RESEARCH ARTICLE

ENVIRONMENTAL MICROBIOLOGY | Applied Microbiology International

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

Jan Strauss[1] | Chang Jae Choi[1] | Jonathan Grone[1,2] | Fabian Wittmers[1,2] | Valeria Jimenez[1] | Kriste Makareviciute-Fichtner[1] | Charles Bachy[1,3] | Gualtiero Spiro Jaeger[4] | Camille Poirier[1] | Charlotte Eckmann[1] | Rachele Spezzano[1] | Carolin R. Löscher[5] | V. V. S. S. Sarma[6] | Amala Mahadevan[4] | Alexandra Z. Worden[1,2,7]

[1]Ocean EcoSystems Biology Unit, RD3, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany

[2]Faculty of Mathematics and Natural Sciences, Christian-Albrecht University of Kiel, Kiel, Germany

[3]Sorbonne Université – CNRS, Roscoff Culture Collection, FR2424, Station Biologique de Roscoff, Roscoff, France

[4]Physical Oceanography Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA

[5]Nordcee, DIAS, Department of Biology, University of Southern Denmark, Odense, Denmark

[6]Regional Centre, CSIR-National Institute of Oceanography, Visakhapatnam, India

[7]Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts, USA

**Correspondence**
Alexandra Z. Worden, Ocean EcoSystems Biology Unit, RD3, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel, Germany.
Email: azworden@mbl.edu

## Abstract

The Bay of Bengal (BoB) is a 2,600,000 km$^2$ expanse in the Indian Ocean upon which many humans rely. However, the primary producers underpinning food chains here remain poorly characterized. We examined phytoplankton abundance and diversity along strong BoB latitudinal and vertical salinity gradients—which have low temperature variation (27–29°C) between the surface and subsurface chlorophyll maximum (SCM). In surface waters, *Prochlorococcus* averaged $11.7 \pm 4.4 \times 10^4$ cells ml$^{-1}$, predominantly HLII, whereas LLII and 'rare' ecotypes, HLVI and LLVII, dominated in the SCM. *Synechococcus* averaged $8.4 \pm 2.3 \times 10^4$ cells ml$^{-1}$ in the surface, declined rapidly with depth, and population structure of dominant Clade II differed between surface and SCM; Clade X was notable at both depths. Across all sites, *Ostreococcus* Clade OII dominated SCM eukaryotes whereas communities differentiated strongly moving from Arabian Sea-influenced high salinity (southerly; prasinophytes) to freshwater-influenced low salinity (northerly; stramenopiles, specifically, diatoms, pelagophytes, and dictyochophytes, plus the prasinophyte *Micromonas*) surface waters. Eukaryotic phytoplankton peaked in the south ($1.9 \times 10^4$ cells ml$^{-1}$, surface) where a novel *Ostreococcus* was revealed, named here *Ostreococcus bengalensis*. We expose dominance of a single picoeukaryote and hitherto 'rare' picocyanobacteria at depth in this complex ecosystem where studies suggest picoplankton are replacing larger phytoplankton due to climate change.

Jan Strauss, Chang Jae Choi, and Jonathan Grone are co-first authors.

# 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 (Mahadevan, 2016; Subramanian, 1993). 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 ~$1.5 \times 10^{12}$ m$^3$ 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). Nevertheless, the levels of phytoplankton and zooplankton production are adequately sustained to support active fisheries in the Bay of Bengal (McCreary et al., 2013; Paulmier & Ruiz-Pino, 2009). Indeed, human reliance on a particular ocean region is probably nowhere greater than around the Bay of Bengal, upon which some of the densest populations in India, Bangladesh, Myanmar, and Sri Lanka rely 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 & 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 (Madhupratap et al., 2003; Prasanna Kumar et al., 2002; Thushara et al., 2019). With respect to nutrients, although riverine fluxes bring in nutrients, a substantial portion is consumed within a few kilometres from the coast (Krishna et al., 2016; Sarma et al., 2013). Nutrient limitation is exacerbated by strong salinity gradients (Madhupratap et al., 2003) and the system is substantially independent of thermally-driven seasonal variations, that occur in some subtropical gyres, given its consistently warm surface waters (~26–29°C). Furthermore, strong stratification also contributes to a quiescent environment, which may further limit the overall nutrient supply to the system, restricting primary productivity (Prasanna Kumar et al., 2002; Sarker et al., 2020). Thus, in many respects, salinity is considered a leading control on the dynamics of the photic zone, and shapes its shallow, spatially and temporally variable mixed layer (ML) of oligotrophic conditions (Mahadevan et al., 2016).

However, there is still much to learn with respect to the phytoplankton taxa that underpin productivity in the system.

Multiple studies have tackled the primary producer communities in the Bay of Bengal or localized regions within it. These have used a variety of methodologies including satellite data (which reflects the surface layer), pigment analyses, and microscopy-based observations (Gregg & Rousseaux, 2019; Jyothibabu et al., 2015, 2018; Roxy et al., 2016). While important, the methods employed have generally lacked resolution for organisms in the small size fraction, such as *Prochlorococcus* and eukaryotic picoeukaryotes, and those that have used amplicon approaches have generally focused on the surface, for example, 7 m (Larkin et al., 2020, 2022). Another recent amplicon-based study focused on the Bay of Bengal showed high relative abundances of picocyanobacteria and photosynthetic picoeukaryotes at the six stations analysed (Angelova et al., 2019). Additionally, a satellite imaging-based study concluded there has been a shift in the primary producer community from larger eukaryotic phytoplankton to picocyanobacteria, specifically *Prochlorococcus*, in response to decreasing nitrate concentrations (Gregg & Rousseaux, 2019). This community shift has further been projected to critically impact bulk primary production, picoeukaryotic taxa (Chen et al., 2022), export production, and the expansion and intensity of the oxygen minimum zone in the Bay of Bengal (Löscher, 2021). While these studies shed light on possible scenarios and, in some cases, particular phytoplankton community members, they have generally not been performed in a depth-resolved integrated manner nor over the latitudinal expanse of the bay, which has regional variations that are especially sensitive to changes in the global water cycle.

The summer monsoon is the period in which the Bay of Bengal receives the highest rainfall (in the world), contributing to major differences in its physicochemical attributes compared to the other ocean regions at similar latitudes (Wijesekera et al., 2016). Therefore, we examined microbial phytoplankton communities in the Bay of Bengal during the peak southwest summer monsoon in 2015. Samples were collected from 29 sites along the surface, 18 of which were also sampled for vertical depth profiles. These were analysed by flow cytometry and, at two depths (surface and subsurface chlorophyll maximum [SCM]), V1-V2 16S rRNA gene amplicon sequencing was performed. Data from the latter were analysed using phylogenetic approaches to reveal fine-scale phytoplankton distributions, to the level of sub-species for some ecotypes/species. To this end, we further developed high-resolution full-length 16S rRNA gene phylogenetic reconstructions for two major planktonic microbes in this ecosystem, the cyanobacterium *Prochlorococcus* and the picoeukaryote *Ostreococcus*, to provide

insights into new or rarely detected members and their spatiotemporal patterns in connection with environmental conditions. Collectively, our results reveal a 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

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, and riverine nutrient input regions; Figure 1A). Within the top 10 m, salinity declined from 34.2 to 29.4 g kg$^{-1}$ moving northwards (Figures 1B, 2B, and S1A). This large differential lessened in the southward direction, such that surface salinity was significantly higher at stations S1–S4 (33.3 ± 0.8 g kg$^{-1}$) than at stations S5–S18 (30.6 ± 0.7 g kg$^{-1}$; $p < 0.001$, Wilcoxon test). Unlike the strong gradients observed in salinity, surface temperatures were similar across the transect (29.1 ± 0.2°C; Table S1A). It should be noted that we had planned the return to port from the northern sampling region as the period for intensive profile sampling, however, this was negated by the loss of the Niskin-rosette at sea after Station 18. Therefore, bucket samples (B1–B11) were collected at the surface along the return transect of the cruise and environmental data from these generally aligned with the samples from the Niskin rosette that had been collected previously as the ship moved northwards after the start of the cruise (Figures 2A and S1A). Because of the complex water dynamics in this region some samples collected at two different time points (way northward versus return south) exhibited variability in temperature and salinity at the surface, particularly in the mid-latitudes.

Vertical distribution of salinity and temperature across stations indicated that S1–S4 had distinct water mass characteristics from other stations. Specifically, these stations were saltier and warmer at a given density than others, exposing the influence of Arabian Sea High Salinity water upon them (Figure S1B). With respect to the photic zone, stations S5 and S7-18 exhibited a lens of low salinity waters in the upper 10 m overlying more saline waters (Figures 2B and S1A,D). Temperature over the top 100 m ranged from 20.2 to 29.5°C across the transect. It was less variable (≥27°C) between the SCM and shallower depths (Figures 2C,D and S1C).
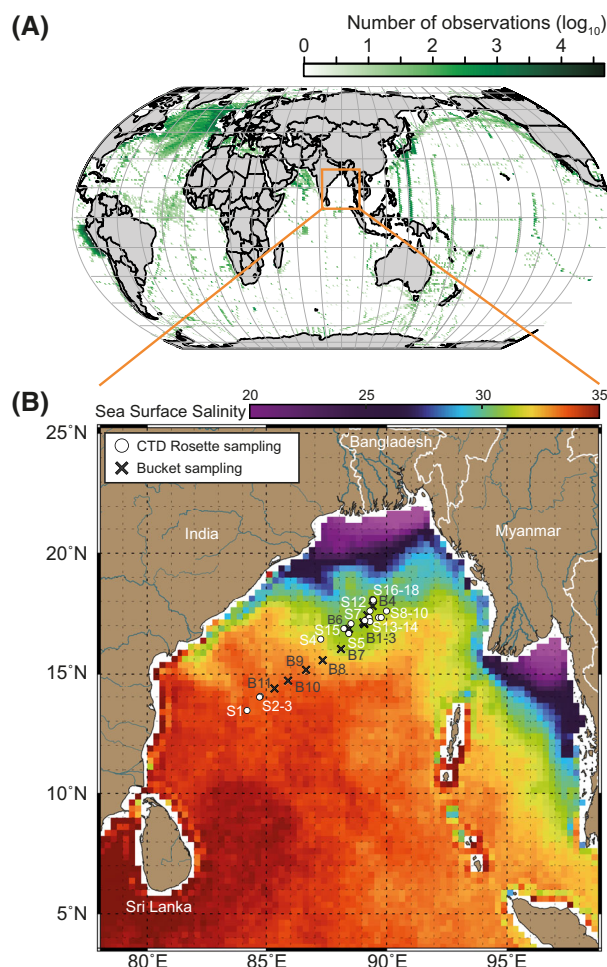


**FIGURE 1** Oceanographic study context. (A) Worldwide phytoplankton observations indicating global spatial biases in biological oceanographic sampling intensity as derived from PHYTOBASE database (https://doi.org/10.1594/PANGAEA.904397) by normalization of observations per area of 1° grid. Note that 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 during the study period. Oceanographic and environmental data up to 200 m depth were measured for the 18 hydrographic stations (S1–S18, profiles, white circles) and 11 bucket samples (B1–B11, surface, black x-marks) were collected from surface waters for salinity, temperature and microbial plankton analysis. Note that due to crowding not all sampling points are labelled with their respective number.

In vivo chlorophyll *a* fluorescence showed that a SCM was present at all stations, ranging from between 30 and 60 m (Figure 2D); the average temperature at the SCM was (27.8 ± 0.6°C). Chlorophyll fluorescence followed patterns in nutrients which varied with depth in the photic zone and along the transect (Figure S2, Table S1A). Dissolved inorganic nitrogen (DIN, sum of nitrate, nitrite, and ammonium) concentrations in surface waters varied from 0.7 μM (station S1) to 4.6 μM
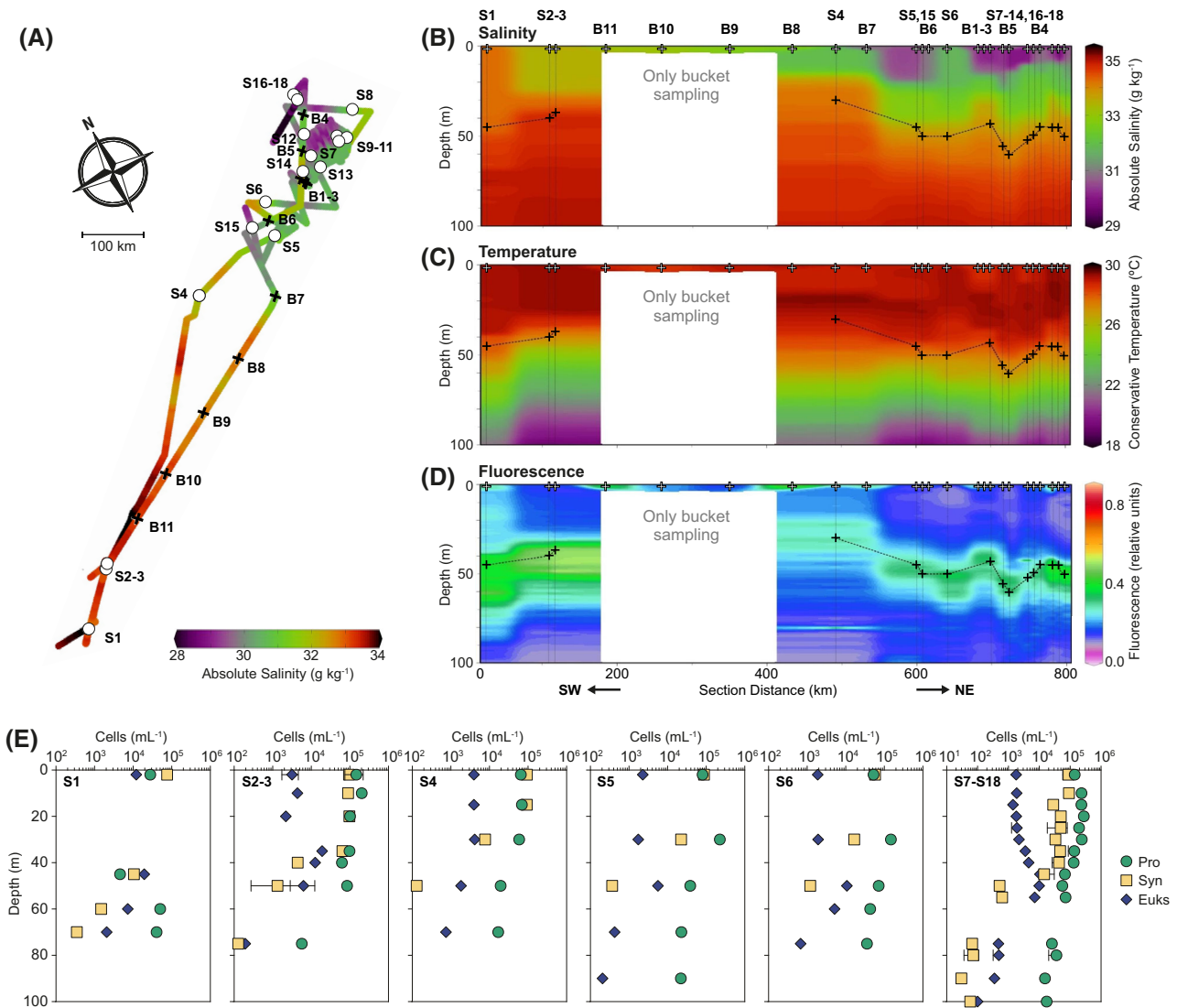
**FIGURE 2** Characteristics of the water column and overall phytoplankton cell abundance. Data is shown in relation to depth from the southernmost (S1) to northernmost (S18) stations in the Bay of Bengal as well as from bucket samples at the surface. (A) Surface salinity plotted with respect to section distances. Interpolated depth profiles for (B) salinity, (C) temperature and (D) *in vivo* fluorescence from chlorophyll *a*. Depths where DNA was collected are indicated (surface, white cross; subsurface chlorophyll maximum, black cross). Positions of stations (S) and buckets (B) are indicated by letters followed by numbering. (E) Average cell abundance of cyanobacteria, specifically *Prochlorococcus* (green), *Synechococcus* (orange) and eukaryotic phytoplankton (blue), from stations with similar characteristics. Note that flow cytometry samples were collected at 4–8 depths depending on the water column structure. Data points and error bars represent the mean and standard deviation, respectively.

(station S15) (Figure S2A–C, Table S1A). With depth, nutrient concentrations generally increased such that in the SCM DIN averaged 8.6 ± 4.0 μM. Similarly, silicate concentrations were 2.1 ± 0.9 μM and 4.0 ± 2.4 μM in surface waters and the SCM, respectively (Figure S2E, Table S1A). Phosphate concentrations averaged 0.4 ± 0.2 μM in the surface waters and 0.6 ± 0.2 μM in the SCM and were generally lower in the northernmost stations (Figure S2D, Table S1A). Finally, in general, oxygen was significantly higher and salinity lower in the surface compared to the underlying waters, again apart from southern stations (Figure S1D,E).

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

In total, four to six depths were analysed by flow cytometry, typically ranging from 1 to 100 m and corresponding to (1) surface, (2) ML, (3) SCM, (4) below SCM, (5) base of oxycline and sometimes an anoxic layer, if reached (Table S1B). Eleven surface samples collected using a bucket (when the CTD was lost overboard; Figures 1B, 2, S3) were also analysed.

Surface waters were numerically dominated by cyanobacteria, with abundances of *Prochlorococcus* and

*Synechococcus* ranging from $2.8 \times 10^4$ to $2.0 \times 10^5$ (mean ± SD = $1.04 \pm 0.46 \times 10^5$) cells ml$^{-1}$ and $4.8 \times 10^4$ to $1.3 \times 10^5$ ($8.36 \pm 2.34 \times 10^4$) cells ml$^{-1}$ (surface CTD samples only, $n = 18$), respectively, across sites (Figure 2E). Abundances in bucket samples B1–B7 were akin to those in nearby Niskin-rosette stations, and in the intervening buckets (B8–11) ranged from $1.1 \times 10^5$ to $2.0 \times 10^5$ cells ml$^{-1}$ and $5.7 \times 10^4$ to $8.9 \times 10^4$ cells ml$^{-1}$ for *Prochlorococcus* and *Synechococcus*, respectively (Table S1A). Over the vertical dimension, *Synechococcus* dropped dramatically in cell abundance, typically by depths of 15–20 m (Table S1B). At the SCM, *Prochlorococcus* ranged from $4.6 \times 10^3$ to $9.8 \times 10^4$ ($6.01 \pm 2.13 \times 10^4$) cells ml$^{-1}$ ($n = 18$), while *Synechococcus* ranged from $3.7 \times 10^2$ to $6.4 \times 10^4$ ($8.30 \pm 15.56 \times 10^3$) cells ml$^{-1}$ ($n = 18$), with the large standard deviation in the latter arising from the fact that seven stations showed <1000 *Synechococcus* ml$^{-1}$ (i.e., $5.51 \pm 1.43 \times 10^2$ cells ml$^{-1}$) at the SCM. Three factors stood out from these analyses; first, the peak abundance of *Prochlorococcus* was typically between 14 and 30 m, not at the surface or SCM. Second, at station S1 the highest abundances of *Prochlorococcus* were seen at 63 m ($5.0 \times 10^4$ cells ml$^{-1}$) and 72 m ($4.0 \times 10^4$ cells ml$^{-1}$), not shallower in the photic zone. Finally, at stations S7–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 (Figure 2E, rightmost panel).

Eukaryotic phytoplankton averaged $2.1 \pm 0.8 \times 10^3$ cells ml$^{-1}$ ($n = 17$) in surface (rosette) samples with station S1 excluded due to it having considerably higher concentrations ($1.2 \times 10^4$ eukaryotic phytoplankton cells ml$^{-1}$; Table S1A). Bucket samples exhibited $2.0 \pm 0.6 \times 10^3$ cells ml$^{-1}$ ($n = 11$). At all stations across the transect, eukaryotic phytoplankton cell abundances manifested a peak coincident with the depth of the SCM, in which they averaged $1.1 \pm 0.5 \times 10^4$ cells ml$^{-1}$ ($n = 18$; Figure 2E). 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 S1A). 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 $101.2 \pm 25.8$, $4.6 \pm 2.1$, and $18.3 \pm 5.4$, respectively

(Table S1A). Furthermore, eukaryotic FALS in the surface waters increased within the transect from south to north.

Our results thus far revealed differences in phytoplankton abundances and population characteristics, such as FALS, the proxy for cell size. However, the extent to which these were attributable to acclimation to light availability, differences in speciation, or both, was unclear. Therefore, we generated 16S V1-V2 amplicons using primers that capture plastids from photosynthetic eukaryotes and cyanobacterial 16S rRNA genes, as well as those from heterotrophic bacteria, and performed taxonomic assignments for all phytoplankton sequences (Table S1C). For *Prochlorococcus* assignments, we used a previously validated *Prochlorococcus* ecotype placement alignment (Sudek et al., 2015) and 16S rRNA gene sequences generated herein that extended into the ITS, alongside several new sequences from published studies that provided for robust ecotype identification based on SNPs (Figures S4–S5). For *Synechococcus* we utilized a reference alignment that had been validated for correct *Synechococcus* Clade placement based on information contained in V1-V2 regions of the 16S rRNA gene (Sudek et al., 2015). For photosynthetic eukaryotes we leveraged published reference alignments (Choi et al., 2020), except for diatoms (Figure S6).

## 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 (Figure S7, Table S1D). *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 (Figure S7C, Table S1D). *Prochlorococcus* dominance was observed at the SCM where it represented $93 \pm 6\%$ of cyanobacteria ($n = 15$, S1–S2 excluded). In the southernmost stations S1 and S2, *Synechococcus* SCM amplicon contributions to cyanobacteria ($57 \pm 10\%$) were higher than those of *Prochlorococcus* ($42 \pm 10\%$) (Figure S7C).

General cyanobacterial patterns were dissected by examining specific ecotypes and clades. Within *Prochlorococcus*, the high-light (HL) II ecotype dominated the surface water community, comprising $99 \pm 1\%$ of *Prochlorococcus* amplicons in both niskin- and bucket-collected samples (Figure 3A, Figure S8A). HLII contributed 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 (Table S1D). At the SCM the HLVI
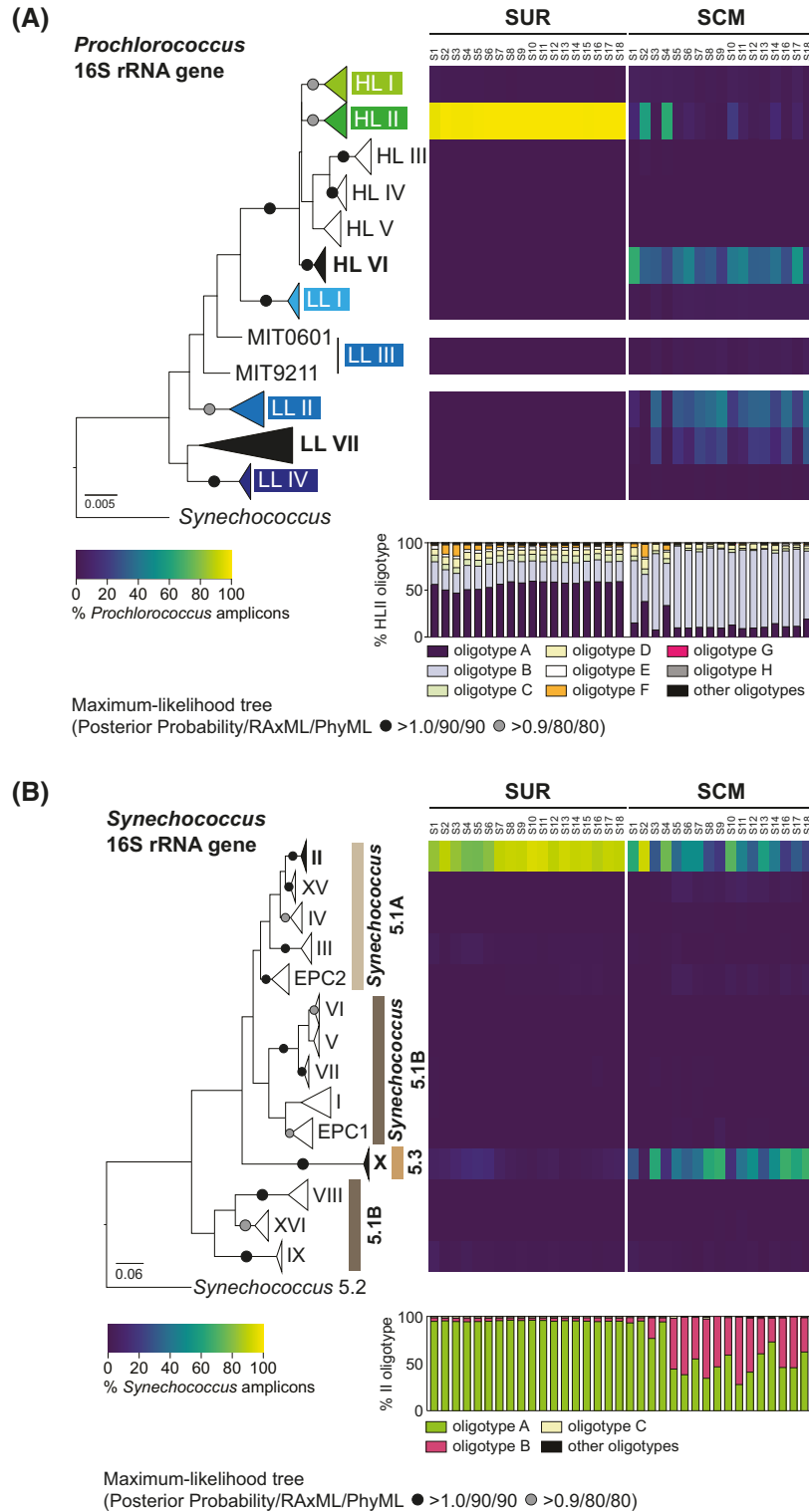
**FIGURE 3** Diversity and relative abundance of cyanobacteria in the Bay Of Bengal. (A) *Prochlorococcus* and (B) *Synechococcus* clade amplicon relative abundances at the surface and at the subsurface chlorophyll maximum (SCM). Each reference tree is based on the full-length 16S rRNA gene sequences and shows clades with a corresponding row in the heat plots which represent the percentage amplicons assigned to the respective *Prochlorococcus* or *Synechococcus* ecotype/clade out of the total of that genus in that sample (station numbers are indicated at the top of each 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.

ecotype formed on average 36 ± 12% of *Prochlorococcus* amplicons, with up to 66% at station S1, and across all stations was followed by the low-light (LL) II ecotype (27 ± 12%). LLVII contributed 15 ± 10% of SCM *Prochlorococcus* amplicons across the stations, with up to 31% at the northernmost station S18, comprising 13 ASVs across all samples (Figure 3A). Systematic or significant changes in *Prochlorococcus* LLVII ecotype amplicon relative abundances over the surface salinity gradient were not observed.

Within *Synechococcus*, Clade II dominated the surface waters by forming 88 ± 5% of the total *Synechococcus* amplicons, followed by *Synechococcus* Clade X at 6 ± 5% and the latter contributed 2 ± 1% in buckets spanning the CTD-section gap (i.e., B8-B11; Figure 3B, Figure S8B, Table S1E). Increased contributions from Clade X were observed at the SCM, forming 46 ± 19% (*n* = 15 excluding stations S2 and S4, 42 ± 21% based on all stations) of the total *Synechococcus* amplicons except for the stations S2 and S4 where it contributed 10 ± 8%. Contributions from Clade II were still high, averaging 43 ± 19% (*n* = 15 excluding stations S2 and S4 where it represented 84 ± 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 (Figure 3A, B). Multiple oligotypes were observed in the HLII amplicons, with two dominants, oligotypes A and B, representing 56 ± 4% and 23 ± 1% of the total HLII amplicons, respectively. At the SCM oligotype B dominated HLII, representing 80 ± 5% (*n* = 15) of HLII amplicons followed by oligotype A (11 ± 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 ± 1% of the Clade II amplicons. Increased contributions from oligotype B were observed at the SCM, averaging 50 ± 13% except for stations S1-S4 where oligotype A represented 90 ± 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 (Figure 4A, B, Figure S8C, Table S1D, E). The surface community mostly comprised stramenopiles forming 70 ± 7% (*n* = 14, stations S5–S18; 63 ± 12%, B8-B11) of plastid amplicons, except for the four southernmost stations S1–S4 where prasinophytes contributed 69 ± 19%

(*n* = 4) and stramenopiles 26 ± 15%. At station S1, 91% of plastid amplicons were from prasinophytes (Figure 4A). In the remaining stations (S5–S18) prasinophytes averaged 17 ± 9% of plastid amplicons at the surface. Such clear shifts were not observed for eukaryotic groups of lower relative abundances between stations S1–S4 and S5–S18. Thus, across the transect, prymnesiophytes (mostly with 95% nt identity to *Chrysochromulina* sp.) contributed 8 ± 4% of plastid amplicons at stations (*n* = 18) and 5 ± 1% in buckets. Less than 1% came from cryptophytes and 2 ± 2% were from other eukaryotic phytoplankton (*n* = 18). At the SCM, prasinophytes dominated eukaryotic phytoplankton amplicons (83 ± 5%), while stramenopiles dropped to 15 ± 5% and amplicons from other eukaryotic groups (prymnesiophytes, cryptophytes, and others) were near detection levels (Figures 4B and S8C).

Among photosynthetic stramenopiles, diatoms contributed 58 ± 11% of the stramenopile amplicons in the surface, followed by dictyochophytes and pelagophytes forming 19 ± 6% and 16 ± 8%, respectively (Figure 4A, Table S1D,E). At the SCM, pelagophytes had higher relative contributions, forming 57 ± 11% of stramenopile amplicons whereas diatoms and dictyochophytes formed 33 ± 11% and 6 ± 2%, respectively (Figure 4B, Table S1C,D). Other stramenopiles, including chrysophytes, represented <5% of stramenopile amplicons (Figure 4A,B, Table S1D).

We further analysed diatoms, dictyochophytes, and pelagophytes using reference trees for high-resolution analysis of these groups (Choi et al., 2020). The diatoms with the highest relative abundance belonged to clade 32, and more specifically the dominant diatom amplicon had >99% nt identity to the pennate *Tryblionella apiculata* (also in clade 32; Figure S6). Dictyochophyte ASVs at the surface were largely from uncultured clades DEC-II and DEC-VI, but also Florenciellales and Rhizochromulina were notable. In the SCM, dictyochophytes (like all stramenopiles) had low relative abundances and generally appeared to be Pedinellales, followed by Florenciellales and uncharacterized groups (Table S1C). 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 ± 10% (*n* = 6) of pelagophyte amplicons, whereas at S7–S18 PEC-VIII contributed 76 ± 8% of pelagophyte amplicons. Pelagophyte amplicons at the SCM were largely assigned to PcC, forming 97 ± 2% of pelagophyte amplicons, and the rest were mostly PEC-VIII (2 ± 2%) (Figure S9).

Viridiplantae (green algal) at the surface were mostly prasinophytes and formed on average 27 ± 12% and 17 ± 8% of plastid amplicons in buckets B1–B7 and stations S5–S18, that is, northwards of S4, respectively. For buckets in the gap between southern and

northern stations the green algal average was 31 ± 12% of plastid amplicons (B8–B11). In these samples the surface prasinophyte community was generally dominated by Class II picoplanktonic *Micromonas* Clade A/B (Figures 4 and S8C, Table S1C–E), represented by the genome sequenced species *M. commoda* (van Baren et al., 2016). Dynamics in southerly stations differed in that prasinophytes were the dominant eukaryotic phytoplankton, and high relative abundances of both *Micromonas* Clade A/B as well as *Ostreococcus* Clade OII were observed.

Class II prasinophytes dominated at the SCM of all stations sampled. *Ostreococcus* Clade OII in particular dominated the SCM, comprising 90 ± 5% of Viridiplantae amplicons and 75 ± 8% of eukaryotic phytoplankton (Figure 4B, Table S1D). *Bathycoccus* and *Micromonas* Clades A/B were also detected at the SCM, albeit at lower relative abundances. However, at other stations (S3–S18), contributions from *Ostreococcus* Clade OII were low in the surface (9 ± 12% of Viridiplantae, $n = 16$), and, instead, *Micromonas* Clades A/B

exhibited the highest relative abundances, together forming 56 ± 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–C, Figure 4C). OII oligotype A dominated SCM OII amplicons, forming 94 ± 3% ($n = 14$, stations S5-S18), but less at stations S1–S4 (56 ± 17%). In the latter, increased contributions from oligotypes B and C were observed, forming 37 ± 19% and 6 ± 4% of OII amplicons, respectively. At the surface, oligotype A generally contributed the most, comprising 55 ± 6% (S4–S6, $n = 3$) and 92 ± 13% (S7–S18, $n = 13$) of Clade OII amplicons. At the three southernmost stations, Oligotype A contributions were lower (20 ± 6% S1–S3, $n = 3$), and, instead, oligotype B dominated (65 ± 5% at S1–S3, $n = 3$) and C was also notable.

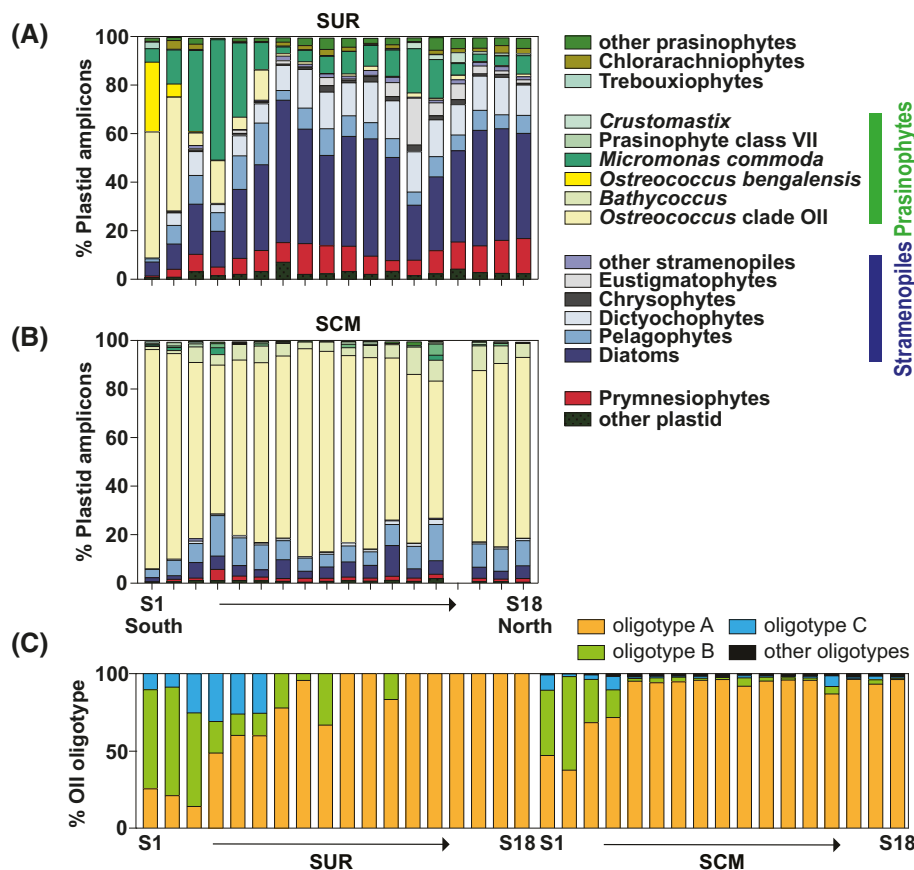As noted above, station S1 stood out from all other stations at the surface (Figure 4A). At this station



**FIGURE 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) with detailing of groups within the prasinophytes and stramenopiles expressed as the percentage of their amplicons relative to all eukaryotic phytoplankton amplicons in surface waters and the subsurface chlorophyll maximum (SCM), respectively. (C) Microdiversity within the most abundant prasinophyte taxon, *Ostreococcus* Clade OII, in the Bay of Bengal. Note that dinoflagellates are poorly resolved using plastid-based methods, see also (Bennke et al., 2018; Green, 2011). They have been reported in low abundances in the Bay of Bengal and can be abundant at sites in the eastern bay (Naik et al., 2011) based on microscopy. Additionally, while not established, it is thought that the primers used herein may under-represent prymnesiophytes (haptophytes).

*Ostreococcus* OII had high relative abundances at the surface, but a new group of prasinophyte amplicons was notable as well. The latter were placed at a last common ancestor (LCA) node in the *Ostreococcus* region of the prasinophyte tree in initial phylogenetic analyses and due to their uniqueness were then explored further using additional sequencing approaches.

## Phylogenetic identification of a novel *Ostreococcus* clade

We next investigated the LCA sequences detected at an *Ostreococcus* node in the prasinophyte reference tree. These amplicons ranged from 91% to 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 16S V1-V2 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 statistically supported clade distinct from other *Ostreococcus* clades/species (Figure S10A). Next, we incorporated 16S V1-V2 rRNA amplicon sequences from the newfound LCA *Ostreococcus* clade, as well as sequences identified as potentially novel prasinophytes from the Arabian Sea (Fuller et al., 2006) and the Bay of Bengal (Angelova et al., 2019), into the phylogenetic analysis (Figure S10B). Sequences from the prior reports clustered within *Ostreococcus* Clade B/OII, which included the cultured species *Ostreococcus* sp. RCC809 and *Ostreococcus* RCC393, whereas the LCA 16S V1-V2 ASVs we found clustered within statistically supported Clade E.

Due to the 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, because this region is often used to delineate Class II prasinophyte species (Marin & Melkonian, 2010). To accomplish this, we developed a forward primer specific to *Ostreococcus* 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 that was distinct from the other *Ostreococcus* and from environmental full-length 18S rRNA gene sequences (Figure 5A). Phylogenetic analysis of the

ITS supported these results (Figure S11A,B). We next analysed the ITS2 secondary structure, which contained the extra helix so far only seen in *Ostreococcus* and *Bathycoccus* (Figure 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 (Figure 5C) pointing to designation as a new species, which we propose to be informally named *Ostreococcus bengalensis* nomen nudum.

To this point the only other *O. bengalensis* full-length sequences recovered from GenBank had little contextual information apart from location. For the sequences from Georgia and San Diego USA, 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 Genbank 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. *O. bengalensis* appears in V4 data from the 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 (Figure 6A). *O. lucimarinus* exhibited highest relative abundances in some continental shelf regions and in upwelling zones in the eastern Pacific (Figure 6C). Lastly, *Ostreococcus* Clade OII was observed at open ocean sites of the Atlantic and Pacific (Figure 6B).

## Statistical analysis of communities and potential environmental drivers

Statistical analyses showed, not surprisingly, that the taxonomic composition between the surface and SCM differed at a significance level of $p < 0.001$ for both cyanobacteria (ANOSIM test statistic $R = 0.943$) and eukaryotic phytoplankton (ANOSIM test statistic $R = 0.937$). Relative abundance plots supported that microbial communities were most different at the southernmost stations S1–S4, and this difference was especially strong at the surface (Figures 3 and 4A–C and S7). Compositional data analysis confirmed that the cyanobacterial ecotypes *Synechococcus* Clade II and *Prochlorococcus* HLII most strongly influenced taxonomic compositions at the surface, whereas the
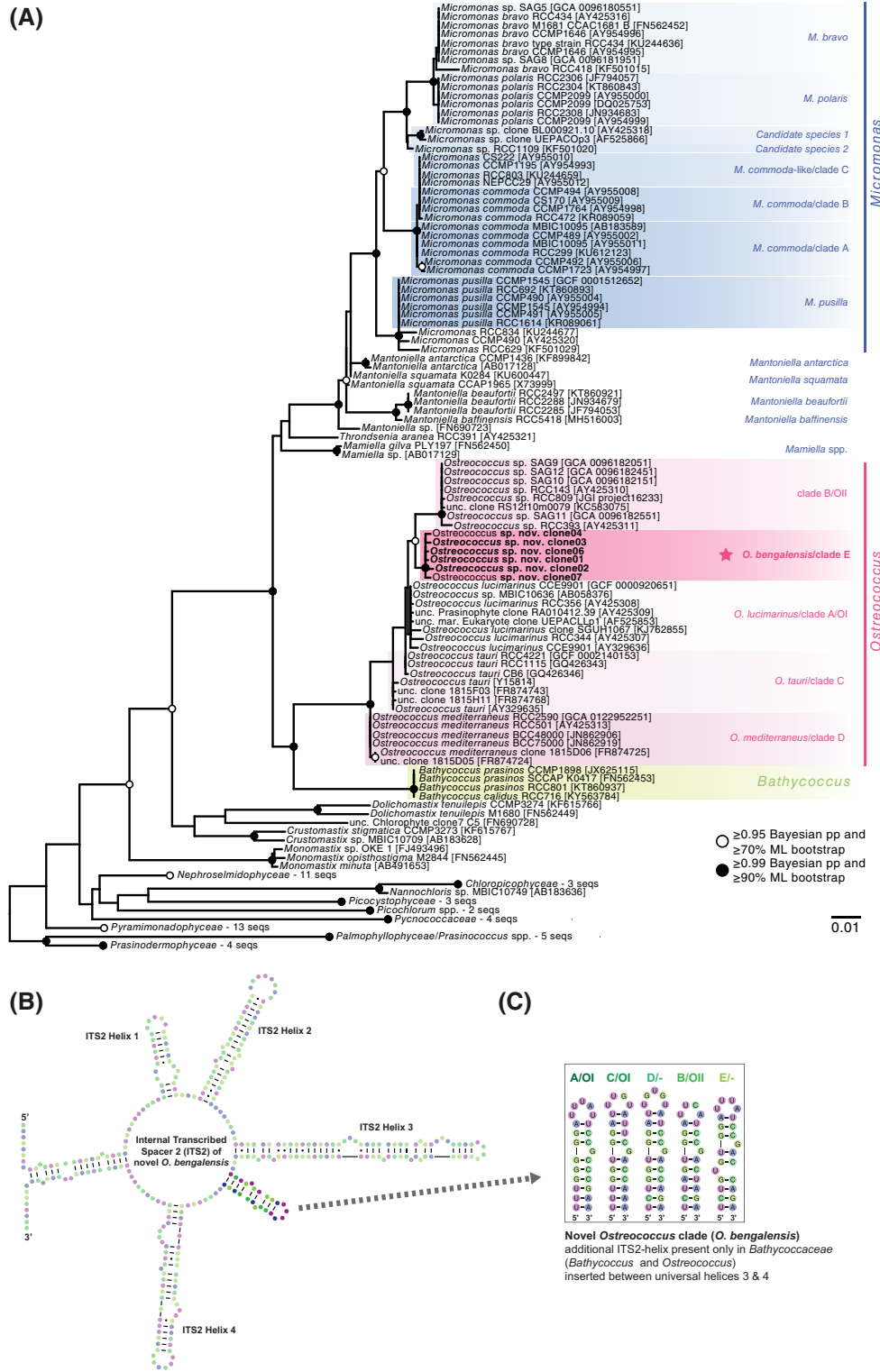
**FIGURE 5** Phylogeny of the prasinophytes with an emphasis on Mamiellophyceae (Class II) and specific attributes of the 'new' *Ostreococcus* species. (A) Maximum likelihood (ML) phylogenetic reconstruction of Mamiellophyceae and other prasinophytes inferred from an alignment of nearly complete nuclear 18S rRNA gene sequences (i.e., >1450 bp) that incorporated 145 total sequences and analyzed 1466 homologous positions. The 18S rRNA gene sequences generated herein for the new *Ostreococcus* are included (bold). Node support was evaluated from 1000 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 ≥70% ML bootstrap and ≥0.90 Bayesian posterior probability (open circles) and ≥ 90% ML bootstrap and ≥0.99 Bayesian posterior probability (black filled circles) are indicated. *Ostreococcus* clades (pink) are highlighted together with other widespread Mamiellophyceae genera, *Micromonas* (blue) and *Bathycoccus* (green), and the entire *O. bengalensis* clade/Clade E is denoted by a star. The basal branches were collapsed, and we used the prasinodermophyte sequences as an outgroup for display purposes only. (B) The secondary structure for the Internal Transcribed Spacer 2 (ITS2) in the newly discovered *Ostreococcus* (i.e., *O. bengalensis*) is depicted as a reference for defining Mamiellophyceae speciation, as established by (Marin & Melkonian, 2010). (C) Comparison of the Bathycoccaceae family-specific additional ITS2 helix, located between universal helices 3 and 4 from each *Ostreococcus* clade.
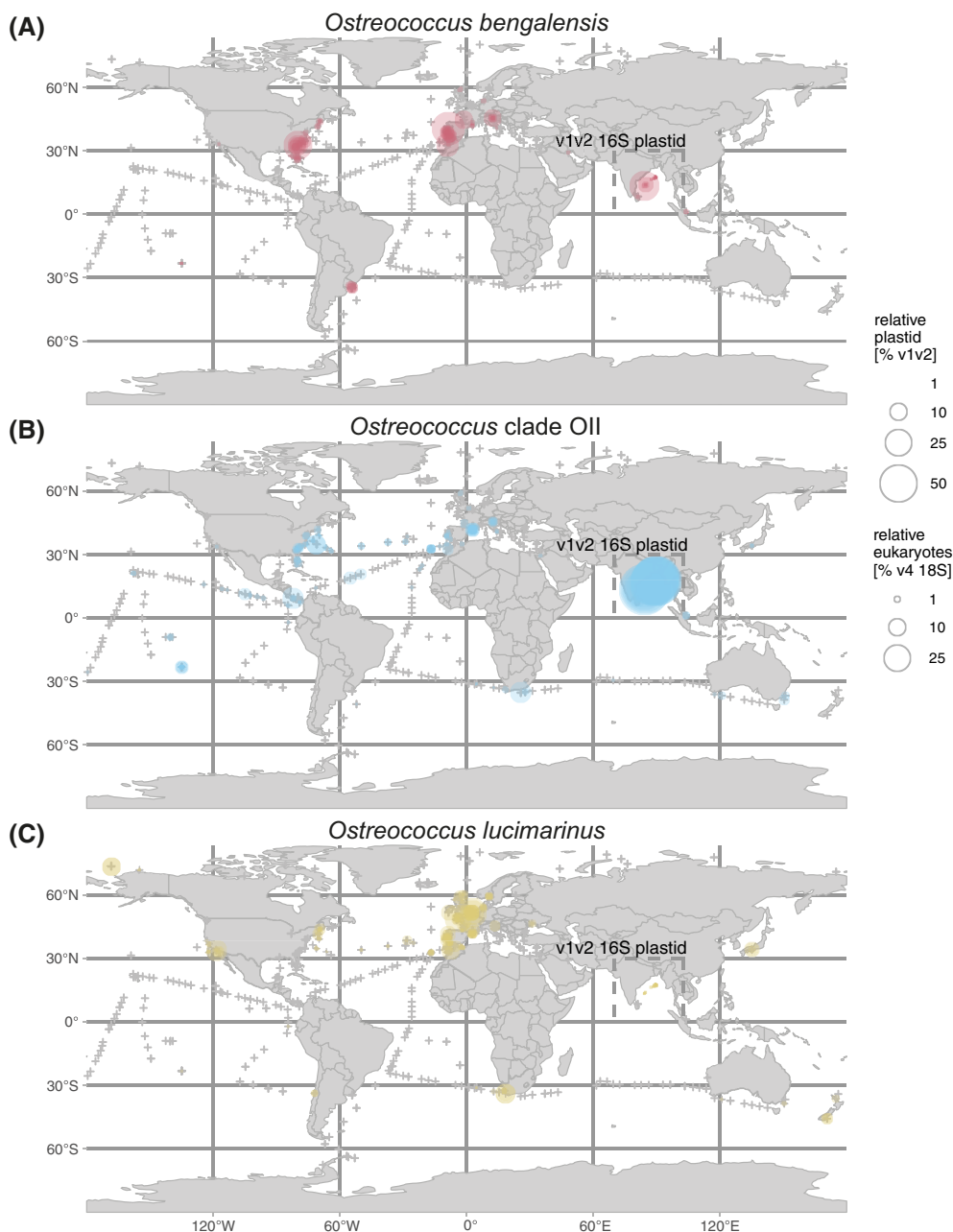
**(A)** *Ostreococcus bengalensis*

**(B)** *Ostreococcus* clade OII

**(C)** *Ostreococcus lucimarinus*

**FIGURE 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), which is biased towards surface samples. 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.

*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 (Figure S12A).

We next performed a permutation test for canonical correspondence analysis (CCA) and found the ordination was significant and the environmental variables salinity, nitrite, depth, temperature, silicate, and fluorescence, with the first four parameters explaining most of

the variance in the observed ASV matrix (Figures 7, S12B–D, Table S1F). The CCA for *Prochlorococcus* revealed two broad clusters of ASVs indicating surface and SCM communities (Figure 7A). Surface ASVs consisted exclusively of HL ecotypes, with HLII accounting for most data points. Some HL ASVs, specifically from HLI, HLII, HLIV, and HLVI, clustered apart from the CCA region most indicative of surface waters. Additional Spearman's rank correlation analyses between
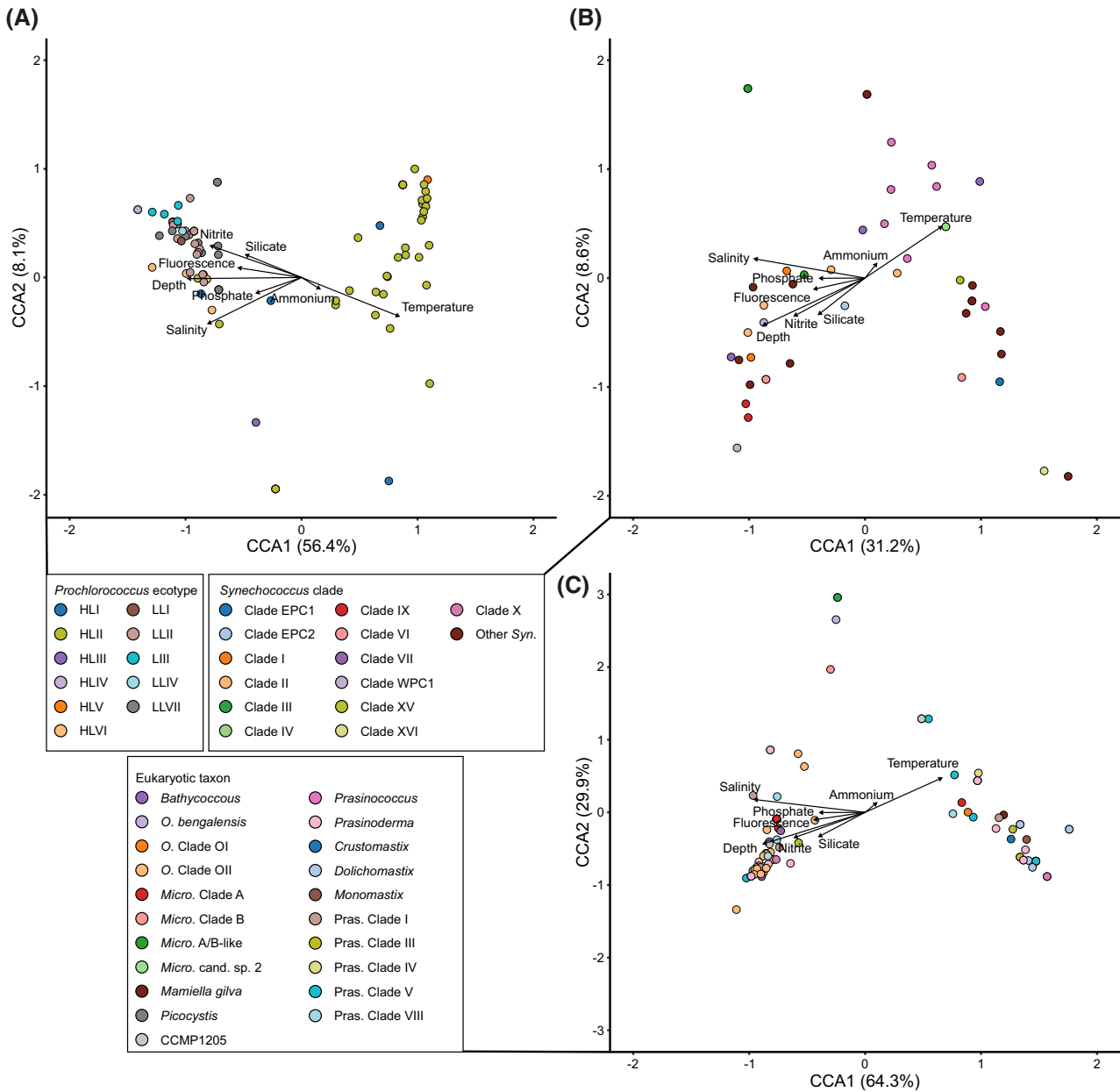
**FIGURE 7** Canonical correspondence analysis (CCA) of amplicons assigned to phytoplankton taxa at the 18 CTD stations including samples from the surface and subsurface chlorophyll maximum (SCM) layers and environmental parameters. (A) CCA plot of distribution of amplicon sequence variants assigned to *Prochlorococcus* ecotypes along environmental parameters. (B) CCA plot of distribution of amplicon sequence variants assigned to *Synechococcus* clades along environmental parameters. (C) CCA plot of distribution of amplicon sequence variants assigned to prasinophytes along environmental parameters.

relative abundances of *Prochlorococcus* ecotypes (including all HLII oligotypes) and environmental variables indicated that the most abundant HLII oligotype, oligotype B, exhibited significantly positive correlations with salinity ($\rho(34) = 0.70$), nitrite ($\rho(36) = 0.69$) and nitrate ($\rho(35) = 0.59$) 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: $\rho(34) = -0.69$; oligotype C: $\rho(34) = -0.60$), nitrite (oligotype A: $\rho(36) = -0.68$; oligotype C: $\rho(36) = -0.69$) and nitrate (oligotype A: $\rho(35)$

$= -0.69$; oligotype C: $\rho(35) = -0.62$) at a significance level of $p < 0.001$. In the case of *Synechococcus* CCA analysis, the ASVs did not show clearly distinguishable clusters in association with the environmental vectors, apart from Clade X which was distinctly surface associated (Figure 7B). Apart from Clade X, other *Synechococcus* clades were not exclusively associated with either surface or SCM communities.

CCA analyses of eukaryotic phytoplankton rendered patterns that generally delineated ASVs that appeared to be surface associated, from others

associated with variables connected to deeper portions of the photic zone. Most of the prasinophyte ASVs clustered together within the SCM community (Figure 7C). However, most taxa were represented in both surface and SCM clusters. *Ostreococcus* Clade OII ASVs were associated with depth and salinity as the primary parameters, followed by nitrate and silicate, although seen at the surface in the southernmost station. Additional CCAs for stramenopile groups as well as detailed breakdowns for clades within the dictyochophytes and pelagophytes were performed (Figure S12). Both breakdowns suggested distinct surface and SCM communities (Figure S12A–C) that allowed us to discern that *Pseudochattonella* and *Rhizochromulina* ASVs were mainly associated with characteristics of the surface layer (Figure S12B). For the pelagophytes, members of the abundant *P. calceolata* Clade (PcC) were strictly associated with SCM characteristics (Figure S12C). Despite its low variation, temperature (or co-associated factors) was a significant factor in distinction of surface from SCM phytoplankton communities (both cyanobacterial and eukaryotic), whereas salinity, nitrite, depth, fluorescence, and silicate were each significant for SCM-associated ASVs (Figure 7, Table S1F).

## DISCUSSION

Substantial efforts have been dedicated to characterizing water mass dynamics of the Bay of Bengal and the South Asian summer monsoon, with its tremendous ocean–atmosphere–land feedback (Mahadevan et al., 2016). These dynamical attributes influence the biota that reside in this marine ecosystem, including the primary producers that form the base of the food network. Here, we strove to elucidate Bay of Bengal phytoplankton communities within the context of physicochemical dynamics during the southwest summer monsoon. To this end, we examined the phytoplankton community using flow cytometry and V1-V2 16S rRNA amplicon analyses. This marker was advantageous for capturing both cyanobacteria and photosynthetic eukaryotes, and because the 16S rRNA gene has less copy number variability than, for example, the 18S rRNA gene (Choi et al., 2020; Fuller et al., 2006; Gong et al., 2020; Needham & Fuhrman, 2016). Combining this approach with targeted full-length small subunit rRNA gene and internal transcribed spacer sequencing enhanced insights into phytoplankton diversity and factors that shape shifts in the primary producer community.

A dominant feature during our study was strong salinity-driven stratification which has been reported to inhibit nutrient supplies from subsurface to surface waters (Sarma et al., 2016). We also observed low near-surface concentrations of nitrate and phosphate,

particularly in the northerly region of the transect, where the distinction between the surface layer and SCM was largest in terms of salinity and density, and riverine freshwater input was strongest (Figures 2, S1, and S2, Table S1A). For southern stations, where the influence of Arabian Sea waters was observed (Figure S1), gradients in salinity were less sharp, 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.

Recent publications have pointed to the potential importance of *Prochlorococcus* in the Bay of Bengal, including possible amplification due to climate change (Gregg & Rousseaux, 2019; Löscher, 2021). Here, we examined the 13 *Prochlorococcus* lineages previously resolved through phylogenetic analysis of the ITS, and termed HLI to VI and LLI to VII (Biller et al., 2015; Johnson et al., 2006; Rocap et al., 2003). We validated our placement methods by characterizing SNPs in the V1-V2 region of the 16S rRNA gene (Figure S4) and, for some lineages, the ITS (Figure S5), building upon a validated approach for *Prochlorococcus* ecotype (and *Synechococcus* Clade) discrimination (Sudek et al., 2015). In previous studies, several ecotypes have been dissected using genomic analyses, which have highlighted differing gene sets for use of inorganic nitrogen sources (Berube et al., 2019; Johnson et al., 2006; Rocap et al., 2003). Additionally, ecotype distributional patterns have been connected to oceanic regions and environmental conditions, such as light availability and temperature (Biller et al., 2015), as well as primary production contributions (Thompson & Kouba, 2019). Thus, it is well recognized that a complex set of factors underpin the niche differentiation of *Prochlorococcus* ecotypes.

Extending beyond Bay of Bengal surface water studies showing relative amplicon dominance by HLII during the spring intermonsoon season (Larkin et al., 2020; Pujari et al., 2019; Wei et al., 2020), we observed differing HLII oligotype ratios between the surface and SCM (Figure 3A). This potentially reflects niche differentiation given that SNPs can reflect population biology that connects to ecology (Cordero & Polz, 2014) and that, for example, nitrate assimilation genes have been suggested for some HLII ecotype members (Berube et al., 2019). Potential differences in niche occupation on a sub-ecotype level within HLII have been reported in the North Pacific and North Atlantic based on ITS amplicons (Kashtan et al., 2017; Larkin et al., 2016). Since our HLII 16S V1-V2 amplicons used for oligotyping do not extend into the ITS, the observed microdiversity could not be compared with the prior studies, but the possibility of similar/ identical oligotypes being present in multiple oceanic regions warrants further investigation. More generally, *Prochlorococcus* at the surface showed similarities to

communities at the Bermuda Atlantic Time Series (BATS) and Hawaii Ocean Time Series (HOT, station ALOHA) sites. During strong seasonal thermal stratification of the BATS photic zone (Steinberg et al., 2001), HLII also dominates in the surface, followed by HLI (Kashtan et al., 2014; Malmstrom et al., 2010; Zinser et al., 2007), and it appears to always dominate in the surface at ALOHA (Thompson et al., 2018) more akin to our Bay of Bengal findings.

Our results capture an unprecedented view into *Prochlorococcus* diversity below the surface skin (e.g., upper 7 m) of the Bay of Bengal. At depth, we found that *Prochlorococcus* ecotypes differed substantially from the DCM at BATS and ALOHA. Specifically, we observed high relative abundances of two *Prochlorococcus* ecotypes, HLVI (the SCM dominant) and LLVII (Figure 3A), which have been considered 'rare' (Biller et al., 2015; Huang et al., 2012; Jiao et al., 2014). Interestingly, in CCA analysis *Prochlorococcus* HLVI ASVs clustered closely with low-light adapted clades in the SCM (Figure 7A). Adaptation of HLVI to low-light conditions has been suggested before (Berube et al., 2019; Huang et al., 2012), as indicated by our statistical analyses. Unfortunately, little else is known about the ecology and distributions of HLVI which were discovered in the middle-to-lower photic zone (75–150 m) of the South China Sea and transcripts from an HL-like *Prochlorococcus* population from 125 m at ALOHA (Shi et al., 2011) may come from HLVI cells. Additional complexity comes in HLVI having been reported in North Pacific surface communities, albeit at low relative abundances (Larkin et al., 2016). The 16S rRNA gene phylogeny constructed herein supports a basal position among HL ecotypes (Figure 3A), corresponding with the possibility that HLVI may be functionally closer to LL clades and potentially capable of nitrate assimilation (Berube et al., 2019; Huang et al., 2012). We also found high relative abundances of LLII in the Bay of Bengal SCM, which is also present at BATS and ALOHA (Malmstrom et al., 2010). In the context of the current knowledge of these ecotypes, our findings raise questions about the aspects of genomic and metabolic potential in HLVI that lead to a rivalling of LLII at the SCM in the Bay of Bengal (Figure 3A).

Another largely unknown *Prochlorococcus* ecotype, LLVII, was also observed at the SCM in the Bay of Bengal, forming ≤31% of *Prochlorococcus* amplicons (Figure 3A). *Prochlorococcus* LLVII (formerly clade NC1, Non-Cultured 1) was first noted based on ITS sequences in the lower photic zone of subtropical waters (Martiny et al., 2009). It is highly diverged from other ecotypes and formed a monophyletic group based on our 16S rRNA phylogenetic analysis, although with weak statistical support (Figure 3A). Three LLVII reference sequences were generated herein; two others came from 125 m at Station ALOHA, and two came from an Arabian Sea OMZ study (see

Experimental procedures). In our samples, one ASV assigned to LLVII dominated across all stations and depths sampled (69 ± 42% of LLVII). Statistical analyses indicate that LLVII (as well as the other low-light adapted clades and HLVI) relative amplicon abundances correlate positively with salinity and nitrite + nitrate concentrations, but negatively with temperature and oxygen (Table S1F). Nitrate availability may therefore play a role in vertical distributions observed here, although other factors or co-associated environmental factors could be responsible. Overall, the ecotype distributions at depth appear to be influenced in part by the degree of Arabian Sea water, but broader scale differences from, for example, BATS and ALOHA, likely arise from the different mechanisms of stratification (leading to warmer waters at depth in the Bay of Bengal).

*Synechococcus* is generally less abundant than *Prochlorococcus* in the open ocean (Flombaum et al., 2013), but in the Bay of Bengal these genera had similar average abundances at the surface, below which *Synechococcus* declined sharply (Figure 2E). As occurs in the Pacific subtropical gyre (Kent et al., 2019; Xia et al., 2019), *Synechococcus* Clade II dominated, with oligotype partitioning between surface and SCM waters (Figure 3B). Notable contributions for Clade X at the SCM (Figure 3B) fit with the inference that it occupies a similar niche with Clade II, inhabiting warm, oligotrophic tropical and subtropical open oceans (Ahlgren & Rocap, 2012; Sohm et al., 2016). Herein, statistical analyses indicated that Clade X is more widely present deeper in the photic zone in association with increasing inorganic nitrogen and salinity ($p < 0.001$, Spearman's rank correlation). Our results indicate complex population biology and environmental variables drive *Synechococcus* fine-scale diversity.

Although several studies have investigated eukaryotic phytoplankton in the Bay of Bengal, little is known about their molecular diversity. 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. Prior amplicon-analysis suggested *Ostreococcus tauri* was highly abundant, followed by the diatom *Ditylum brightwellii*, pelagophytes and prymnesiophytes within the Braarudosphaeraceae (Angelova et al., 2019). Prior microscopy-based studies of diatoms in the bay have reported *Thalassiontrix*, *Chaetoceros*, *Thalassionema*, *Navicula*, *Thalassiosira, Coscinodiscus*, *Skeletonema* and *Nitzschia angularis* as dominants (Paul et al., 2008; Ramaiah et al., 2010). Here, among stramenopiles, an ASV matching the pennate diatom *T. apiculata* (unofficial synonym *Nitzschia apiculata*) had highest relative abundances among diatoms at several stations (Figure S6, clade 32). Until now, *T. apiculata* is known from, for example, rivers along the west coast of North America (Gastineau

et al., 2021) and has not been reported in the Bay of Bengal or other oceanic photic zones. Our blast searches using 16S rRNA gene sequences from cultured *T. apiculata* as a query recovered hits with 98% nt identity to multiple marine environmental clones, indicating *T. apiculata* like diatoms may be present in warm, low salinity waters akin to the lower salinity stations herein although physical transport from coastal regions of the bay may also have influenced our results. Northwards of station S6, diatoms (particularly *T. apiculata*-like ASVs) generally formed higher proportions of total eukaryotic phytoplankton amplicons (surface only, Figure 4). The spatial differences in diatom proportions within surface waters were, however, not reflected in nutrient data as both Si:N and silicate concentrations were lowest in diatom rich northern stations S17 and S18. This may relate to enhanced Si drawdown, since freshwater inputs were strongest there, and riverine nutrients generally seemed to be taken up rapidly in close proximity to the shore (Krishna et al., 2016; Sarma et al., 2013).

Among non-diatom stramenopiles, 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 dominant dictyochophytes in our study were uncultured Clades (Choi et al., 2020), but Florenciellales were also present—an order that includes *Florenciella parvula*, a mixotrophic flagellate that feeds on picoplankton (Li et al., 2021) (Table S1C). Likewise, *P. calceolata* Clade (PcC) and Pelagophyte Environmental Clade VIII (PEC-VIII) have not been reported in the Bay of Bengal. We observed sharp transitions among pelagophytes at the surface, with PcC contributing the most at stations S1-S6 whereas PEC-VIII dominated S7-S18 (Figure S9). *P. calceolata* has been increasingly recognized as a widely distributed oceanic photosynthetic picoeukaryote (Duerschlag et al., 2022; Guérin et al., 2022; Worden et al., 2012) that is particularly important at the SCM/DCM in mesotrophic and oligotrophic stations in the Eastern North Pacific, the Sargasso Sea (Choi et al., 2020; Dupont et al., 2015), and the South Pacific Subtropical Ocean (Duerschlag et al., 2022). Surprisingly, unlike these prior studies, here, *P. calceolata* was not among the dominant eukaryotic phytoplankton at the SCM. Rather, the prasinophyte *Ostreococcus* Clade OII was the clear dominant (representing 75 ± 8% of plastid amplicons) at the SCM and *Micromonas commoda* was relatively abundant in the surface (Figure 4A).

Chlorophytes (often used to refer to all green algae or prasinophytes) and even 'Mamiellales' (now Mamiellophyceae) have been reported in the Bay of Bengal. *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). QPCR studies have demonstrated that Clade OI (Clade A; represented by genome sequenced *O. lucimarinus*) is present in cooler mesotrophic and coastal waters versus Clade OII (Clade B), the dominant herein. Clade OII is a more oceanic ecotype, which is present throughout the water column during periods of mixing at BATS and ALOHA before being confined to the DCM during stratified periods (Demir-Hilton et al., 2011; Limardo et al., 2017). In a recent study of the Bay of Bengal, abundant eukaryotic plastid amplicons were further classified using PhytoRef (Decelle et al., 2015), which rendered identification of one dominant as being *O. tauri* (Angelova et al., 2019). Upon checking the sequences attributed to *O. tauri* in that study, we found they clustered instead with environmental sequences from Fuller et al. (2006), as well as with *Ostreococcus* sp. RCC393 and *Ostreococcus* sp. RCC809 (Figure S9), thereby being more closely related to *Ostreococcus* Clade OII (also known as Clade B). This brings amplicon data from Angelova et al. (2019) into alignment with our findings. Dominance based on relative amplicon abundances is supported by low FALS (flow cytometry measurements) for SCM eukaryotic phytoplankton cells, again indicative of *Ostreococcus*. Collectively, to our knowledge, this is the first time such a massive dominance of a single eukaryotic species, that is, *Ostreococcus* Clade OII, has been observed across multiple spatially separated SCM/DCM stations.

While we found OII and not OI, consistent with prior work showing these two species rarely co-occur, multiple lines of evidence supported presence of a fifth *Ostreococcus* Clade in the Bay of Bengal (Figures 5 and S10–S11). After acquiring full-length 16S rRNA gene sequences for the LCA taxon seen in amplicon data, phylogenetic analyses demonstrated it forms a statistically supported, independent *Ostreococcus* clade, Clade E. It should be noted that PhytoRef has a sequence (FJ744977) that is placed within the *O. bengalensis* clade in our phylogenetic analysis of partial-length sequences (Figure S10) but is misnamed "*O. tauri*" in that database. Other lines of evidence supporting the new Bay of Bengal clade being an own species came from our environmental 18S-ITS1-5S-ITS2-to partial 28S rRNA clone sequencing and phylogenetic analysis (Figures 5A and S11). Because *Bathycoccus* and *Ostreococcus* have an additional helix located between universal helices 3 and 4 (Figure 5B) not present in other eukaryotic phytoplankton (Marin & Melkonian, 2010), we further validated our findings with analysis of the ITS2 secondary structure which showed Clade E sequences contained this helix, and possessed a unique nucleotide composition compared to other *Ostreococcus* clades (Figure 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 (Figure 5A). Collectively, these analyses validated the informal designation of the novel Clade E as being a fifth species, *O. bengalensis*.

The geographic distribution of *O. bengalensis* amplicons indicates it is predominantly in mesotrophic, potentially moderate salinity, continental shelf regions of the Atlantic Ocean and in the Bay of Bengal (Figure 6A). Our analysis of previously published V4 18S rRNA amplicons indicate 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 Ocean Sampling Day data. These were proposed to represent an *Ostreococcus* variant based on 18S V4 amplicons alone and also termed Clade E (Tragin & Vaulot, 2019), which has also been reported along the western coast of the Philippines (dela Peña et al., 2021). The distribution of *O. bengalensis* in the Bay of Bengal samples contrasted with the broader dominance of *Ostreococcus* Clade OII, which we found had high relative abundances in published V4 18S amplicon data from open ocean sites in the Pacific and North Atlantic (Figure 6B), as well as high qPCR-based abundances in the Sargasso Sea, North Pacific, and western North Atlantic (Bolaños et al., 2020; Limardo et al., 2017; Treusch et al., 2012). A caveat to these 18S amplicon-based analyses is that the relative abundances are subject to changes in any and all other eukaryotic taxa, whether phytoplanktonic or not. Moreover, the relative abundances we observed using MetaPR2 were (oddly) sometimes extraordinarily high in some Ocean Sampling Day samples (e.g., OSD111, surface water off Portugal), especially in contrast to a recent analysis of prasinophytes in TARA V9 data (Bachy et al., 2022). Hence, we take these analyses to serve as a broad-brush view of distributions, as opposed to indication of relative importance (Figure 6).

As shown by broad phytoplankton abundances and distributions, phytoplankton diversity patterns and physical characteristics of the southern stations were different from those closer to freshwater inputs in more northerly regions. The southern region is where the largest influence of Arabian Sea waters can be seen, especially in surface salinity of southern stations (Figure S1B). This influence is well known for the southeastern portion of the bay, and the Arabian Sea itself has traditionally been considered more biologically productive than the Bay of Bengal (Madhupratap et al., 2003; Prasanna Kumar et al., 2002; Thushara et al., 2019). Here, we found that *Prochlorococcus* and *Synechococcus* showed differences from the northern stations largely in the SCM, where higher contributions of high-light adapted *Prochlorococcus* ecotypes and surface associated clades of *Synechococcus* were observed compared to northern SCMs. The composition of *Prochlorococcus* HLII and *Synechococcus*

Clade II oligotypes in these samples also shared more similarities with surface samples than elsewhere. With respect to eukaryotic phytoplankton, relative abundances of Viridiplantae show that surface samples from southern stations have higher contributions of *Ostreococcus* than northern stations. This observed dominance of *Ostreococcus* over other prasinophyte taxa resembles typical characteristics of our SCM samples with two distinctions. First, microdiversity analyses of *Ostreococcus* Clade OII showed major contributions of oligotypes B and C in surface and SCM as a unique feature of the southern stations. Secondly, both *O. bengalensis* and a yet unidentified *Micromonas* related to Clade A/B lineage members, appear in surface waters of station S1. Here, our statistical analyses did not identify a significant environment variable driving this distribution, but these two species had ASVs that fell apart from other clusters of prasinophyte ASVs in the CCA (Figure 7C). Collectively, these observations suggest penetration of Arabian Sea waters into the bay has an important influence on the composition of the picoplanktonic phytoplankton communities, and selection of these groups, such that picoeukaryotes dominate the eukaryotic contributors to primary production.

Eukaryotic phytoplankton communities were highly differentiated across the physicochemical gradients encountered in our study (Figures 2–4). The low-density surface layer in the northern and central stations supported coastal eukaryotic species, adapted to brackish conditions and fluctuating environments and generally groups that have plastids derived from a secondary endosymbiosis event (Bachy et al., 2022). The region of stronger Arabian Sea influences and weaker stratification (Figures 1, 2, and S1), harboured communities with similarities between surface and SCM, even at the oligotype level, and mostly taxa derived from the primary symbiosis event (Figure 4A, Table S1C,E). Cyanobacterial communities, including *Synechococcus* Clade II and *Prochlorococcus* HLII oligotypes, also exhibited shifts, but with more similarity in dominants across surface waters, and seemingly greater community complexity at depth (Figure 3). Despite noted patterns, the observed phytoplankton dynamics was not fully reconciled by statistical analyses using the environmental data acquired. Studies that target seasons outside the summer monsoon will be valuable for further resolving the environmental drivers that shape these communities and potential changes.

# CONCLUSIONS

Comprehensive knowledge of the Bay of Bengal primary producer community is essential given the importance of this ecosystem. The phytoplankton communities observed here showed distinctions from

the gyre ecosystems that reach similar temperatures, that appear to be driven by the unique physicochemical properties of Bay of Bengal waters. Highlights include notable relative abundances of two *Prochlorococcus* ecotypes previously considered rare (HLVI and LLVII). Additionally, *Synechococcus* Clade X occurs at similar relative abundances as Clade II, indicating partially overlapping niches. In the south, where Arabian Sea influences were strongest, *Ostreococcus* Clade OII and an uncultivated species characterized herein, *O. bengalensis*, dominated plastid-16S relative abundances in the surface, whereas *M. commoda* was the more important prasinophyte in other surface samples—alongside several stramenopiles, including uncultivated dictyochophytes. In general, eukaryotic communities exhibited seemingly greater variations in composition than cyanobacteria both vertically and along the transect. These patterns have not been captured in photic zone surface-skin survey sampling, especially in biologically underexplored oceans like the Bay of Bengal. Importantly, our results support a proposal suggesting that picophytoplankton dominate in the Bay of Bengal relative to micro- and nanoplankton size fractions, based on pigment analyses (Sarma et al., 2016, 2020). Moreover, our study supports the projection that picophytoplankton will continue to increase concurrent with decreases in larger eukaryotic phytoplankton (Löscher, 2021). Our insights provide a springboard for more targeted studies and experimentation to understand picophytoplankton niche partitioning in the Bay of Bengal, monsoonal influences, and community transitions.

## EXPERIMENTAL PROCEDURES

### Oceanographic sampling

Samples were collected from the *RV* Roger Revelle between 24 August 2015 and 18 September 2015 at 18 hydrographic stations using a CTD Niskin Rosette as well as 11 samples collected by bucket after loss of the rosette (Table S1A–E). Vertical profiles of temperature, salinity, fluorescence, and nutrients were measured at 0.5 m intervals. Continuous thermosalinograph underway data was consulted to derive parameters not directly measured for samples collected by bucket. Seawater was collected at four to six specific depths ranging from 2 to 100 m depth corresponding to surface, ML, SCM, below SCM, base of the oxycline, and anoxic layer. DNA samples were collected by filtering 500 ml seawater onto 0.2 μm pore size polyethersulfone membrane filters (Supor 200, Pall Gelman). Filters were placed into sterile cryovials, flash-frozen in liquid nitrogen and transferred to −80°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°C until further use. Corresponding nutrient samples were taken as previously described and analysed for concentrations of dissolved inorganic nutrients (nitrate, nitrite, ammonium, phosphate and silicate) using an autoanalyser (Skalar, San series; Netherlands) (Kremling et al., 1983). The detection limits for nitrate, nitrite, ammonium, phosphate, and silicate were ±0.02, ±0.02, ±0.02, ±0.01, and ±0.02 μM, respectively.

### Flow cytometry

Samples were analysed on a BD Influx flow cytometer equipped with a 488 nm 200 mW laser and running on 0.2 μm filtered 1× PBS as sheath fluid. Scatter and fluorescence signals were collected using the BD FACS Software (software v1.0.0.650) with FALS (a proxy for cell size) set as the trigger parameter. Prior to measurement, fluorescent polystyrene beads (0.75 μm yellow-green, Polysciences) were added as internal standards. Each sample was run for 8 min at 25 μl 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 analysed 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 S15 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 μl$^{-1}$ with TE pH 8.0. PCR reactions (50 μl) were performed in 5 μl of 10× buffer, 1 U of HiFi-Taq, 1.6 mM 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 s, 55°C for 30 s, 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 × 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 & Fass, n.d.). Paired-end reads were merged using USEARCH v10.0.240 (Edgar & Flyvbjerg, 2015) when reads had a ≥ 50 bp overlap with maximum 5% mismatch and those with a 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 and 443,556 (237,952 ± 87,596) V1–V2 16S amplicons were recovered per sample, resulting in between 58,416 and 391,913 (164,882 ± 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 with published reference alignments (Choi et al., 2020; Sudek et al., 2015) we noted that, at several key nodes, there were multiple amplicons assigned to LCA. 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 internal transcribed spacer (ITS) 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 & Melkonian, 2010) was used. For recovery of sequences from *Prochlorococcus* ecotypes HLVI and LLVII, the 16S rRNA gene, ITS and partial 23S were 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 s, 60°C for 30 s, 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 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 (Figure S10).

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 (>1500 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. 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 using phylogenetic methods (Choi et al., 2020; Sudek et al., 2015). 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., 2020). Cyanobacteria amplicons were next analysed using a cyanobacteria reference alignment (Sudek et al., 2015) and *Prochlorococcus* was further analysed with reference alignment developed herein after recognizing presence of atypical clades. Reference alignments used herein are available on github.

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 (Katoh & 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. A total of 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 (Katoh & 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).

## 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 & Reid, 2016) to determine taxa driving the differences in microbial compositions at different stations. A CCA of Hellinger-transformed amplicon count data was used to identify the main environmental parameters that may influence the distribution of six taxonomic groups. A variance inflation factor (VIF) analysis was performed using the car package (Fox & Weisberg, 2019) in order to check for multicollinearity among variables: oxygen was discarded, while temperature, depth, nitrite, phosphate, silicate, salinity, and fluorescence were kept in the modelling framework. 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).

## AUTHOR CONTRIBUTIONS
**Jan Strauss:** Formal analysis (equal); writing – original draft (equal). **Chang Jae Choi:** Data curation (equal); formal analysis (equal); validation (equal); visualization (supporting); writing – original draft (supporting). **Jonathan Grone:** Formal analysis (equal); visualization (equal); writing - review and editing (equal). **Fabian Wittmers:** Data curation (supporting); Formal analysis (equal); methodology (equal); writing – review and editing (supporting). **Valeria Jimenez:** writing – review and editing (supporting). **Kriste Makareviciute-Fichtner:** Formal analysis (supporting); writing – review and editing (supporting). **Charles Bachy:** Formal analysis (equal); writing – review and editing (supporting). **Gualtiero Spiro Jaeger:** Investigation (equal). **Camille Poirier:** Formal analysis (supporting); writing – review and editing (supporting). **Charlotte Eckmann:** Formal analysis (supporting); writing – review and editing (supporting). **Rachele Spezzano:** Formal analysis (supporting); methodology (supporting). **Carolin R. Loscher:** Writing – review and editing (supporting). **VVSS Sarma:** Data curation (supporting); formal analysis (supporting); writing – review and editing (supporting). **Amala Mahadevan:** Conceptualization (equal); data curation (supporting); funding acquisition (equal), supervision (equal), writing – review and editing (equal). **Alexandra Z. Worden:** Conceptualization (equal); data curation (supporting); formal analysis (equal); funding acquisition (equal); methodology (equal); resources (equal); supervision (equal); writing – original draft (equal); writing – review and editing (equal).

## CONFLICT OF INTEREST STATEMENT
The authors declare no conflicts of interest.

## 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 PRJNA905425. 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.

## ORCID

*Jan Strauss* 🔘 https://orcid.org/0000-0002-6208-791X
*Jonathan Grone* 🔘 https://orcid.org/0000-0001-5074-265X
*Amala Mahadevan* 🔘 https://orcid.org/0000-0002-7522-4100
*Alexandra Z. Worden* 🔘 https://orcid.org/0000-0002-9888-9324

## REFERENCES

Ahlgren, N.A. & 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. *Frontiers in Microbiology*, 3, 213.

Andersen, K.S., Kirkegaard, R.H., Karst, S.M. & Albertsen, M. (2018) ampvis2: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv*, 299, 537.

Andrews, S. (2012) FastQC: a quality control tool for high throughput sequence data, Babraham Institute.

Angelova, A.G., Ellis, G.A., Wijesekera, H.W. & Vora, G.J. (2019) Microbial composition and variability of natural marine planktonic and biofouling communities from the Bay of Bengal. *Frontiers in Microbiology*, 10, 2738.

Bachy, C., Wittmers, F., Muschiol, J., Hamilton, M., Henrissat, B. & Worden, A.Z. (2022) The land-sea connection: insights into the plant lineage from a green algal perspective. *Annual Review of Plant Biology*, 73, 585–616.

Bennke, C.M., Pollehne, F., Müller, A., Hansen, R., Kreikemeyer, B. & Labrenz, M. (2018) The distribution of phytoplankton in the Baltic Sea assessed by a prokaryotic 16S rRNA gene primer system. *Journal of Plankton Research*, 40, 244–254.

Berube, P.M., Rasmussen, A., Braakman, R., Stepanauskas, R. & 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. & Chisholm, S.W. (2015) Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology*, 13, 13–27.

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. *The ISME Journal*, 14, 1663–1674.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A. & Holmes, S.P. (2016) DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581–583.

Capella-Gutiérrez, S., Silla-Martínez, J.M. & 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. & Knight, R. (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26, 266–267.

Chen, Z., Gu, T., Wang, X., Wu, X. & Sun, J. (2022) Oxygen gradients shape the unique structure of picoeukaryotic communities in the Bay of Bengal. *Science of the Total Environment*, 814, 152862.

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. *Frontiers in Microbiology*, 11, 542372.

Cordero, O.X. & Polz, M.F. (2014) Explaining microbial genomic diversity in light of evolutionary ecology. *Nature Reviews Microbiology*, 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.

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. *Molecular Ecology Resources*, 15, 1435–1445.

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. *The Philippine Journal of Science*, 150, 27–42.

Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C.L., Zehr, J.P. & Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote Ostreococcus. *The ISME Journal*, 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 $CO_2$-fixation across the oligotrophic South Pacific Subtropical Ocean. *The ISME Journal*, 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. *The ISME Journal*, 9, 1076–1092.

Edgar, R.C. & 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. et al. (2013) Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. *Methods in Ecology and Evolution*, 4, 1111–1119.

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. *Proceedings of the National Academy of Sciences of the United States of America*, 110, 9824–9829.

Fox, J. & Weisberg, S. (2019) *An R Companion to Applied Regression*, 3rd edition. Sage, Thousand Oaks CA. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fuller, N.J., Campbell, C., Allen, D.J., Pitt, F.D., Zwirglmaier, 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. *Aquatic Microbial Ecology*, 43, 79–93.

Gastineau, R., Hamedi, C., Baba Hamed, M.B., Abi-Ayad, S.-M.E.-A., Bąk, 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. *Ecological Indicators*, 121, 107104.

Gloor, G.B. & Reid, G. (2016) Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Canadian Journal of Microbiology*, 62, 692–703.

Gomes, H.R., Goes, J.I.G. & Saino, T. (2000) Influence of physical processes and freshwater discharge on the seasonality of phytoplankton regime in the bay of Bengal. *Continental Shelf Research*, 20, 313–330.

Gong, W., Hall, N., Paerl, H. & Marchetti, A. (2020) Phytoplankton composition in a eutrophic estuary: comparison of multiple taxonomic approaches and influence of environmental factors. *Environmental Microbiology*, 22, 4718–4731.

Green, B.R. (2011) Chloroplast genomes of photosynthetic eukaryotes. *The Plant Journal*, 66, 34–44.

Gregg, W.W. & Rousseaux, C.S. (2019) Global ocean primary production trends in the modern ocean color satellite record (1998–2015). *Environmental Research Letters*, 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. *Communications Biology*, 5, 983.

Huang, S., Wilhelm, S.W., Harvey, H.R., Taylor, K., Jiao, N. & Chen, F. (2012) Novel lineages of Prochlorococcus and Synechococcus in the global oceans. *The ISME Journal*, 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.

Jiao, N., Luo, T., Zhang, R., Yan, W., Lin, Y., Johnson, Z.I. et al. (2014) Presence of Prochlorococcus in the aphotic waters of the western Pacific Ocean. *Biogeosciences*, 11, 2391–2400.

Johnson, Z.I., Zinser, E.R., Coe, A., McNulty, N.P., Woodward, E.M.S. & Chisholm, S.W. (2006) Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. *Science*, 311, 1737–1740.

Joshi, N.A. & 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. & Vinayachandran, P.N. (2018) Response of phytoplankton to heavy cloud cover and turbidity in the northern Bay of Bengal. *Science Reports*, 8, 11282.

Jyothibabu, R., Vinayachandran, P.N., Madhu, N.V., Robin, R.S., Karnan, C., Jagadeesan, L. et al. (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. & Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14, 587–589.

Kashtan, N., Roggensack, S.E., Berta-Thompson, J.W., Grinberg, M., Stepanauskas, R. & Chisholm, S.W. (2017) Fundamental differences in diversity and genomic population structure between Atlantic and Pacific Prochlorococcus. *The ISME Journal*, 11, 1997–2011.

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. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772–780.

Kent, A.G., Baer, S.E., Mouginot, C., Huang, J.S., Larkin, A.A., Lomas, M.W. et al. (2019) Parallel phylogeography of Prochlorococcus and Synechococcus. *The ISME Journal*, 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. & Grasshoff, K. (1983) In: Grasshoff, K., Ehrhardt, M. & Kremling, K. (Eds.) *Methods of Seawater Analysis*, 2nd edition. New York, NY: Verlag Chemie Weinheim.

Krishna, M.S., Prasad, M.H.K., Rao, D.B., Viswanadham, R., Sarma, V.V.S.S. & Reddy, N.P.C. (2016) Export of dissolved inorganic nutrients to the northern Indian Ocean from the Indian monsoonal rivers during discharge period. *Geochimica et Cosmochimica 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. *The ISME Journal*, 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. *Limnology and Oceanography*, 65, S220–S232.

Larkin, A.A., Hagstrom, G.I., Brock, M.L., Garcia, N.S. & Martiny, A.C. (2022) Basin-scale biogeography of Prochlorococcus and SAR11 ecotype replication. *The ISME Journal*, 17, 185–194.

Li, Q., Edwards, K.F., Schvarcz, C.R., Selph, K.E. & Steward, G.F. (2021) Plasticity in the grazing ecophysiology of Florenciella (Dichtyochophyceae), a mixotrophic nanoflagellate that consumes Prochlorococcus and other bacteria. *Limnology and Oceanography*, 66, 47–60.

Li, S., Bronner, G., Lepère, C., Kong, F. & Shi, X. (2017) Temporal and spatial variations in the composition of freshwater photosynthetic picoeukaryotes revealed by MiSeq sequencing from flow cytometry sorted samples. *Environmental Microbiology*, 19, 2286–2300.

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. *Environmental Microbiology*, 19, 3219–3234.

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 Research Part II: Topical Studies in Oceanography*, 50, 881–896.

Mahadevan, A. (2016) The impact of submesoscale physics on primary productivity of plankton. *Annual Review of Marine Science*, 8, 161–184.

Mahadevan, A., Paluszkiewicz, T., Ravichandran, M., Sengupta, D. & 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. et al. (2010) Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific oceans. *The ISME Journal*, 4, 1252–1264.

Marin, B. & 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.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, 17, 10–12.

Martiny, A.C., Tai, A.P.K., Veneziano, D., Primeau, F. & Chisholm, S.W. (2009) Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environmental Microbiology*, 11, 823–832.

McCreary, J.P., Yu, Z., Hood, R.R., Vinaychandran, P.N., Furue, R., Ishida, A. et al. (2013) Dynamics of the Indian-Ocean oxygen minimum zones. *Progress in Oceanography*, 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. *The ISME Journal*, 6, 610–618.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A. et al. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*, 37, 1530–1534.

Monier, A., Worden, A.Z. & Richards, T.A. (2016) Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environmental Microbiology Reports*, 8, 461–469.

Naik, R.K., Hegde, S. & Anil, A.C. (2011) Dinoflagellate community structure from the stratified environment of the Bay of Bengal, with special emphasis on harmful algal bloom species. *Environmental Monitoring and Assessment*, 182, 15–30.

Needham, D.M. & Fuhrman, J.A. (2016) Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology*, 1, 16005.

Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R. et al. (2022) Vegan: community ecology package. The R Project for Statistical Computing.

Paul, J.T., Ramaiah, N. & Sardessai, S. (2008) Nutrient regimes and their effect on distribution of phytoplankton in the Bay of Bengal. *Marine Environmental Research*, 66, 337–344.

Paulmier, A. & Ruiz-Pino, D. (2009) Oxygen minimum zones (OMZs) in the modern ocean. *Progress in Oceanography*, 80, 113–128.

Poretsky, R.S., Sun, S., Mou, X. & Moran, M.A. (2010) Transporter genes expressed by coastal bacterioplankton in response to dissolved organic carbon. *Environmental Microbiology*, 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? *Geophysical Research Letters*, 29, 2235.

Price, M.N., Dehal, P.S. & 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). *Frontiers in Microbiology*, 10, 1501.

R Core Team. (2022) R: a language and environment for statistical computing.

Ramaiah, N., Fernandes, V., Bhaskar, J., Jyothibabu, R., Gauns, M. & Jayraj, E.A. (2010) Seasonal variability in biological carbon biomass standing stocks and production in the surface layers of the Bay of Bengal. *Indian Journal of Marine Sciences*, 39, 369–379.

Rocap, G., Distel, D.L., Waterbury, J.B. & Chisholm, S.W. (2002) Resolution of Prochlorococcus and Synechococcus ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. *Applied and Environmental Microbiology*, 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. & Moreau, H. (2005) Ecotype diversity in the marine picoeukaryote Ostreococcus (Chlorophyta, Prasinophyceae). *Environmental Microbiology*, 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. *Systematic Biology*, 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. *Geophysical Research Letters*, 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. & Sharifuzzaman, S.M. (2020) A biophysicochemical perspective of the Bay of Bengal. *Journal of the Marine Biological Association of the United Kingdom*, 100, 517–528.

Sarma, V.V.S.S., Rajula, G.R., Durgadevi, D.S.L., Kumar, G.S. & Loganathan, J. (2020) Influence of eddies on phytoplankton composition in the Bay of Bengal. *Continental Shelf Research*, 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. *Journal of Oceanography*, 69, 229–243.

Seemann, T. (2018) BAsic Rapid Ribosomal RNA Predictor (barrnap).

Shi, Y., Tyson, G.W., Eppley, J.M. & DeLong, E.F. (2011) Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *The ISME Journal*, 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 Research Part II: Topical Studies in Oceanography*, 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. *The ISME Journal*, 10, 333–345.

Spiro Jaeger, G. & Mahadevan, A. (2018) Submesoscale-selective compensation of fronts in a salinity-stratified ocean. *Science Advances*, 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. & 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 Research Part II: Topical Studies in Oceanography*, 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. *Current Science*, 64, 928–930.

Sudek, S., Everroad, R.C., Gehman, A.-L.M., Smith, J.M., Poirier, C.L., Chavez, F.P. et al. (2015) Cyanobacterial distributions along a physico-chemical gradient in the northeastern Pacific Ocean. *Environmental Microbiology*, 17, 3692–3707.

Thompson, A.W. & Kouba, K. (2019) Differential activity of coexisting Prochlorococcus ecotypes. *Frontiers in Marine Science*, 6, 701.

Thompson, A.W., van den Engh, G., Ahlgren, N.A., Kouba, K., Ward, S., Wilson, S.T. et al. (2018) Dynamics of Prochlorococcus diversity and photoacclimation during short-term shifts in water column stratification at station ALOHA. *Frontiers in Marine Science*, 5, 488.

Thushara, V., Vinayachandran, P.N.M., Matthews, A.J., Webber, B.G.M. & Queste, B.Y. (2019) Vertical distribution of chlorophyll in dynamically distinct regions of the southern Bay of Bengal. *Biogeosciences*, 16, 1447–1468.

Tragin, M. & Vaulot, D. (2019) Novel diversity within marine Mamiellophyceae (Chlorophyta) unveiled by metabarcoding. *Science Reports*, 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. *The ISME Journal*, 6, 481–492.

van Baren, M.J., Bachy, C., Reistetter, E.N., Purvine, S.O., Grimwood, J., Sudek, S. et al. (2016) Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. *BMC Genomics*, 17, 267.

Vaulot, D., Sim, C.W.H., Ong, D., Teo, B., Biwer, C., Jamy, M. et al. (2022) metaPR2: a database of eukaryotic 18S rRNA metabarcodes with an emphasis on protists. *Molecular Ecology Resources*, 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. *The ISME Journal*, 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. *Molecular Biology and Evolution*, 37, 599–603.

Wei, Y., Huang, D., Zhang, G., Zhao, Y. & Sun, J. (2020) Biogeographic variations of picophytoplankton in three contrasting seas: the Bay of Bengal, South China Sea and Western Pacific Ocean. *Aquatic Microbial Ecology*, 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. *Bulletin of the American Meteorological Society*, 97, 1859–1884.

Worden, A.Z. (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquatic Microbial Ecology*, 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. *Current Biology*, 22, R675–R677.

Xia, X., Cheung, S., Endo, H., Suzuki, K. & Liu, H. (2019) Latitudinal and vertical variation of Synechococcus assemblage composition along 170° W transect from the South Pacific to the Arctic Ocean. *Microbial Ecology*, 77, 333–342.

Yu, G., Smith, D.K., Zhu, H., Guan, Y. & Lam, T.T.-Y. (2017) Ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, 8, 28–36.

Zinser, E.R., Johnson, Z.I., Coe, A., Karaca, E., Veneziano, D. & Chisholm, S.W. (2007) Influence of light and temperature on Prochlorococcus ecotype distributions in the Atlantic Ocean. *Limnology and Oceanography*, 52, 2205–2220.

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.