden, A.Z., Janouskovec, J., McRose, D., Engman, A., Welsh, R.M., Malfatti, S., et al. (2012) Global distribution of a wild alga revealed by targeted metagenomics. *Curr Biol* **22**: R675–7.
- Xia, X., Cheung, S., Endo, H., Suzuki, K., and Liu, H. (2019) Latitudinal and vertical variation of *Synechococcus* assemblage composition along 170° W transect from the South Pacific to the Arctic Ocean. *Microb Ecol* **77**: 333–342.
- Yu, G., Smith, D.K., Zhu, H., Guan, Y., and Lam, T.T.-Y. (2017) Ggtree : an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* **8**: 28–36.
- Zinser, E.R., Coe, A., Johnson, Z.I., Martiny, A.C., Fuller, N.J., Scanlan, D.J., and Chisholm, S.W. (2006) *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl Environ Microbiol* **72**: 723–732.
- Zinser, E.R., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D., and Chisholm, S.W. (2007) Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol Oceanogr* **52**: 2205–2220.

## Figure Legends

**Fig. 1.** Oceanographic study context. A. Worldwide phytoplankton observations indicating global spatial biases in biological oceanographic sampling intensity as derived from PHYTOBASE database (<https://doi.org/10.1594/PANGAEA.904397>) by normalization of observations per area of 1° grid. Note that this database is of cyanobacteria and eukaryotic phytoplankton, but is largely from microscopy and flow cytometry with less sequence data; hence it has not been updated with relevant data such as that of (Larkin *et al.*, 2020) on surface *Prochlorococcus* ecotype composition in the Indian Ocean, including Bay of Bengal. B. Sampling sites and the salinity profile in the Bay of Bengal. Oceanographic and environmental data up to 200 m depth were measured for the 18 hydrographic stations (S1-S18, Profile) and 11 additional samples (B1-B11, Surface) were collected from surface waters for salinity, temperature and microbial plankton distribution analysis.

**Fig. 2.** Characteristics of the water column and overall phytoplankton cell abundance. Data is shown in relation to depth from the southernmost (S1) to northernmost (S18) stations in the Bay of Bengal. A. Interpolated depth profile for the salinity and B. *in vivo* fluorescence derived from chlorophyll *a* across the cruise track with the depths sampled for DNA indicated (surface, white cross; Subsurface Chlorophyll Maximum, black cross), whereas flow cytometry samples were collected at 4 to 8 depths depending on the water column structure (not indicated). C. Average cell abundance of cyanobacteria *Prochlorococcus* (green), *Synechococcus* (orange) and eukaryotic phytoplankton (blue) from stations with similar characteristics. Data points and error bars represent the mean and standard deviation, respectively. Note that data from bucket samples is provided in Fig. S3.

**Fig. 3.** Diversity and relative abundance of cyanobacteria in the Bay of Bengal. A. *Prochlorococcus* and B. *Synechococcus* clade amplicons at surface waters and SCM. Each reference tree based on the full-length 16S rRNA genes shows clades with a corresponding row in the heat plots and represented by the percentage abundance of amplicons assigned to *Prochlorococcus* and *Synechococcus* in that sample (sample names are indicated at the top of the vertical column). Oligotype distributions from the most abundant clade within *Prochlorococcus* and *Synechococcus* (*Prochlorococcus* ecotype HLII and *Synechococcus* Clade II) are shown at the bottom of the heat plots.

**Fig. 4.** Major eukaryotic phytoplankton groups in the Bay of Bengal. A. and B. show distributions of major eukaryotic phytoplankton lineages (Viridiplantae mostly represented by prasinophytes, stramenopiles, prymnesiophytes) and detailed groups within the prasinophytes and stramenopiles expressed as the percentage of their amplicons relative to all eukaryotic phytoplankton amplicons in surface waters and the SCM, respectively. C. Microdiversity within the most abundant prasinophyte taxa, *Ostreococcus* OII clade, in the Bay of Bengal. Note that dinoflagellates are poorly resolved using plastid-based methods like ours (Green, 2011; Bennke *et al.*, 2018); however, they have been reported in low abundances in the Bay of Bengal (Paul *et al.*, 2007; Naik *et al.*, 2011) and can be abundant at sites in the eastern bay (Jyothibabu *et al.*, 2003) based on microscopy. Additionally, while not firmly established, it is thought that the primers used herein may under-represent prymnesiophytes (haptophytes).

**Fig. 5.** Phylogeny of the prasinophytes with an emphasis on Mamiellophyceae (Class II) and specific attributes of the ‘new’ *Ostreococcus* species. A. Maximum Likelihood (ML) phylogenetic reconstruction of Mamiellophyceae and other prasinophytes inferred from an alignment of nearly complete nuclear 18S rRNA gene sequences larger than 1,450 bp (145 total sequences, 1466 homologous positions) including the new 18S *Ostreococcus* sequences from this study. Node support was evaluated from 1,000 bootstrap (BS) replicates and Bayesian inference (BI) with 10,000,000 generations per run and 25% burn in using the GTR+I+G model of nucleotide substitution. All branches with support values of  $\geq 70\%$  ML bootstrap and  $\geq 0.90$  Bayesian posterior probability (open circles) and  $\geq 90\%$  ML bootstrap and  $\geq 0.99$  Bayesian posterior probability (black filled circles) are highlighted. *Ostreococcus* clades (pink) are highlighted together with other widespread Mamiellophyceae genera, *Micromonas* (blue) and *Bathycoccus* (green). The new *Ostreococcus* 18S rRNA gene sequences are in bold. The basal branches were collapsed and we used the prasinodermophyte sequences for display purposes only. B. The secondary structure for parts of the Internal Transcribed Spacer 2 (ITS2) in newly discovered *Ostreococcus*, as a reference for defining Mamiellophyceae speciation established by (Marin and Melkonian, 2010). C. Comparison of the Bathycoccaceae family-specific additional ITS2 helix, located between universal helices 3 and 4 from other *Ostreococcus* clades.

**Fig. 6.** Global distributions of the two known major oceanic *Ostreococcus* species as well as the newly proposed *Ostreococcus bengalensis*. Relative abundances of V4 18S rRNA amplicons assigned to each are based on data from TARA Oceans, Malaspina, and Ocean Sampling Day that have been incorporated in metaPR2 (Vaulot *et al.*, 2022). All stations included in the analysis are indicated with a cross, regardless of whether any of the three *Ostreococcus* were detected. Note that all relative abundance values for the *Ostreococcus* species in the Bay of Bengal are based on V1-V2 16S rRNA analysis of the plastid portion of amplicons from our study due to the lack of V4 18S rRNA data for this region.

Fig. 1

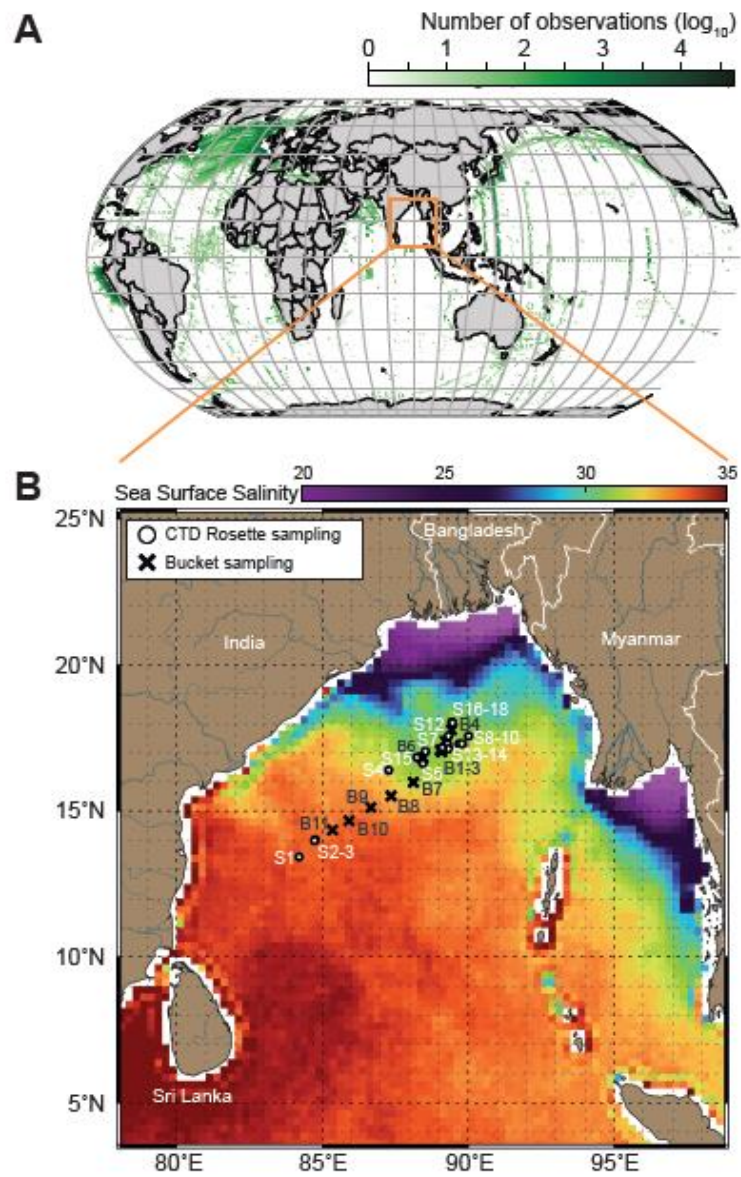


Fig. 2

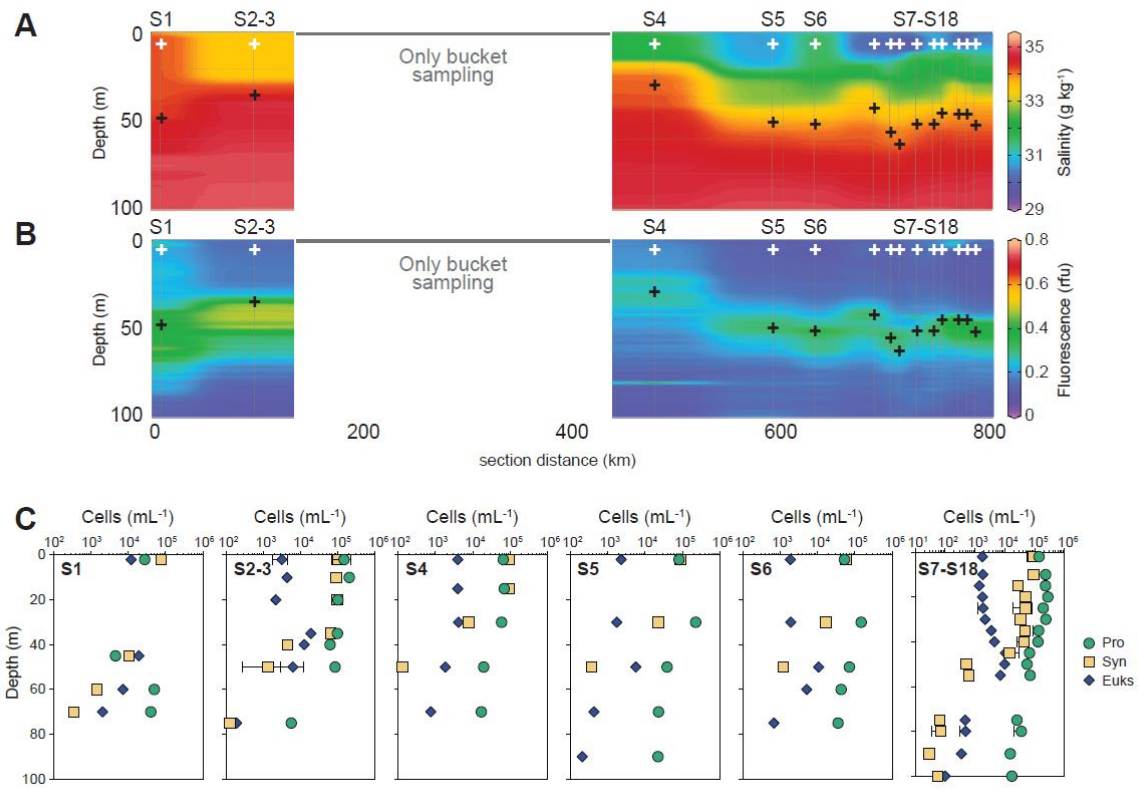




Fig. 3

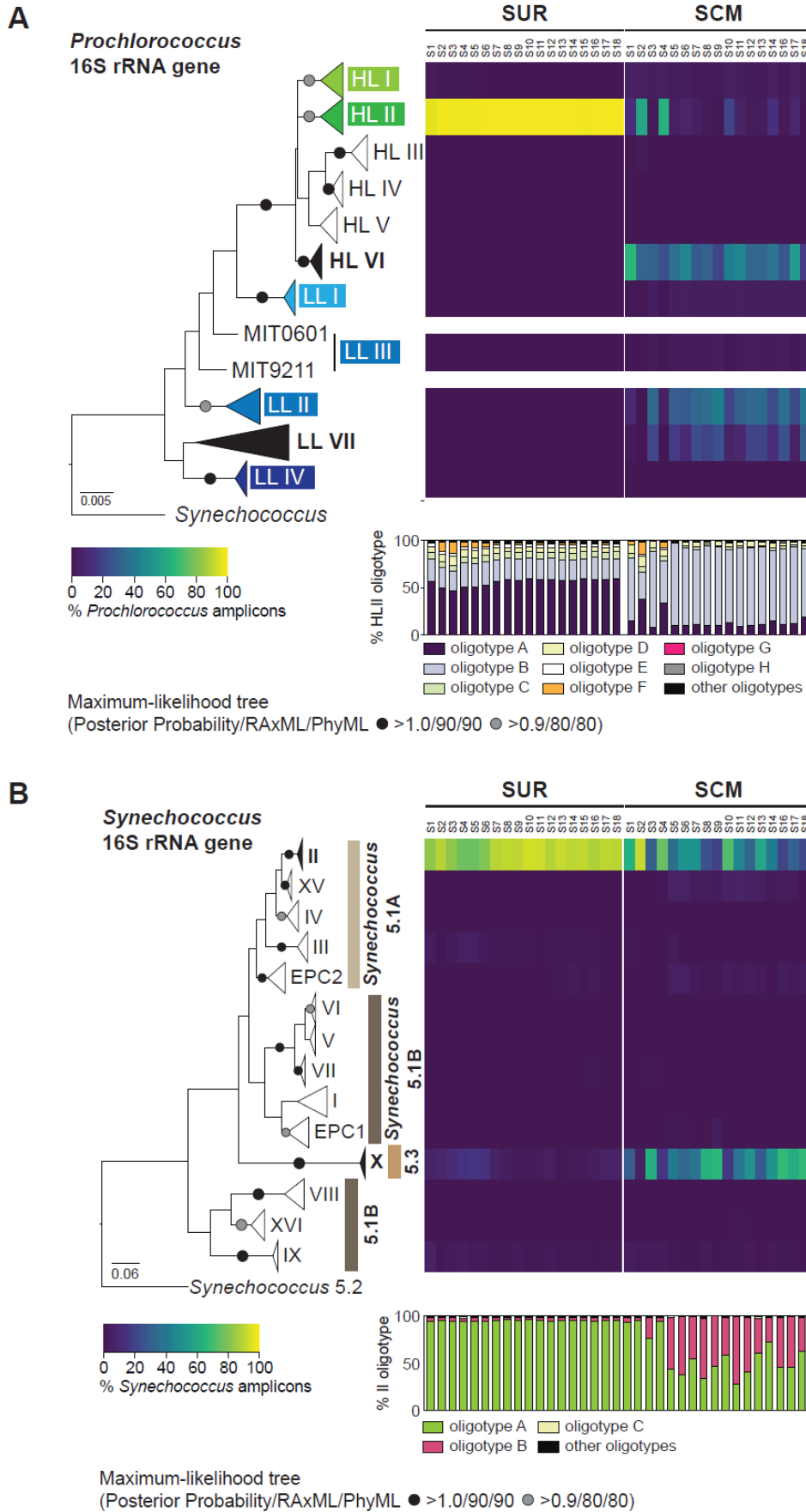


Fig. 4

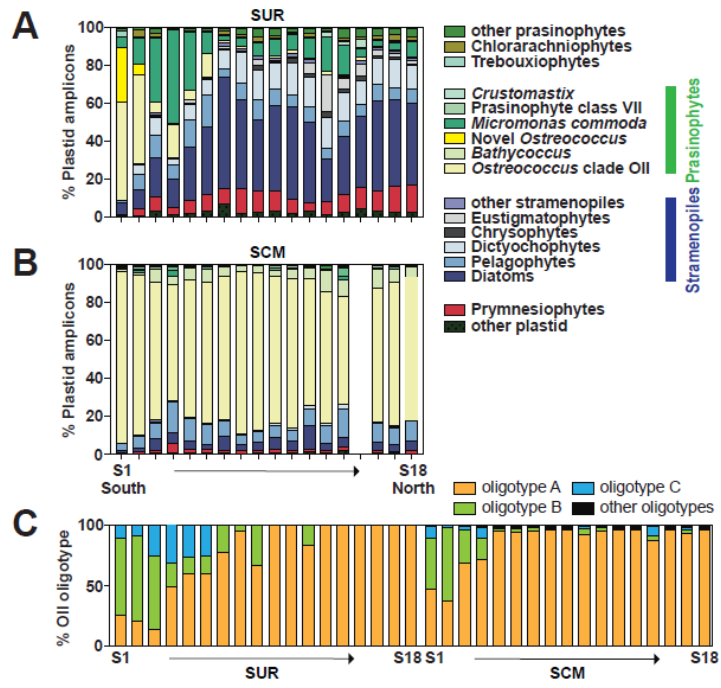
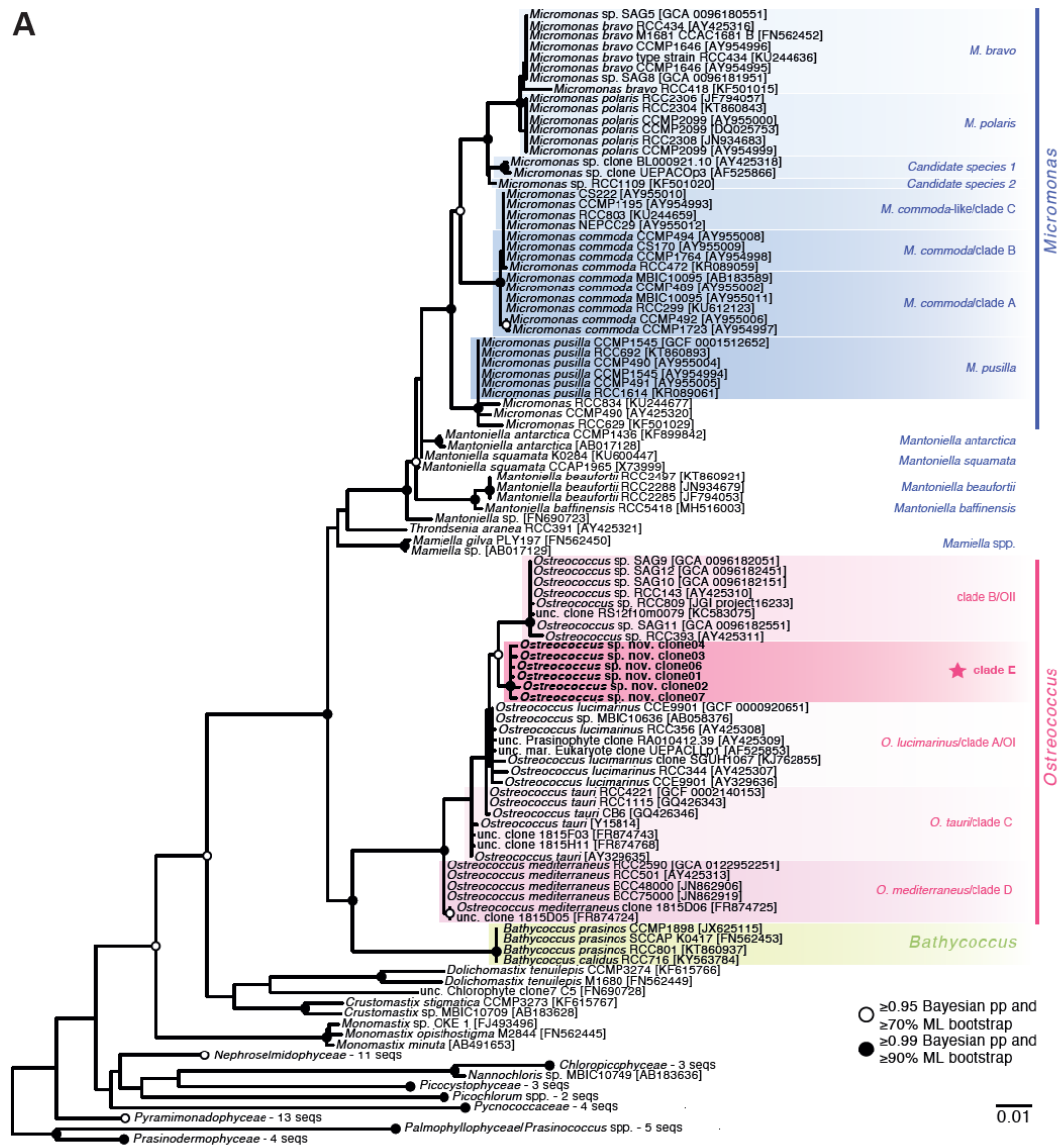
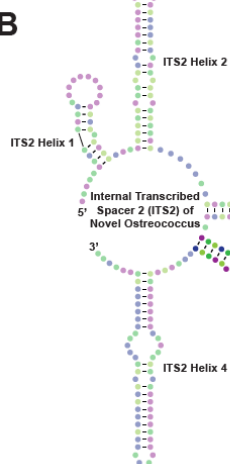


Fig. 5

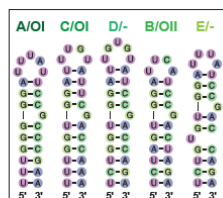
A



B

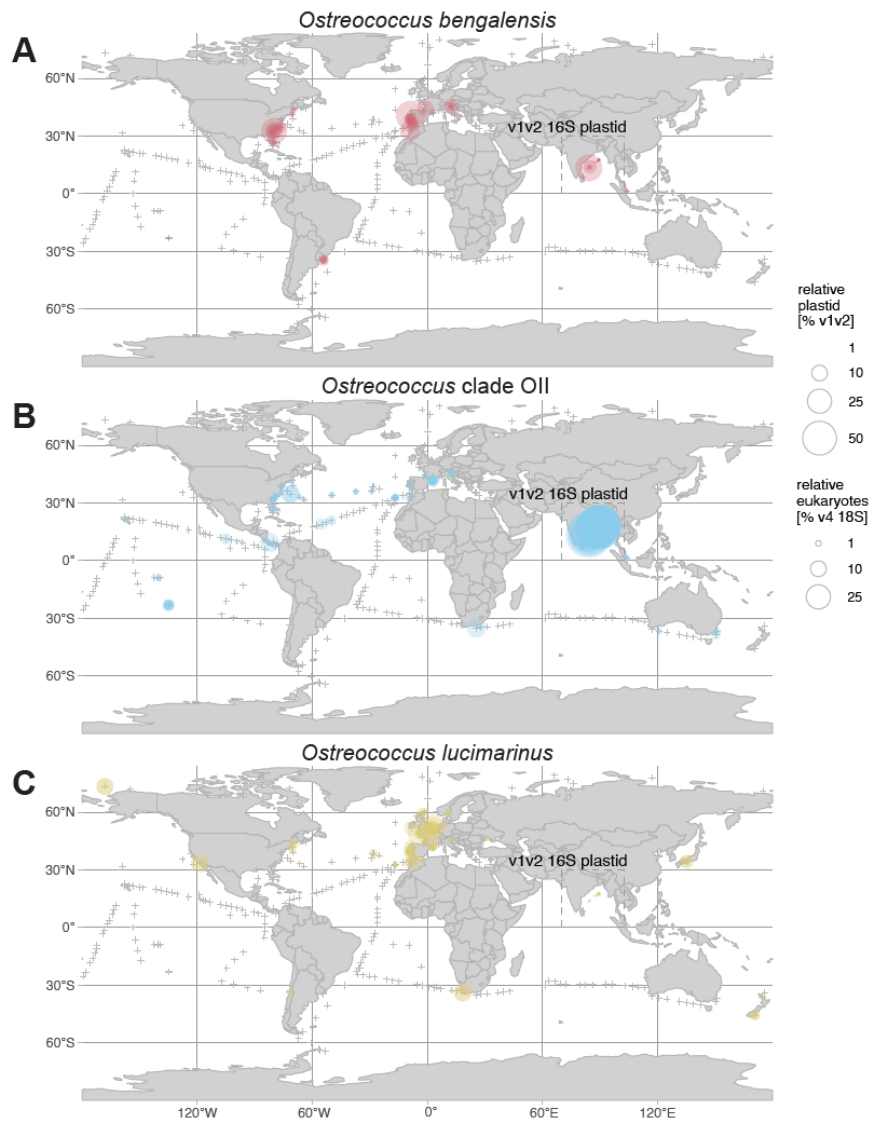


C



Novel *Ostreococcus* clade  
 additional ITS2-helix present only in *Bathycoccaceae*  
 (*Bathycoccus* and *Ostreococcus*)  
 inserted between universal helices 3 & 4

Fig. 6



## Supporting Information

**Table S1.** Metadata and summary tables containing CTD data of 18 hydrographic stations, flow cytometry cell counts and taxonomically assigned amplicons.

**Fig. S1.** A. Sea surface salinity from the Aquarius satellite and curtain plot sections in the water column using an underway CTD system. B. Temperature-Salinity (T-S) diagrams of the 18 CTD stations across the cruise transect showing isopycnal contours and major water masses that play a role in the Bay of Bengal (BBW: Bay of Bengal Water; ASHSW: Arabian Sea High Salinity Water; ICW: Indian Central Water; AAMW: Australasian Mediterranean Water). T-S characteristics of water masses at 100 m (black mark) are indicated for reference only. C. Temperature depth profiles of 18 CTD stations. D. Salinity depth profiles of 18 CTD stations. E. Oxygen depth profiles of 18 CTD stations. Selected sampling sites at the surface and SCM are indicated and labeled (red marks). Mixed layer depth is highlighted as a shaded gray box in C,E.

**Fig. S2.** Interpolated depth profile of nutrient concentrations ( $\mu\text{mol L}^{-1}$ ) from collected samples across stations. A. Nitrate concentrations. B. Nitrite concentrations. C. Ammonium concentrations. D. Phosphate concentrations. E. Silicate concentrations. Black dashed line indicates depth of the SCM.

**Fig. S3.** Cell abundance of *Prochlorococcus*, *Synechococcus* and eukaryotic phytoplankton. A-B. Show conservative salinity and temperature from CTD data for reference. C-E. Flow cytometric analysis of these three groups of phytoplankton. Black lines indicate depth of the SCM. For all panels the mid-section that is primarily white represents stations where the CTD had gone overboard and for which only surface data is available via bucket and intake sampling.

**Fig. S4.** Coarse breakdown of amplicons by domain and photosynthetic grouping. A. Percent of amplicons from non-photosynthetic or photosynthetic taxa in the profile samples at the Surface (left side) and in the SCM (right side). The assignment is based on placement based on the global reference tree provided in (Vergin *et al.*, 2013). B. Percent of amplicons assigned to eukaryotic phytoplankton (plastid 16S rRNA gene) or cyanobacteria out of all amplicons assigned to photosynthetic taxa. C. Percent of amplicons assigned to *Prochlorococcus* or *Synechococcus* out of all cyanobacterial amplicons.

**Fig. S5.** Assignment of diatom V1-V2 16S rRNA gene amplicons against the diatom reference alignment. The heat color depicts the relative abundance of specific diatoms out of total diatom amplicons. The tree shown is based on 371 diatom reference sequences. Note that only surface samples are shown because the SCM samples exhibited less diatom amplicons on average compared to surface waters (surface,  $34.8 \pm 14.6\%$ ; SCM,  $5.2 \pm 2.7\%$  of the total eukaryotic phytoplankton amplicons).

**Fig. S6.** Diversity and distributions of newly resolved pelagophyte clades in the Bay of Bengal. High-taxonomic resolution of pelagophyte distributions in the surface waters (SURFACE) and the SCM. Reference trees to the left of each panel indicate clades. Heatmaps show the percentage of amplicons relative to those of pelagophytes. Bootstrap support and phylogenetic inferences are as detailed in Fig. 2 from Choi *et al.* (2020).

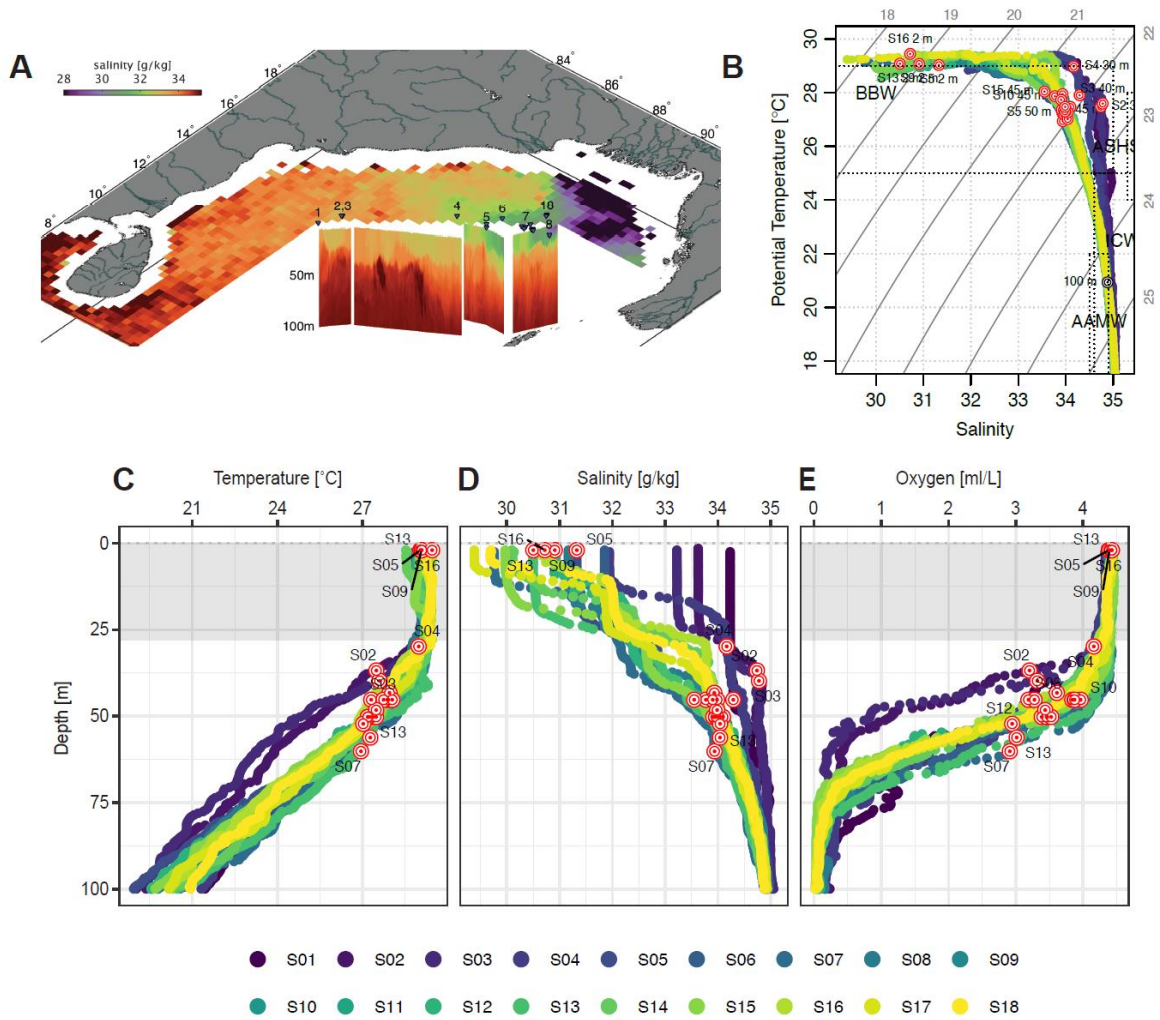
**Fig. S7.** Phylogenetic analysis of the Bathycoccaceae. This family contains *Bathycoccus* and *Ostreococcus* and by analyzing just members from this order, we were able to retain a greater number of homologous nucleotide positions than when performing reconstructions across the diversity of prasinophytes (such as Fig. 5A), thereby strengthening insights into the novel clade. A. ML analysis of near full-length plastid 16S rRNA genes for the phylogenetic placement of clade E. An additional basal clade labeled “F” in this reconstruction is represented by two environmental clones from estuaries, and indicates the presence of additional unrecognized Bathycoccaceae diversity. B. ML analysis of the nuclear 18S-ITS-1-5S-ITS2 and partial 28S rRNA sequence. C.

ML analysis of the nuclear 18S rRNA genes, again with a greater number of homologous positions available than for Fig. 5A. Note that *O. tauri* (Clade C), which was isolated from an oyster lagoon in Southern France (Courties *et al.*, 1994) has been seen in few truly marine settings. Likewise, Clade D, represented by *O. mediterraneus*, seems to be restricted to coastal lagoons, such as where it was first isolated in the Mediterranean (Subirana *et al.*, 2013), and where amplicons have been reported, specifically a lagoon in Uruguay and a freshwater lake in China (Li *et al.*, 2017; Tragin and Vaultot, 2019).

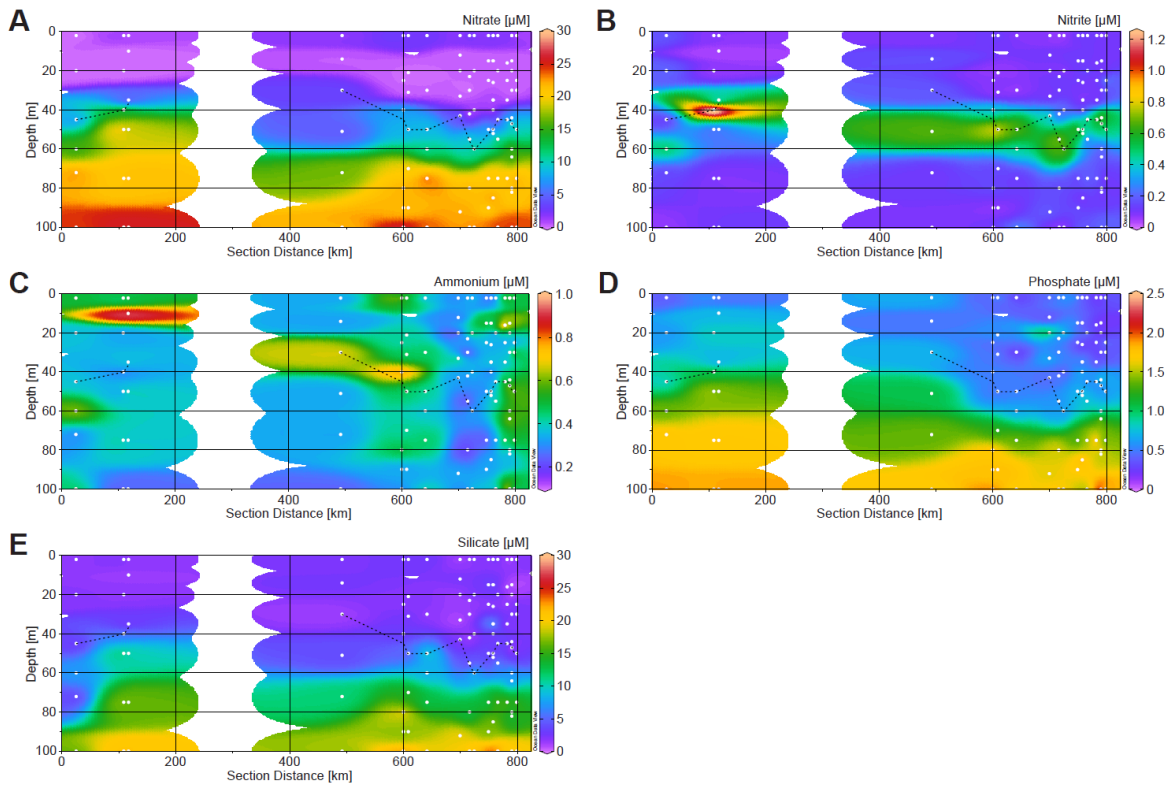
**Fig. S8.** Phylogenetic analysis of plastid 16S rRNA sequences of the Bathycoccaceae family containing *Bathycoccus* and *Ostreococcus* (73 total sequences, 1445 homologous positions) including full-length sequences as in Fig. S7A and additional (short) amplicons from Angelova *et al.* (2019) (V4 amplicons: names beginning with SRR8106743) and environmental clone library sequences, of ~790 nt and spanning the V4-V5 region, from Fuller *et al.* (2006) (AY702185, AY702186, AY702191) as well as a representative V1-V2 sequence read from *O. bengalensis* generated herein (highlighted in bold). Both, amplicons from Angelova *et al.* (2019) and Fuller *et al.* (2006) cluster with *Ostreococcus* Clade B/OII including *Ostreococcus* sp. RCC809 and *Ostreococcus* sp. RCC393, whereas the V1-V2 amplicon of *O. bengalensis* clusters within Clade E. The two environmental clones from estuaries (KX602022, KX602034) forming a basal clade labeled “F” in Fig. S7A are unstable in phylogenetic reconstructions, sometimes clustering outside *Ostreococcus* Clade A/OI. Node support was evaluated from 1,000 bootstrap (BS) replicates and Bayesian inference (BI) with 10,000,000 generations per run and 25% burn in using the GTR+I+G model of nucleotide substitution. Statistical support (ML bootstrap/Bayesian posterior probability) is provided for selected nodes and all branches with support values of  $\geq 70\%$  ML bootstrap and  $\geq 0.90$  Bayesian posterior probability (open circles) and  $\geq 90\%$  ML bootstrap and  $\geq 0.99$  Bayesian posterior probability (black filled circles) are highlighted. The tree is rooted in the *Bathycoccus* Clade BI/BII.

**Fig. S9.** Canonical Correspondence analysis (CCA) of amplicons assigned to phytoplankton taxa at the 18 CTD stations including samples from the surface and surface chlorophyll maximum (SCM) layers and environmental parameters. A. CCA plot of distribution of amplicon sequence variants assigned to cyanobacteria ecotypes along environmental parameters. B. CCA plot of distribution of amplicon sequence variants assigned to major eukaryotic phytoplankton groups (cryptophytes, prymnesiophytes, stramenopiles) along environmental parameters. C. CCA plot of distribution of amplicon sequence variants assigned to green algae and Chlorarachniophyceae along environmental parameters.

Supp. Fig. 1

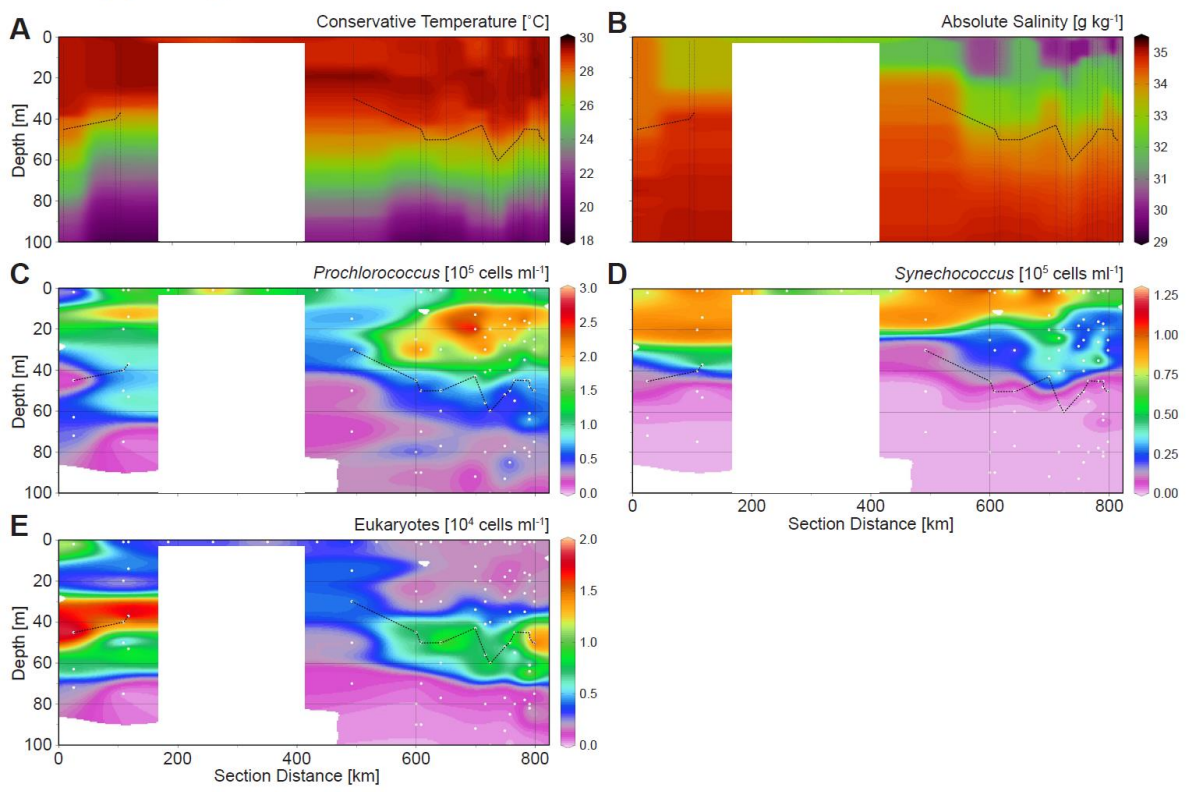


Supp. Fig. 2

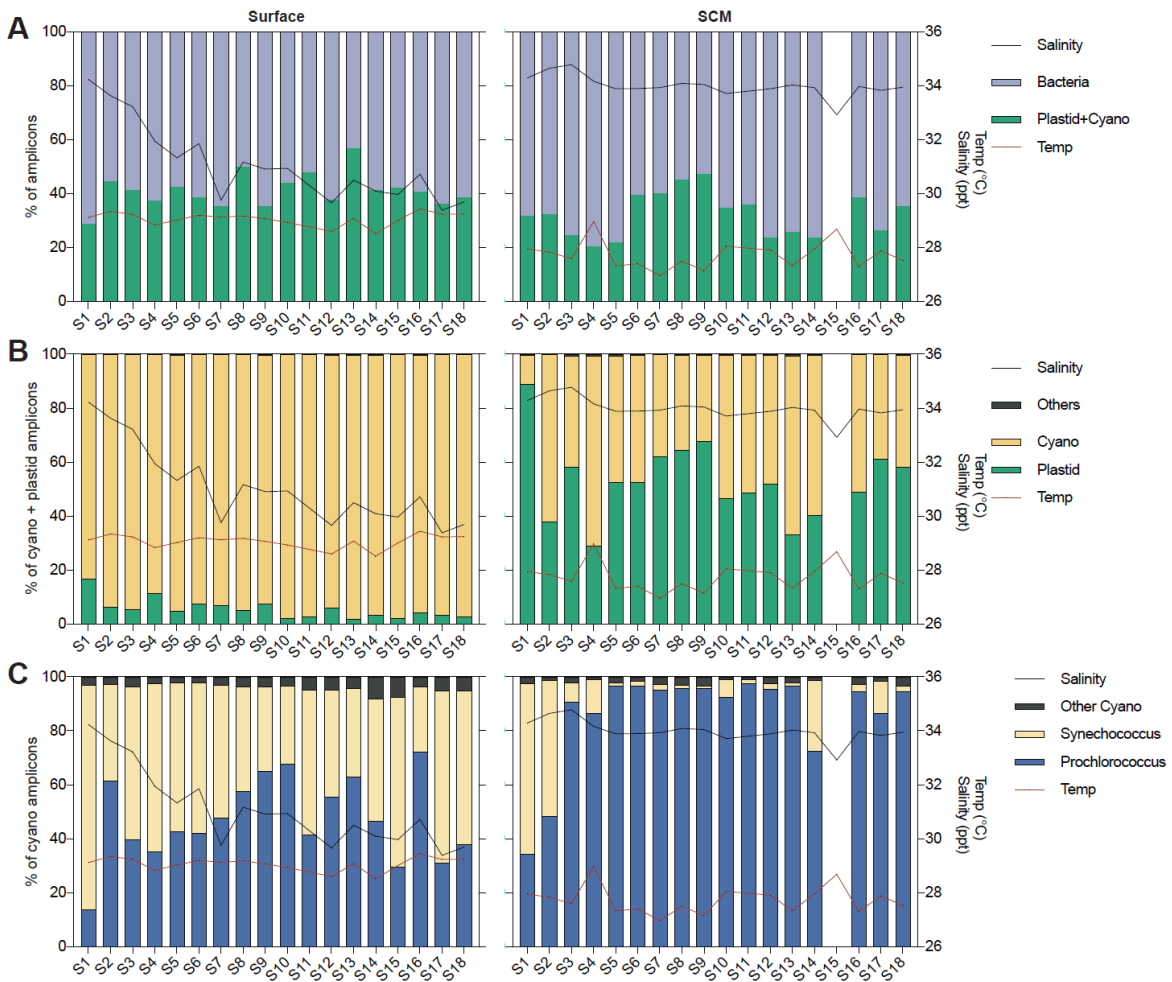




Supp. Fig. 3



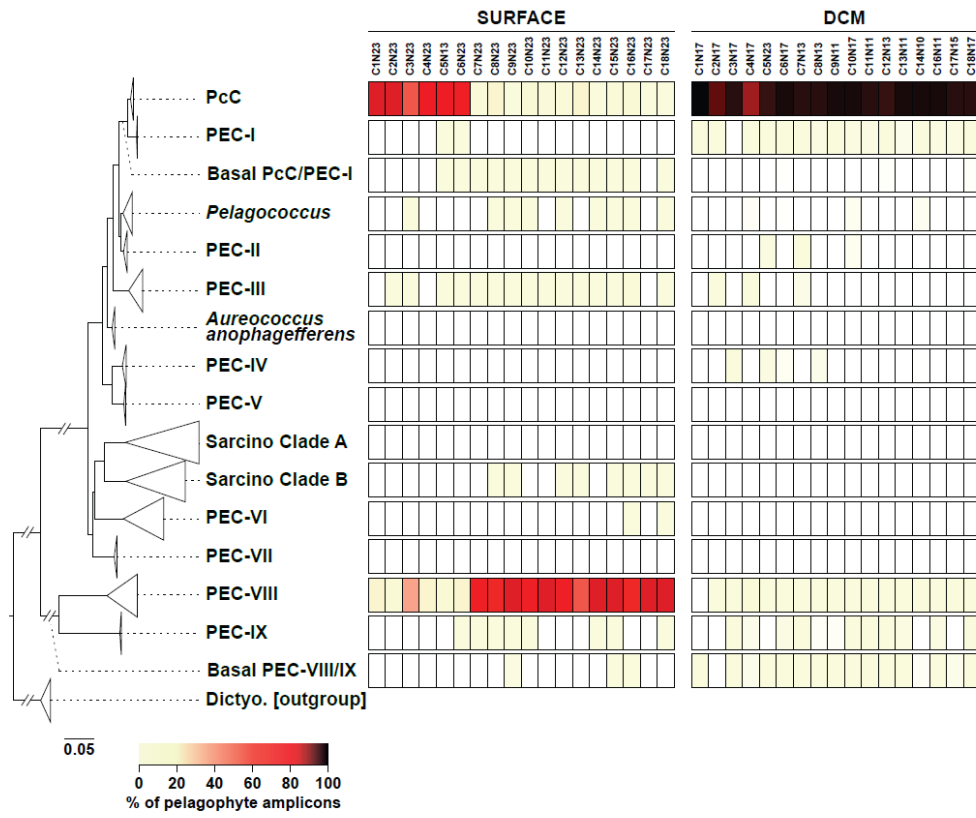
Supp. Fig. 4



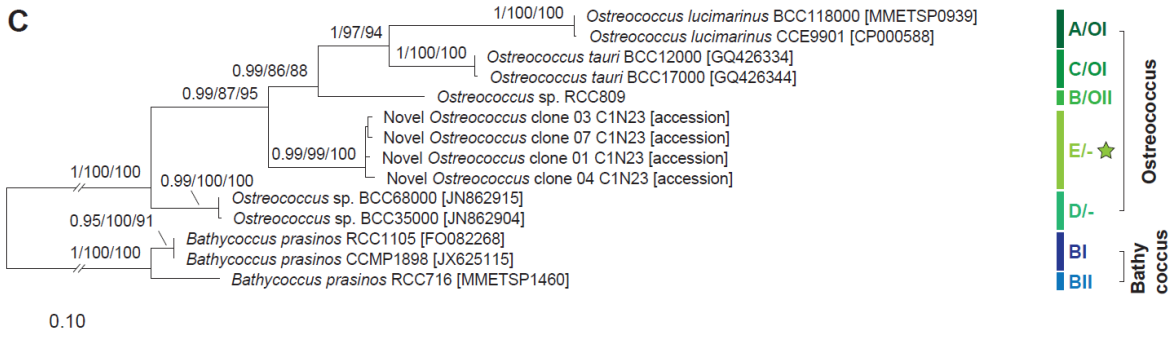
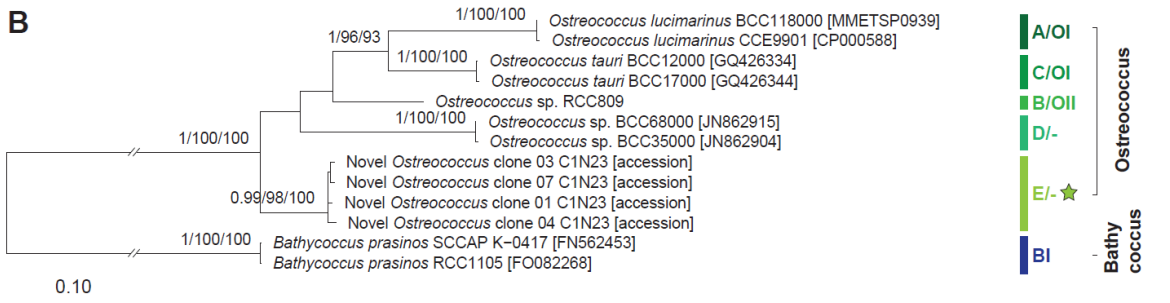
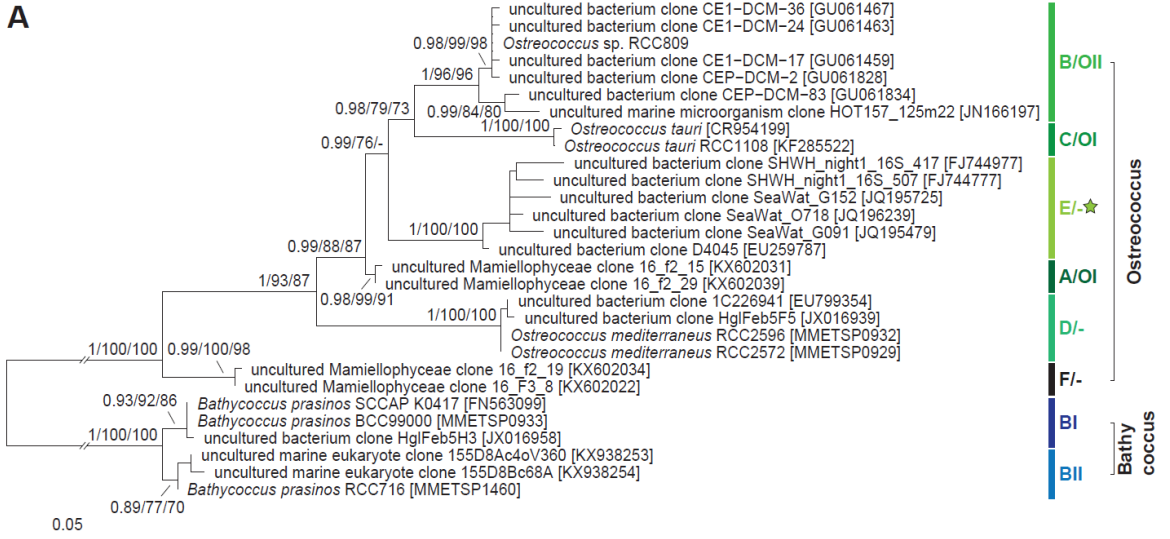
Supp. Fig. 5



# Supp. Fig. 6



# Supp. Fig. 7



# Supp. Fig. 8



# Supp. Fig. 9

