Title: Bacterial biogeography of the Indian Ocean

Authors: Melissa L. Brock[1], Alyse A. Larkin[2], Eric J. Raes[3], and Adam C. Martiny[1,2*]

Institutional Affiliations:

1) Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, CA, USA

2) Department of Earth System Science, University of California Irvine, Irvine, CA, USA

3) Flourishing Oceans, Minderoo Foundation, Broadway Nedlands, Western Australia, Australia

ORCID IDs & Institutional Emails:

Melissa L. Brock: 0000-0003-4032-9241; mlbrock@uci.edu

Alyse A. Larkin: 0000-0003-4466-0791; larkinsa@uci.edu

Eric J. Raes: 0000-0002-4131-9312; eraes@minderoo.org

Adam C. Martiny: 0000-0003-2829-4314; amartiny@uci.edu

***Corresponding Author:** Department of Ecology and Evolutionary Biology and Department of Earth System Science, University of California Irvine, Irvine, CA, USA; amartiny@uci.edu

Running Head: Bacterial biogeography of the Indian Ocean

Keywords: surface ocean, microbial diversity, taxonomic composition, marker-gene biogeography, ecological regions, seascape

Abstract

Historically, our understanding of bacterial ecology in the Indian Ocean has been limited to regional studies that place emphasis on community structure and function within oxygen minimum zones. Thus, bacterial community dynamics across the wider Indian Ocean are largely undescribed. As part of Bio-GO-SHIP, we sequenced the 16S rRNA gene from 465 samples collected on sections I07N and I09N. We found that (i) there were 23 distinct bioregions within the Indian Ocean, (ii) the southeastern gyre had the largest gradient in bacterial alpha-diversity, (iii) the Indian Ocean surface microbiome was primarily composed of a core set of taxa, and (iv) bioregions were characterized by transitions in physical and geochemical conditions. Overall, we showed that bacterial community structure spatially delineated the surface Indian Ocean and that these microbially-defined regions were reflective of subtle ocean physical and geochemical gradients. Therefore, incorporating metrics of in-situ microbial communities into marine ecological regions traditionally defined by remote sensing will improve our ability to delineate warm, oligotrophic regions.

Introduction

The Indian Ocean represents an important region for understanding marine ecology because of its unique physical dynamics, global biogeochemical importance, and historic under-sampling. The Indian Ocean is bounded to the north by the Eurasian land mass and experiences seasonal monsoon dynamics, which have important implications for the geochemistry and ecology of the basin (Hood et al. 2007; Vinayachandran et al. 2021). Monsoon forcing leads to the formation of the world's thickest oxygen minimum zone in the Arabian Sea and thus substantial N-loss (McCreary et al. 2013; Al Azhar et al. 2017). The Indian Ocean also contributes ~15% to global ocean primary productivity (Behrenfeld and Falkowski 1997). Lastly, the region is warming faster than any other ocean basin (Roxy et al. 2014). However, we currently have a limited understanding of microbial biodiversity at large spatial scales across the Indian Ocean.

Our current understanding of bacterial ecology within the Indian Ocean is primarily restricted to regional studies. Emphasis has been placed on microorganisms linked to N-cycling near oxygen minimum zones (Jayakumar et al. 2012). Few studies have examined the bacterial community outside these zones, and they are limited in scale (Jeffries et al. 2015; Zheng et al. 2016; Raes et al. 2018b; Hörstmann et al. 2021). These studies suggest that there is biogeographic partitioning between communities in surface waters of the southwestern Indian Ocean, central Indian Ocean, and Bay of Bengal (Jeffries et al. 2015). There was also spatial heterogeneity within the central (Zheng et al. 2016), southeastern (Raes et al. 2018b), and southwestern Indian Ocean (Hörstmann et al. 2021), indicating fine-scale spatial separation of bacterial communities. The environmental drivers associated with diversity shifts varied between geographic regions. Within the southwestern region, alpha-diversity increased with temperature

(Hörstmann et al. 2021), while biodiversity in the southeastern region was more closely linked to productivity (Raes et al. 2018b). This may suggest that bacterial communities are influenced by different environmental factors within Indian Ocean regions. Additionally, there are distinct microbial "fingerprints" at frontal regions in the southwestern Indian Ocean (Hörstmann et al. 2021), indicating that transitions between water masses represent either unique microbial habitats or a physical barrier to dispersal. Together, these observations suggest regionally distinct communities in the Indian Ocean, but the large-scale microbial biogeography is poorly understood.

Identifying ecological regions in the Indian Ocean has been challenging, because most of the basin is warm and oligotrophic. The pelagic Indian Ocean has previously been divided into two broad biogeochemical provinces (Longhurst 2010), four ecological marine units (Sayre et al. 2017), and ~10 unique seascapes (Kavanaugh et al. 2014) (Figure 1 and Supplemental Figure 1). These regions were primarily delineated using common geochemical and physical metrics derived from remote-sensing and/or in-situ hydrographic measurements, such as temperature, nutrient concentrations, and chlorophyll *a* concentration. However, such metrics have low variability across the surface Indian Ocean, and nutrients are often below the standard detection limit, resulting in low discriminatory resolution. In contrast, microbial communities are highly sensitive to local environments and can be used as "biosensors" for identifying ecological provinces not easily detected with chemical or physical measurements (Larkin et al. 2020). Microbial communities are impacted by a wide variety of biogeochemical factors or the presence of physical barriers (e.g., islands, currents, fronts, and eddies) (Sebastián et al. 2021). A shift in community structure will likely reflect environmental changes and associated ecological functions. For example, when using geochemical measurements, the eastern Indian Ocean was

partitioned into three distinct biomes (Garcia et al. 2018), but when this same region was partitioned using variation in *Prochlorococcus* haplotypes, eight new, distinct biogeochemical regimes were identified (Larkin et al. 2020). Therefore, changes in microbial biogeography can be used to delineate marine ecosystems (i.e., bioregions). Contrary to marine ecological regions, bioregions incorporate the inherent biodiversity of the region into their delineation, which may provide a more nuanced understanding of the Indian Ocean seascape.

Using a high spatial resolution genomics dataset from the Indian Ocean, we ask the following questions: i) Are bacterial communities structured into distinct spatial regions (i.e., bioregions)? ii) Across these bioregions, where do the biggest gradients in bacterial alpha-diversity occur and what environmental factors correspond with changes in alpha-diversity? iii) What lineages are endemic to these regions and what are the major ubiquitous lineages across the Indian Ocean? and iv) How do geochemical and physical dynamics vary across the bioregions? We addressed these questions using intensive DNA sampling from two large meridional Bio-GO-SHIP sections to the eastern (I09N) and western (I07N) Indian Ocean.

Methods

Field sampling and environmental data

Microbial DNA samples were collected on GO-SHIP cruise I09N (n = 215) which ran from Freemont, Australia to Phuket, Thailand in 2016 (March 22 – April 24) and on GO-SHIP cruise I07N (n = 250) which ran from Durban, South Africa to Mormugao, India in 2018 (April 23 – June 6) (Figure 1a). Between 1 – 10 L of surface water were collected every 4 – 6 hours from the ship's circulating seawater system at 7m depth (n = 414) or via Niskin rosette at 3m depth (n = 51) (Supplemental Table 1). Samples were collected using 0.22 μm Sterivex filters

and were preserved with 1620 μL of lysis buffer (23.4 mg mL^{-1} NaCl, 257 mg mL^{-1} sucrose, 50 mmol L^{-1} Tris-HCl, 20 mmol L^{-1} EDTA). Samples were frozen at -20°C until further processing.

Underway temperature and salinity were measured using a mounted near-surface thermosalinograph. At GO-SHIP stations, samples for inorganic nutrients were collected at approximately every latitudinal degree (~11 km) for the entire water column using a Niskin rosette. Nitrate and phosphate concentrations were measured following GO-SHIP protocols (https://cchdo.ucsd.edu/). The detection limits for nitrate and phosphate are typically 0.02 μmol L^{-1} when using standard autoanalytical techniques. For underway sampling points between GO-SHIP stations, nutrient concentrations were linearly interpolated from the nearest GO-SHIP stations. Additional biogeochemical data for these GO-SHIP cruises can be found at (https://cchdo.ucsd.edu/cruise/33RR20160321; https://cchdo.ucsd.edu/cruise/33RO20180423).

Nutricline depth was defined as the depth at which nitrate was $\geq$1 μmol L^{-1} and was used as a proxy for nutrient supply to the mixed layer (Cermeño et al. 2008). Nitrate profiles were interpolated at 1 m resolution at each GO-SHIP station. At underway sampling points between GO-SHIP stations, the nutricline depth was interpolated from the nearest GO-SHIP stations. For underway samples collected before the first GO-SHIP station, World Ocean Atlas climatological nitrate depth profiles were used to estimate the nutricline depth (Garcia et al. 2018).

Samples for particulate organic matter (POM) were collected and measured as described by Garcia et al. (2018). Briefly, 4 – 8 L of seawater was collected from the ship's circulating seawater system and was filtered using a 30 μm nylon mesh to remove large particles. Samples were then collected on a 25 mm pre-combusted (500 °C for 5 h) GF/F filter (nominal pore size = 0.7 μm), stored in pre-combusted aluminum packets, and frozen at -80°C. Concentrations of POC/PON were determined using a CN FlashEA 112 Elemental Analyzer, and concentrations of

130 POP were determined according to a modified ash-hydrolysis protocol. The detection ranges for
131 POC and PON were 0.43 – 43.13 μM and 0.037 – 7.39 μM, respectively. The detection limit for
132 POP was 0.1 nmol L^{-1}. POM values reported here are the mean of the replicates. Ratios of
133 POC/PON, POC/POP, and PON/POP were calculated from the mean concentrations of POC,
134 PON, and POP. POM data is publicly available on BCO-DMO (https://www.bco-dmo.org/).
135 An iron (Fe)-stress parameter, ϕ_{sat}, was estimated from MODIS-Aqua satellite
136 fluorescence data (Behrenfeld et al. 2009). ϕ_{sat} across the Indian Ocean was calculated from a
137 data set of 9 km resolution global distributions as an average of climatological means taken from
138 2003 to 2015 (Larkin et al. 2020). ϕ_{sat} for each sampling point was linearly interpolated from
139 the nearest two ϕ_{sat} data points.
140 Daily mean sea surface height relative to the geoid was extracted from 0.25-degree
141 gridded data obtained from Copernicus Marine Environment Monitoring Service. Sea surface
142 height for each sampling point was interpolated from the nearest two gridded data points. To
143 visualize surface current patterns, daily mean horizontal velocity (meridional and zonal
144 component at 0 m depth) was also extracted from 0.25-degree gridded data obtained from
145 Copernicus Marine Environment Monitoring Service. Sea surface height anomalies, geostrophic
146 current direction, nutricline depth, and temperature were used to identify cold- and warm-core
147 eddies. Cold-core eddies in the northern/southern hemisphere were defined by depressed sea
148 height, anticlockwise/clockwise rotation, shoaling nutricline, and decreased temperatures. Warm-
149 core eddies in the northern/southern hemisphere were defined by increased sea height,
150 clockwise/anticlockwise rotation, depressed nutricline, and increased temperatures.
151 Transitions between water masses along the transects were identified by determining
152 where changes in density occurred. Specifically, temperature and salinity data were fitted with a

polynomial regression using the loess function (span = 0.1) in R (R Core Team 2019). The first derivative of the fitted data was smoothed with a polynomial regression fit. The local minima and maxima of the smoothed first derivative were then used to define where the biggest changes in temperature and salinity occurred. Locations where a local minimum/maximum for both temperature and salinity occurred indicated a transition between water masses. Currents and fronts were identified by daily mean horizontal velocity direction and changes in density.

Dynamic seascape analysis

The number and distribution of seascape classes (Kavanaugh et al. 2014) across the Indian Ocean were extracted from monthly composite data obtained from NOAA CoastWatch. Data were obtained for April, a representative month from the intermonsoon season, for 2016 and 2018. The relative abundance of each seascape class was calculated, and a 2% relative abundance threshold was used to define presence of a seascape class. The geographic distributions of seascape classes were plotted to compare spatial patterns between the two years.

DNA extraction, 16S rRNA amplification, and sequencing

Microbial DNA was extracted following methods previously described (Larkin et al. 2020). Briefly, Sterivex filters were incubated with lysozyme (50 mg mL^{-1} final concentration) at 37°C for 30 minutes. Proteinase K (1 mg mL^{-1}) and 10% SDS buffer were added to the Sterivex filters, and samples were incubated at 55°C overnight. Sodium acetate (245 mg mL^{-1}, pH 5.2) and ice-cold isopropanol (100%) were used to precipitate DNA. Samples were pelleted via centrifuge at 15,000×*g* at 4°C for 30 minutes and resuspended in TE buffer (10 mmol L^{-1} Tris-HCl, 1 mmol L^{-1} EDTA) at 37°C for 1 hour. DNA was purified and concentrated (Zymo

genomic DNA Clean and Concentrator kit; Zymo Research Corp., Irvine, CA). DNA concentration was checked using a Qubit dsDNA HS Assay and a Qubit fluorometer (ThermoFisher, Waltham, MA). Samples were diluted to 2 ng μL^{-1}.

The V4-V5 region of the 16S rRNA bacterial gene was amplified using the 515F-C and 926R primer set (Needham and Fuhrman 2016). A total of 4 μL of 2 ng μL^{-1} DNA was added to 20 μL reactions (0.4 μM of each primer, 1X AccuStart II PCR Supermix; final concentration). Amplification occurred in a 2-step process. The first polymerase chain reaction (PCR) occurred as follows: 94°C for 3 min and 26 cycles of 94°C for 30 s, 55°C for 30 s, 68°C for 40 s. Free primers and primer dimers were removed using a magnetic bead cleanup (10 μL Milli-Q, 10 μL PCR product, 20 μL Sera-mag SpeedBeads). A total of 4 μL of bead-cleaned product was added to 20 μL reactions (0.3 μM each i5 and i7 Nextera v2 indices, 1X AccuStart II PCR Supermix; final concentration). Barcodes were annealed to the bead-cleaned products during the second PCR: 12 cycles of 94°C for 30 s, 55°C for 30 s, 68°C for 40 s, and a final extension of 68°C for 10 min. Final PCR products were visualized with a 1% agarose gel and pooled. Unincorporated barcodes were removed from the pooled library with a final magnetic bead cleanup (60 μL pooled product, 60 μL Sera-mag SpeedBeads). Quality of the library was checked using a Bioanalyzer (Agilent, Santa Clara, CA). Amplicons were pair-end sequenced (2×300 bp) with the Illumina MiSeq platform (Illumina, San Diego, CA). Sequence files are available at NCBI Sequence Read Archive under BioProject ID PRJNA656268, and BioSample accession numbers are reported for each sample in Supplemental Table 2.

Quality filtering and amplicon clustering

198 Initial library sizes ranged from 19,619 to 181,435 sequences (Mean ± SD = 55,796 ±
199 17,552) (Supplemental Table 2). Primers were removed from reads using cutadapt (Martin 2011)
200 in QIIME2 (Bolyen et al. 2019). Forward and reverse reads were quality filtered with fastq-mcf
201 (Aronesty 2013). A window-size of 10 was used to calculate mean quality score. Reads were
202 truncated when the mean quality score was less than 20. After trimming, reads that were shorter
203 than the minimum length threshold (150 bp) and reads that contained N-calls were removed.
204 Forward and reverse reads were merged based on a minimum overlap threshold (10 bp),
205 minimum merge length threshold (350 bp), and number of maximum differences allowed in the
206 overlapping region (5 bp) using usearch (Edgar 2010). Final trimming, quality filtering,
207 clustering of amplicons, and removal of chimeras was performed using DADA2 (Callahan et al.
208 2016) in QIIME2 (Bolyen et al. 2019). The merged reads were trimmed to a length threshold
209 (304 bp) to maintain alignment. Reads that matched to the PhiX genome or that contained more
210 than 3 expected errors were removed. The error model was trained using a minimum of 800,000
211 reads. Samples were then dereplicated, reads were clustered into amplicon sequence variants
212 (ASVs), and chimeric ASVs were removed using a consensus procedure. ASVs were clustered
213 into de novo 99% operational taxonomic units (OTUs) using the VSEARCH plugin (Rognes et
214 al. 2016) in QIIME2 (Bolyen et al. 2019) to minimize the effect that rare ASVs resulting from
215 differences in sequencing runs may have on diversity metrics. After quality filtering, final library
216 sizes ranged from 10,299 to 101,783 sequences (Mean ± SD = 26,893 ± 11,533) (Supplemental
217 Table 2).
218
219 *Taxonomic assignments*

Taxonomy was assigned to 99% OTUs using the SILVA138 reference database (Quast et al. 2012) with RDP's Naïve Bayesian classifier implemented in the R package "dada2" (Callahan et al. 2016). Taxonomic assignments were based on a minimum bootstrap confidence of 80 out of 100 bootstraps. The taxonomic composition of the samples was dominated by bacterial OTUs (Mean ± SD = 96.21% ± 1.88%) (Supplemental Table 2), while eukaryotic and archaeal OTUs were low in most samples (Eukaryotes: Mean ± SD = 0.004% ± 0.010%; Archaea: Mean ± SD = 3.78% ± 1.88%) (Supplemental Table 2). OTUs that matched to eukaryotes and archaea were removed from all subsequent diversity analyses. All bacterial taxonomic assignments for each sample are reported at the genus level in Supplemental Table 3.

Diversity analyses and bioregion delineation

To ensure that diversity patterns were not an artifact of sequencing depth, samples were rarefied to a depth of 10,000 sequences. Alpha-diversity was calculated using the Shannon Index. Richness was calculated as the number of OTUs in a sample, and evenness was calculated using Pielou's Index. Prior to calculating these metrics, singletons were removed to discard a small number of very low-abundance sequences that may represent spurious sequences. Removing singletons did not have a large impact on overall diversity metrics. Removal of singletons caused a small reduction in the Shannon index (Mean ± SD = 0.011 ± 0.007) as well as richness (Mean ± SD = 16 ± 10) and led to a small increase in Pielou's Index (Mean ± SD = 0.006 ± 0.005). Correlations of alpha-diversity with ubiquitous and cosmopolitan genera were calculated using Pearson's correlation coefficient, and p-values were adjusted for multiple comparisons using the Benjamini-Hochberg correction method. Beta-diversity was calculated using the Bray-Curtis dissimilarity index. Distinct biological regions (i.e., bioregions) across the Indian Ocean were

defined by partitioning samples into groups of similar bacterial communities through agglomerative hierarchical clustering using the unweighted pair group method with arithmetic mean (UPGMA) on the Bray-Curtis dissimilarity matrix of the GO-SHIP samples. A high degree of longitudinal separation was observed between the eastern and western transects (Supplemental Figure 2a). Additionally, the two transects exhibited different distributions of Bray-Curtis dissimilarities, where the eastern transect was right-skewed while the western transect was symmetric (Supplemental Figure 3). Therefore, the two transects were analyzed separately. Agglomerative hierarchical clustering using UPGMA was performed on the Bray-Curtis dissimilarity matrices of GO-SHIP I09N and I07N samples. The cophenetic correlation coefficient is the correlation between the Bray-Curtis dissimilarity matrix and the cophenetic matrix, which contains distances between clusters. The cophenetic correlation coefficient ranges from 0 – 1 with values close to 1 indicating that the dendrogram preserved the pairwise distances of the original data points. It was therefore calculated to confirm that the dendrograms were good visual representations of the dissimilarity matrices (I09N = 0.865, I07N = 0.744). The agglomerative coefficient describes the strength of the clustering patterns and ranges from 0 – 1 with values close to 1 indicating a balanced clustering structure. The agglomerative coefficient was therefore calculated to confirm that strong clustering patterns existed within the data (I09N = 0.874, I07N = 0.862). To define clusters, the dendrograms were cut using dissimilarity thresholds of 0.20 and 0.17 for I09N and I07N, respectively (Supplemental Figure 2b and 2c). Dissimilarity thresholds for each section were selected by calculating the -1 standard deviation from the upper half of the dissimilarity matrix. These dissimilarity thresholds resulted in 54 clusters along the I09N transect and 44 clusters along the I07N transect. Clusters with fewer than 5 samples were removed because changes in community structure at this spatial resolution likely

resulted from transient, submesoscale processes and not broader environmental gradients. After removing clusters with fewer than 5 samples from the analysis, a total of 26 clusters remained (I09N = 14 clusters, I07N = 12 clusters). Cluster stability was determined through bootstrapping (1000 iterations using "clusterboot" in the "fpc" package, R). Bootstrapped clusterwise means ranged from 0.524 – 0.960. Clusters with clusterwise means less than 0.60 were considered unstable and were discarded, resulting in a total of 23 remaining clusters that were designated as bioregions. Samples in the 23 bioregions were back-projected to their spatial coordinates to examine their geographic trends. Using higher dissimilarity thresholds (dissimilarity = 0.225, 0.250, 0.275, and 0.300) resulted in cluster instability across large geographic regions or in poor separation of geographic regions (Supplemental Table 4 and Supplemental Figure 4), indicating that higher dissimilarity thresholds were not suitable for partitioning this dataset. The 23 bioregions were plotted on maps along with Longhurst provinces (Longhurst 2010) and ecological marine units (Sayre et al. 2017) to compare the bioregions with previously defined marine ecological regions in the Indian Ocean.

Taxonomic patterns

Differential abundance of genera across the eastern Indian Ocean biomes (i.e., southern gyre, equatorial region, and Bay of Bengal) and the western Indian Ocean biomes (i.e., southern gyre, equatorial region, and Arabian Sea) was performed on taxa count tables (function "DESeq" in the "DESeq2" package, R) (Love et al. 2014). Additionally, heatmaps showing changes in genera abundance (>20 total counts) according to bioregion were generated using "plot_heatmap" in the "phyloseq" package (trans = log_trans(4)) (McMurdie and Holmes 2013). Heatmaps were constructed using count data so that they could be directly compared to the

289 results from differential abundance analysis. Ubiquitous, cosmopolitan, biome-associated,
290 region-associated, and endemic taxa were also identified. Ubiquitous taxa were defined as genera
291 that were found in all samples across the Indian Ocean. Microdiversity of the three most
292 abundant, ubiquitous taxa was examined through stacked bar plots of 99% OTU relative
293 abundance for each taxon and through heatmaps of 99% OTU counts according to bioregion.
294 Cosmopolitan taxa were defined as genera that were found in the majority of samples (≥75% but
295 <100% of samples). Region-associated taxa were defined as genera that were found in ≥75% of
296 samples in one biological region (i.e., bioregion) and ≤25% of samples in all other bioregions.
297 Lastly, endemic taxa were defined as genera that exist in ≥75% of samples within one bioregion
298 and 0% of samples in all other bioregions.
299
300 *Flow cytometry and primary production*
301 Samples for flow cytometry and primary production were collected at 29 GO-SHIP
302 stations along the I09N transect (Baer et al. 2019). For flow cytometry analysis, samples were
303 collected directly from Niskin bottles and preserved with a 0.2 μm-filtered 10%
304 paraformaldehyde solution (final concentration of 0.5% (v/v)). *Prochlorococcus* was enumerated
305 using forward scatter and red fluorescence. *Synechococcus* was enumerated by emission in the
306 orange wave lengths. Heterotrophic bacteria were stained with SYBR Green (Marie et al. 1997)
307 and enumerated. All samples were counted using a BD FacsJazz flow cytometer. The total
308 number of bacteria were calculated by summing the absolute abundances of *Prochlorococcus*,
309 *Synechococcus*, and heterotrophic bacteria. Primary production was measured by ^{13}C-
310 bicarbonate uptake as described in detail in (Baer et al. 2019). Briefly, carbon uptake rates were
311 divided by the total phytoplankton biomass and were normalized by decomposition rates (34.2 ±

312 3.3%) and by the proportion of daylight during the incubation (nmol C L^{-1} $daylight^{-1}$), which was
313 calculated as the percentage of PAR during the incubation compared to the total daily PAR.
314 Correlations of alpha-diversity with the total number of bacteria, absolute abundance of
315 *Prochlorococcus*, absolute abundance of heterotrophic bacteria, and primary production were
316 calculated using Pearson's correlation coefficient.
317
318 *Statistical analysis of environmental data*
319 General additive models (GAMs) were constructed to determine the relationships
320 between alpha-diversity and environmental factors. The GAMs were applied separately to the
321 western and eastern transects using the "gam" function in the "mgcv" package. The GAMs were
322 constructed using the restricted maximum likelihood method with thin plate regression spline
323 smooths applied to each explanatory variable (temperature, nutricline depth, phosphate
324 concentrations, and Fe-stress) using the "s" function. The ratio of the squared Euclidean norms
325 of the vectors for each pair of explanatory variables was calculated using the "concurvity"
326 function in "mgcv". All variables had low concurvity (<0.90), indicating that they could not be
327 approximated by one or more of the other variables in the model. Additionally, all variables had
328 significant relationships ($p < 0.05$) with alpha-diversity.
329 Boxplots were constructed for temperature, salinity, nutricline depth, Fe-stress, phosphate
330 concentrations, nitrate concentrations, POC, PON, POP, C:N, C:P, and N:P to visualize how
331 these factors varied across the bioregions along each transect. A one-way analysis of variance
332 followed by a *post hoc* Tukey's test was performed for each factor on each transect to determine
333 if the means differed significantly ($p < 0.05$) among the bioregions. Results of the Tukey's tests
334 were displayed on the boxplots using compact letter display.

Results

We quantified bacterial biogeography across the Indian Ocean by analyzing the 16S rRNA bacterial gene from 465 samples collected on Bio-GO-SHIP meridional sections. Samples were collected along a western section (I07N, n = 250) from Durban, South Africa to Mormugao, India (April 23 – June 6, 2018), and along an eastern section (I09N, n = 215) from Freemont, Australia to Phuket, Thailand (March 22 – April 24, 2016) – both during the spring intermonsoon season. We classified distinct bioregions based on bacterial community structure, analyzed similarities and differences in phylogenetic composition and alpha-diversity in each bioregion, and identified environmental drivers in order to understand the complex bacterial biogeography of the Indian Ocean.

Geography of the bioregions

Microbially defined bioregions partitioned latitudinally along both the western and eastern sections. We identified 23 bioregions with significantly different community assemblages, leading to 11 and 12 bioregions in the western and eastern sections, respectively (Figure 1a and 1b). There was clear geographic separation between bioregions in the southern Indian Ocean (30°S – 12°S), with four bioregions occurring off the coast of Madagascar and four bioregions occurring in the southeastern gyre. Additionally, there was one bioregion in the southeastern and southwestern gyre that overlapped with the equatorial region (12°S – 5°N). Within each section, there was some overlap between bioregions in both the equatorial and northern Indian Ocean (5°N – 18°N). On the western side of the basin, there were three bioregions distinct to the equatorial zone and two overlapping with the Arabian Sea. On the

eastern side of the basin, there were two bioregions that were distinct to the equatorial zone and two that were spread throughout the equatorial zone and the Bay of Bengal. Different latitudinal trends were observed in the northern Indian Ocean for the Arabian Sea and Bay of Bengal. In the Arabian Sea, one large bioregion exhibited minimal overlap with the equatorial-Arabian Sea bioregions. In the Bay of Bengal, three smaller bioregions had a high degree of variability and overlap between them. The Bay of Bengal overall exhibited the most heterogeneity in bioregion structure. Thus, we observed clear geographic separation of bioregions at southern latitudes and increasing overlap between bioregions occurring at mid- and northern-latitudes.

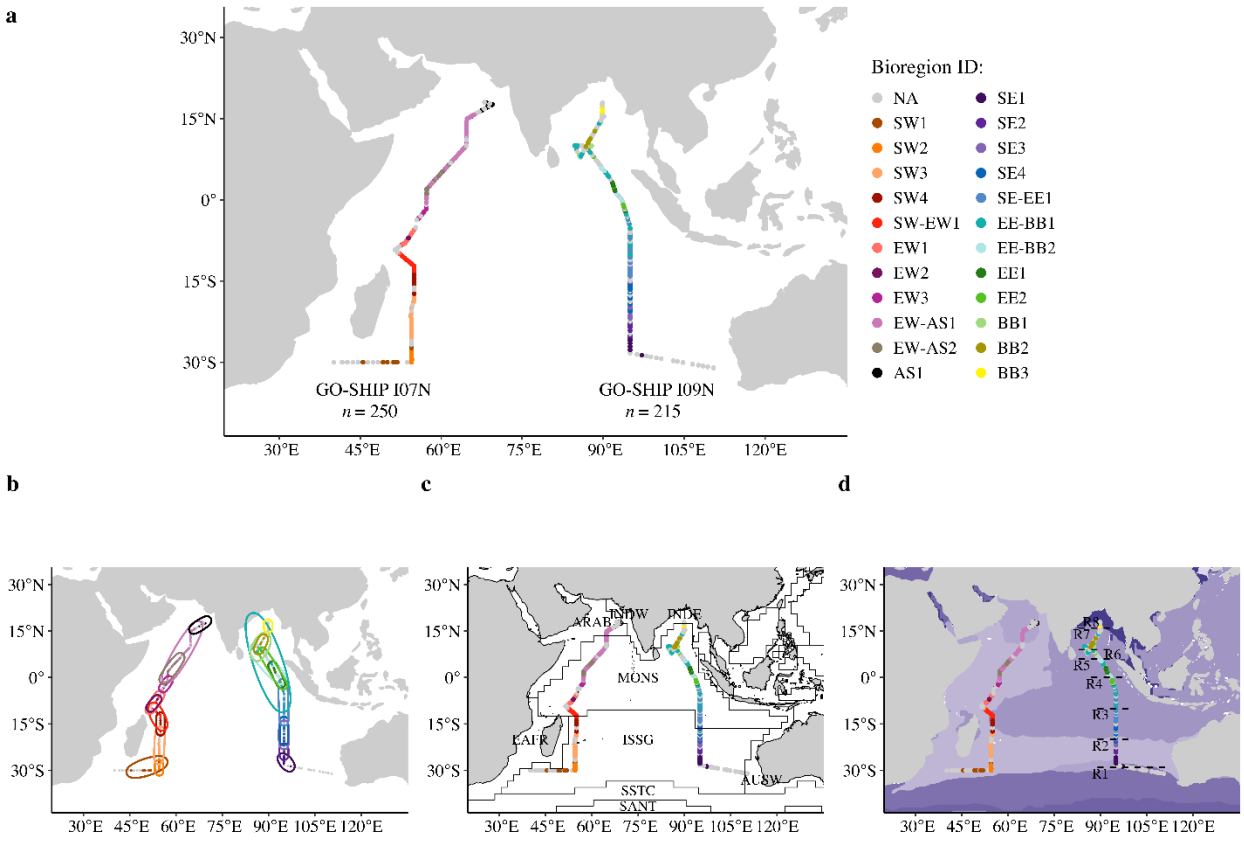


Figure 1: Bioregion geography. (**a** and **b**) Clustering analysis revealed partitioning of surface bacterial communities across the Indian Ocean. Each color represents a distinct bioregion, and

grey dots represent samples that did not cluster into a bioregion. Ellipses show geographic extent of bioregions. (**c**) Comparison of bioregions with Longhurst provinces. (**d**) Comparison of bioregions with ecological marine units (EMUs). Each shade of purple represents a different EMU. SW = southwestern gyre, SW-EW = southwestern-equatorial western, EW = equatorial western, EW-AS = equatorial western-Arabian Sea, AS = Arabian Sea, SE = southeastern gyre, SE-EE = southeastern gyre-equatorial eastern, EE-BB = equatorial eastern-Bay of Bengal, EE = equatorial eastern, and BB = Bay of Bengal.

Microbial ecology of the bioregions

Taxonomic diversity

Multiple ubiquitous and cosmopolitan taxa dominated the Indian Ocean, but only a few taxa were uniquely associated with each bioregion. The phyla Actinobacteria, Bacteroidetes, Cyanobacteria, Marinimicrobia (SAR406 clade), Proteobacteria, and Verrucomicrobia were ubiquitous (i.e., present in all samples) across the Indian Ocean, with Proteobacteria (35 - 61%) and Cyanobacteria (18 - 48%) being the most frequent (Supplemental Figure 5). At a finer phylogenetic level, seven genera were ubiquitous across the Indian Ocean, including *Candidatus* Actinomarina (Actinobacteria), *Prochlorococcus* (Cyanobacteria), SAR11 Clade 1a and Clade 1b (Proteobacteria), and the NS2b, NS4, and NS5 marine groups (Bacteroidetes) (Figure 2c and 2d). Of these ubiquitous genera, *Prochlorococcus,* SAR11 Clade 1a, and SAR11 Clade 1b were the most abundant. They had minimal variation across the bioregions (Figure 3) and were dominated by a single OTU (57.7% - 100% relative frequency), with the exception of SAR11 Clade 1b on the eastern transect (Supplemental Figure 6). Beyond these dominant OTUs, the less common OTUs of *Prochlorococcus*, SAR11 Clade 1a, and SAR11 Clade 1b composed unique

392 microdiverse communities with no bioregions sharing the same community (Supplemental
393 Figure 7). Thus, although these genera were ubiquitous across the Indian Ocean, microdiversity
394 within these genera may have played an important role in structuring communities across the
395 Indian Ocean. There were 13 cosmopolitan genera (i.e., present in ≥75% but <100% of samples),
396 including several OM clades (Proteobacteria), SAR92 (Proteobacteria), and *Synechococcus*
397 (Cyanobacteria) (Figure 2e and 2f). Combined, the ubiquitous and cosmopolitan genera
398 composed a large fraction of the bacterial community (46% – 78%). There were no endemic
399 genera (i.e., present in ≥75% of samples within a single bioregion and 0% of samples in all other
400 bioregions). However, there was one genus that was uniquely associated with a particular
401 bioregion (i.e., found in ≥75% of samples in one bioregion and ≤25% of samples in all other
402 bioregions). *Aurantivirga* was uniquely associated with a southeastern gyre bioregion (SE1) but
403 was found in low relative frequency (<0.36%). Overall, we observed that a small number of
404 ubiquitous and cosmopolitan genera dominated the Indian Ocean bioregions and that a single
405 taxon was bioregion-specific.
406 Bioregion shifts in taxa with important biogeochemical functions were also observed.
407 Significant changes in the number of nitrogen-fixers were observed along the western section.
408 UCYN-A was common in southwestern Indian Ocean gyre bioregions (Figure 3, Supplemental
409 Table 5), while *Trichodesmium* was in higher abundances within the Arabian Sea bioregions
410 (Figure 3, Supplemental Table 5). Significant differences in the number of sulfur-oxidizers were
411 also observed along the western transect with *Thiomicrorhabdus* occurring in higher abundances
412 within southwestern gyre bioregions compared to equatorial and Arabian Sea bioregions (Figure
413 3, Supplemental Table 5). *Sulfitobacter* occurred in higher abundances within Arabian Sea
414 bioregions compared to southwestern gyre bioregions (Supplemental Table 5). Significant

changes in the number of putative alkane degraders were observed along the eastern transect with *Alcanivorax* and *Oleibacter* occurring in lower abundances in the southeastern gyre bioregions compared to the equatorial and Bay of Bengal bioregions (Figure 3, Supplemental Table 5). Thus, bioregions across the Indian Ocean contained different abundances of several genera with important nitrogen, sulfur, and carbon cycle potentials.

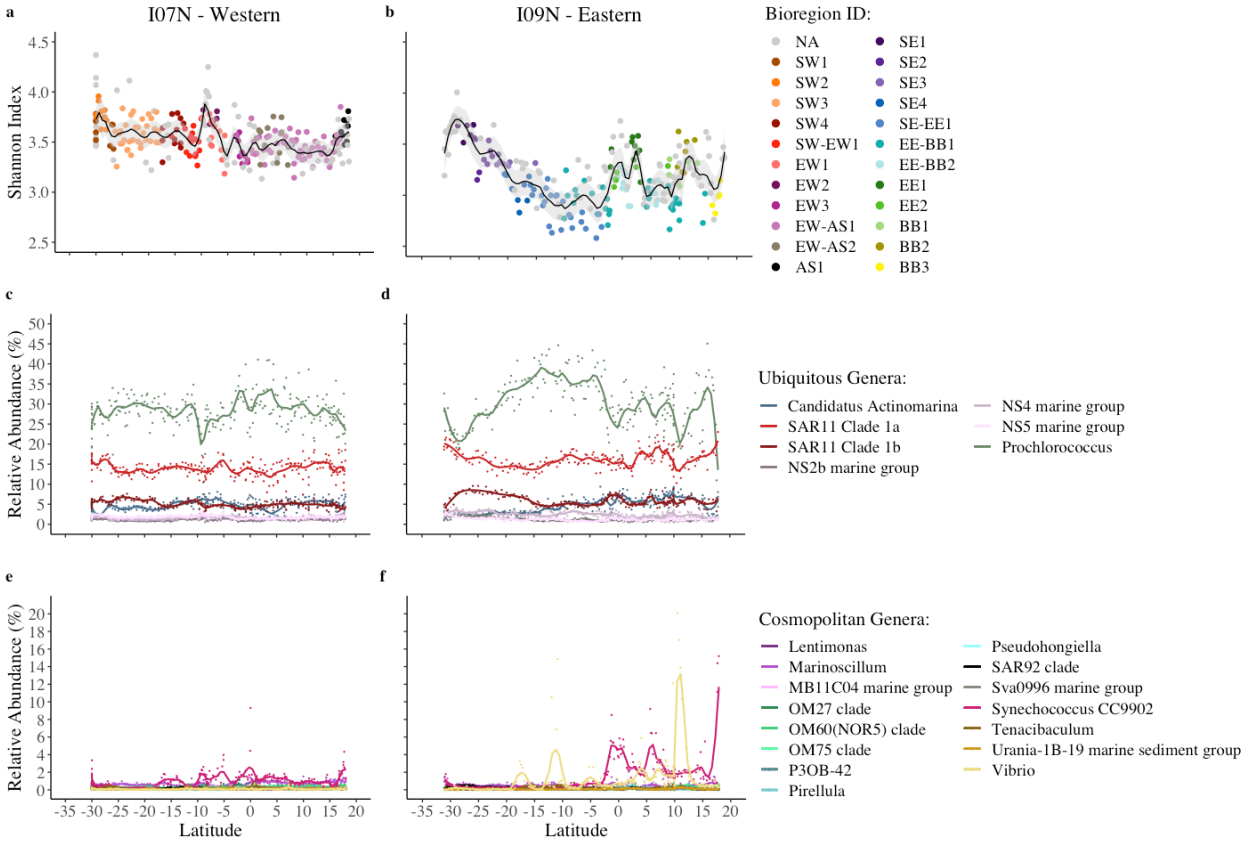


Figure 2: Diversity and taxonomic gradients. (**a** and **b**) Trends in alpha-diversity showed low variability along the western transect and systematic latitudinal variation along the eastern transect. Black lines represent smoothing curves fit with a polynomial regression (span = 0.1), and grey area represents the 95% confidence intervals of the smoothed curves. (**c, d, e,** and **f**) Relative abundances of ubiquitous and cosmopolitan genera showed that there is a primary community composition across the Indian Ocean. Solid lines represent smoothed curves fit with a polynomial regression (span = 0.1) for each genus.

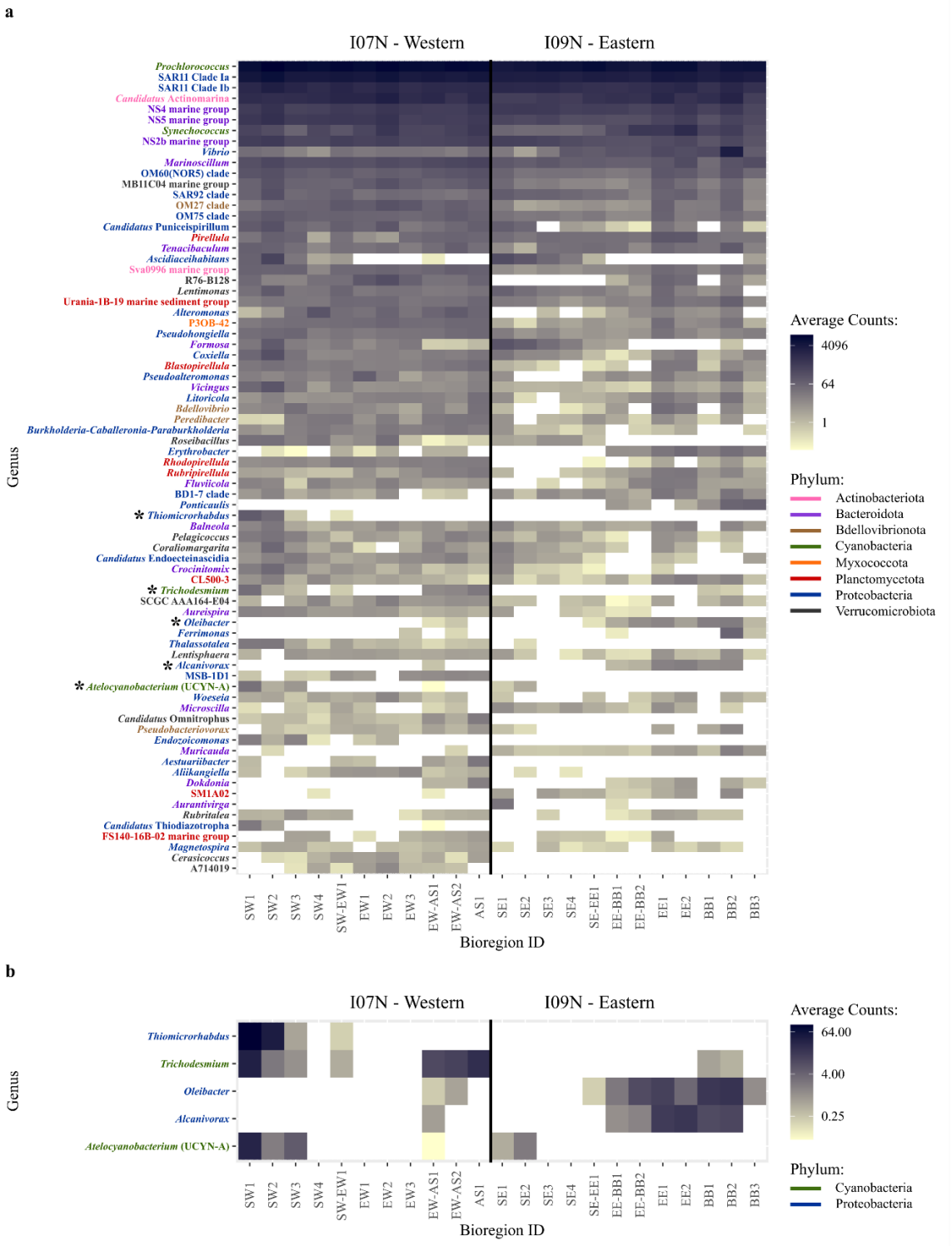


Figure 3: Heatmap of genera according to bioregion. Bioregions contained different numbers of genera with known biogeochemical functions. (**a**) Taxa of interest are denoted by *. (**b**) Subset plot showing only taxa of biogeochemical interest.

Alpha-diversity

Moderate changes in bacterial alpha-diversity were observed across bioregions. Low variability in alpha-diversity was observed along most of the western transect, as bioregions commonly had a Shannon Index near 3.5 (Figure 2a). The exception was the somewhat elevated alpha-diversity in a southwestern gyre bioregion (SW2) and an equatorial bioregion (EW2) (Supplemental Figure 8a). Alpha-diversity was more variable along the eastern transect (Figure 2b). The southeastern Indian Ocean gyre had a similar Shannon Index as the western side but declined northwards through several bioregions (SE2, SE3, and SE4). There was a sharp increase in alpha-diversity near the equator, peaking in bioregions EE1 and EE2 (Supplemental Figure 8b). Alpha-diversity was more variable north of the equator (Supplemental Figure 8b). To better understand how community dynamics contributed to variability in alpha-diversity, the Shannon Index was decomposed into richness and evenness components. On the western transect, the increase in alpha-diversity in bioregion SW2 corresponded with an increase in richness (i.e., the number of OTUs) (Supplemental Figure 9c), while the increase in bioregion EW2 was tied to evenness (Supplemental Figure 9e). On the eastern transect, the large decrease in alpha-diversity in the southeastern gyre from SE1 through SE4 corresponded with decreases in evenness (Supplemental Figure 9f). In contrast, the peak in alpha-diversity within bioregion EE1and EE2 corresponded with an increase in richness (Supplemental Figure 9d). Thus, changes in alpha-diversity across the bioregions were impacted by shifts in both composition and relative abundance.

Alpha-diversity gradients significantly correlated with changes in the relative frequency of dominant genera. Along both transects, alpha-diversity negatively correlated with the relative abundance of *Prochlorococcus* (I07N: $r = -0.638$, *adj*. $p < 0.001$, df = 249; I09N: $r = -0.800$, *adj*.

$p < 0.001$, df = 213). This negative relationship was also seen when using absolute *Prochlorococcus* cell counts measured using flow cytometry ($r = -0.63$, $p < 0.001$, df = 29) (Supplemental Figure 10). In contrast, alpha-diversity positively correlated with the relative abundance of many of the heterotrophic ubiquitous and cosmopolitan genera (Supplemental Table 6) but did not have a significant relationship with absolute cell counts of HNA heterotrophs ($r = -0.16$, $p = 0.405$, df = 29). This suggests an important role of *Prochlorococcus* abundance for Indian Ocean bacterial alpha-diversity.

Trends in alpha-diversity were weakly tied to environmental factors for both the western and eastern transect. A combination of temperature, nutricline depth, phosphate concentrations, and Fe-stress explained 32.6% and 32.8% of the deviance in alpha-diversity for I07N and I09N, respectively (Supplemental Figure 11 and Supplemental Figure 12). Of these variables, Fe-stress explained the most deviance. We observed a parallel decrease in alpha-diversity with Fe-stress on the eastern side. However, the correspondence between Fe-stress and alpha-diversity was non-monotonic on the western side. Temperature, nutricline depth, and surface nutrient concentrations each explained less than 10% of the deviance in alpha-diversity and had varying relationships with alpha-diversity on each transect. Additionally, primary production had a non-significant relationship with alpha-diversity ($p = 0.089$, df = 19) (Supplemental Figure 13). Overall, our observations suggested that shifts in alpha-diversity were not tied to common environmental drivers.

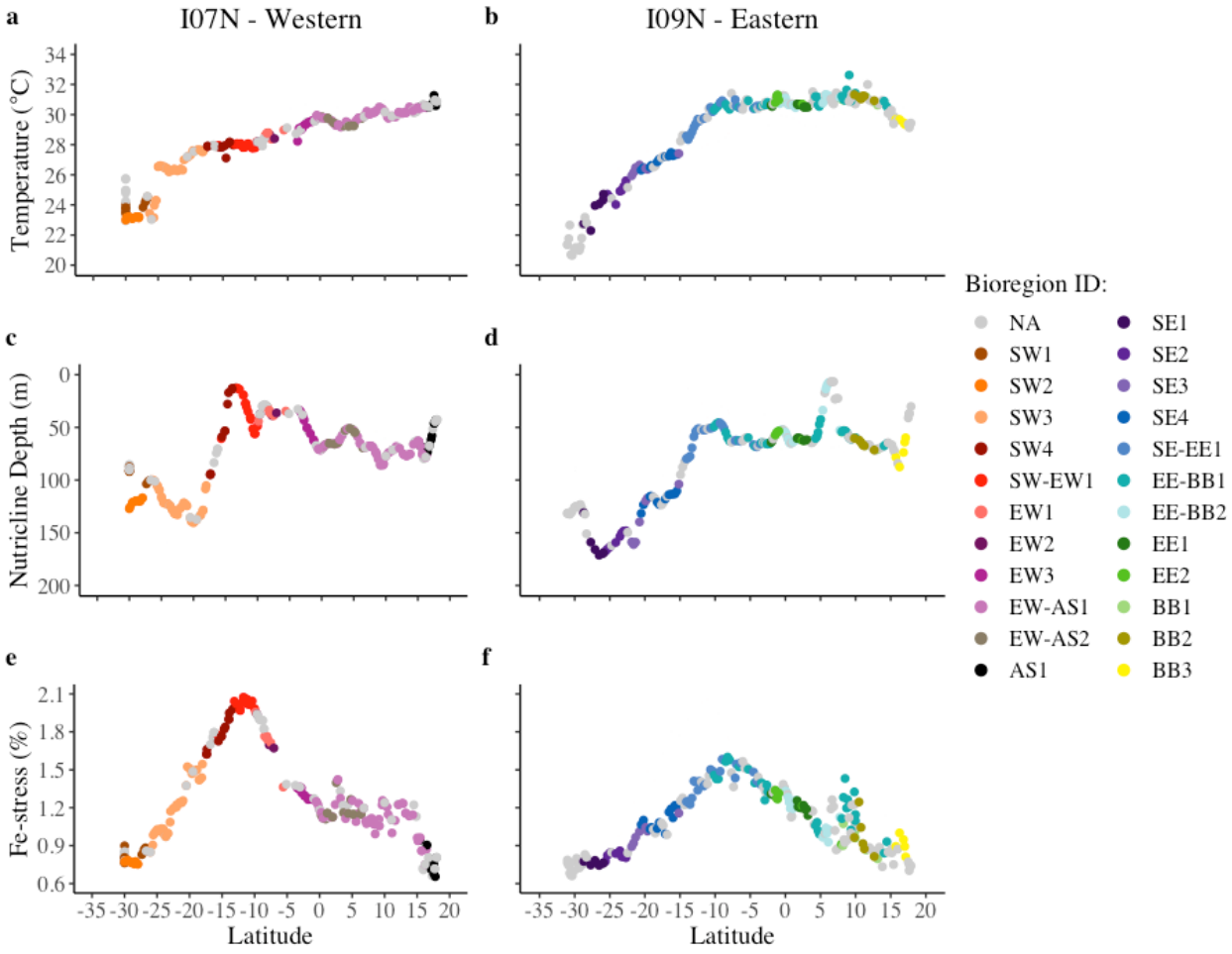
Environmental characteristics of the bioregions

Bioregions were characterized by subtle transitions in physical and geochemical conditions. To link beta-diversity with environmental drivers, changes in environmental

conditions across the bioregions were identified using boxplots and ANOVAs. The southwestern gyre bioregions varied in temperature and nutricline depth (Figure 4a and 4c, Figure 5, Supplemental Figure 14, Supplemental Table 7). Further north, temperature and nutricline depth exhibited minimal variation within the western equatorial and Arabian Sea bioregions. However, these bioregions showed clear differences in Fe-stress with Fe-stress being high in western equatorial bioregions and low in Arabian Sea bioregions (Figure 4e, Figure 5, Supplemental Figure 14, Supplemental Table 7). The eastern transect showed parallel patterns, whereby the southeastern gyre bioregions varied in temperature and nutricline depth, while the equatorial and Bay of Bengal bioregions varied by the type of nutrient stress (Figure 4f, Figure 5, Supplemental Figure 15, Supplemental Table 7). Therefore, although clear drivers of beta-diversity (i.e., bioregions) could not be identified, perhaps because the communities are very similar, there were subtle gradients in temperature as well as nutrient stress type and severity within unique bioregions.

Physical dynamics, such as eddies, fronts, and topographical barriers, also influenced the bioregion distribution. A large, persistent eddy led to the formation of a small bioregion on the eastern transect (BB3) (Supplemental Figure 16b). Smaller, more transient eddies led to intra-bioregion variability within the southern Indian Ocean gyre and the Bay of Bengal (Supplemental Figure 16). Currents and fronts that intersected the transects also resulted in different community structures and bioregions. Along the western transect, the southwestern gyre bioregion, SW4, aligned with two currents formed by the westward flowing South Equatorial Current, namely the Southeast Madagascar Current and the Northeast Madagascar Current (Supplemental Figure 17a). Along the eastern transect, variability within the southeastern bioregion, SE3, at 22.5°S corresponded with a front formed by the eastward flowing

Eastern Gyral Current. Within the southeastern bioregion, SE4, at 16.5°S, variability corresponded with a front formed by the westward flowing South Equatorial Current (Supplemental Figure 17b). Lastly, topographical barriers led to transitions between bioregions. For example, Seychelles Island intersected the western transect resulting in different bioregions south (EW1) versus north (EW3) of the island (Supplemental Figure 17a). Thus, physical dynamics have a pronounced impact on community structure and result in interspersion of bioregions or in the formation of a new bioregion.



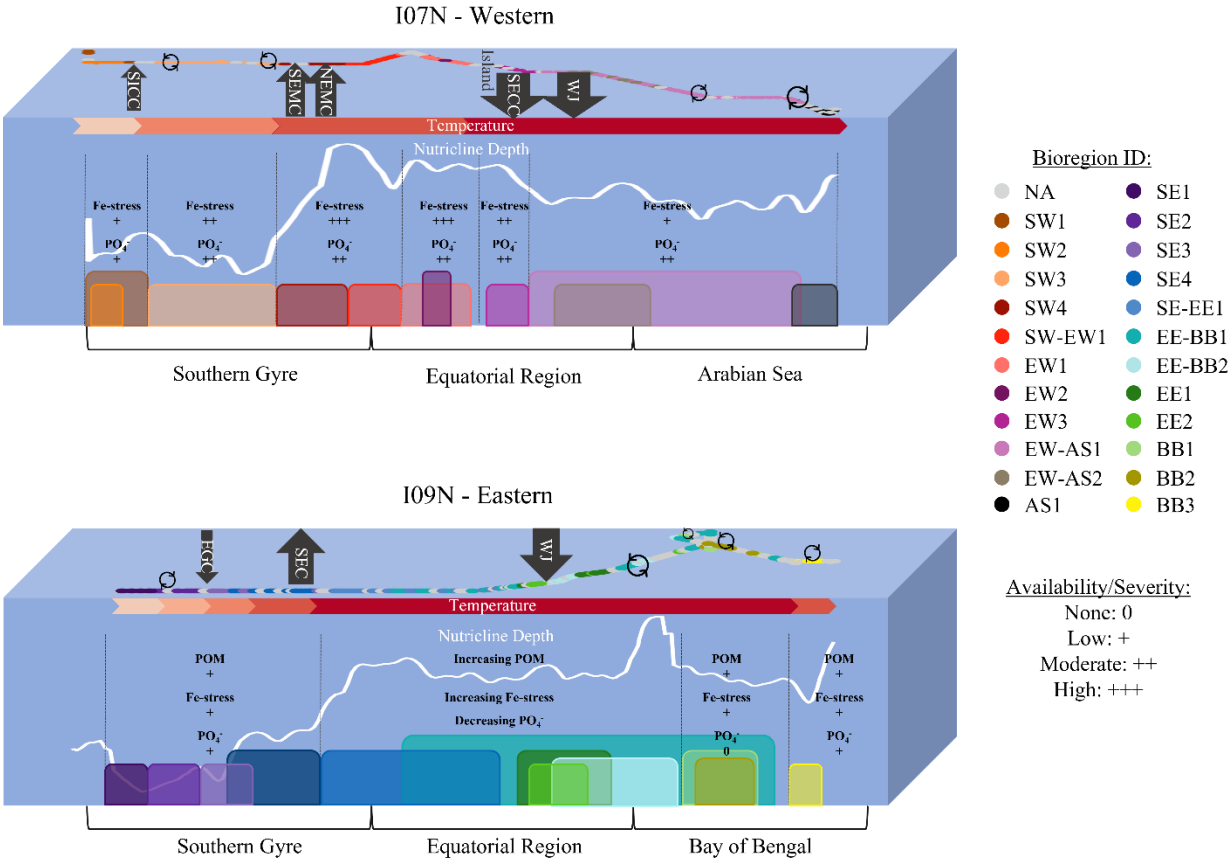
508

509 **Figure 4: Environmental conditions of the bioregions.** (**a** and **b**) Temperature (°C) measured

510 using a mounted near-surface thermosalinograph. (**c** and **d**) Nutricline depth (m) which was

511 defined as the depth at which nitrate was ≥1 μmol L^{-1}. (**e** and **f**) Fe-stress (%) parameter, ϕ_{sat},

512 estimated from MODIS-Aqua satellite fluorescence data.



513

514 **Figure 5: Conceptual diagram of Indian Ocean bioregions.** Southern Indian Ocean gyre
515 bioregions were characterized by differences in temperature and nutrient supply, whereas the
516 equatorial and northern Indian Ocean bioregions were influenced by physical mixing and were
517 characterized by differences in nutrient availability on the western and eastern transects and by
518 POM concentrations on the eastern transect. Arrows represent known currents in the Indian
519 Ocean and are pointing in the direction that they flow: SICC = South Indian Countercurrent,
520 SEMC = South East Madagascar Current, NEMC = North East Madagascar Current, SECC =
521 South Equatorial Countercurrent, WJ = Wyrtki Jets, EGC = Eastern Gyral Current, and SEC =
522 South Equatorial Current.

Discussion

Here, we first asked if bacterial communities were structured spatially and identified 23 distinct bioregions across the Indian Ocean seascape. In terrestrial systems, biomes are defined as areas with similar climate and plant communities, with the plant communities (i.e., phototrophs) functioning as a bottom-up control on animal communities (i.e., heterotrophs). In marine systems, similar relationships are observed between phototrophs and heterotrophs, where the composition of the heterotrophic community is shaped by the composition of the phototrophs and phototroph exudates (Mühlenbruch et al. 2018). At the genus level, the Indian Ocean is dominated by *Prochlorococcus* with *Synechococcus* and nitrogen-fixers appearing in low abundances within specific regions. Thus, based on phytoplankton community composition, the Indian Ocean would appear uniform, whereas we observed distinct communities based on bacterial diversity. A variety of metrics have previously delineated the pelagic Indian Ocean into broad ecological regions such as two broad provinces (Longhurst 2010) (Figure 1c), four ecological marine units (EMUs) (Sayre et al. 2017) (Figure 1d), and ~10 seascapes (Kavanaugh et al. 2014) (Supplemental Figure 1). There was strong longitudinal separation between the eastern and western microbially defined bioregions, which agrees with the longitudinal separation represented by EMUs (Sayre et al. 2017) and seascapes (Kavanaugh et al. 2014). Furthermore, microbially defined bioregions suggest that there are more regions than have been previously described and that these regions have subtle environmental drivers which could provide novel insights to finer-scale ecosystem changes. For example, there were clear bioregion transitions across the southwestern and southeastern Indian Ocean gyre that were masked with remote-sensing derived metrics. Regional analyses of Indian Ocean bacterial community structures have also detected finer ecosystem partitioning, supporting our findings (Jeffries et al.

546 2015; Zheng et al. 2016). Additionally, mesoscale processes, such as eddies, led to intra-
547 bioregion variability or the formation of a unique bioregion, indicating that bioregions capture
548 transient, dynamic features and are not always permanent features of the seascape. Thus,
549 partitioning bioregions using fine-scale sampling of bacterial communities across large
550 geographic gradients generated a highly resolved Indian Ocean seascape that was shaped by
551 dynamic features and subtle environmental changes.

552 Second, we asked where the biggest gradients in bacterial alpha-diversity occurred and what
553 environmental factors corresponded with changes in alpha-diversity? The more pronounced
554 latitudinal alpha-diversity gradient on the eastern side followed trends previously observed off
555 the western coast of Australia (34°S – 12°S), where bacterial richness peaked at 34°S and
556 decreased towards the tropics (Raes et al. 2018b). Additionally, a diagonal transect taken from
557 the southwestern Indian Ocean gyre to the western coast of Indonesia also showed that alpha-
558 diversity remained fairly constant throughout the southwestern gyre (Zheng et al. 2016), despite
559 strong gradients in temperature and nutrient supply in this region. Alpha-diversity was expected
560 to increase with temperature (Fuhrman et al. 2008) or primary productivity (Raes et al. 2018a).
561 Within the Indian Ocean, temperature and nutrient supply are uniquely positively correlated with
562 the warmest regions being the most nutrient replete (Garcia et al. 2018). Additionally, within the
563 eastern Indian Ocean, primary production increases northwards from the gyre (Baer et al. 2019).
564 However, we did not detect a clear positive northward trend in alpha-diversity on either the
565 western or eastern side suggesting that temperature, nutrient supply, or productivity are not the
566 primary drivers here. In our study, Fe-stress explained the most deviance in alpha-diversity, but
567 the different relationships observed on the western versus eastern side makes the role of Fe-stress
568 as a regulator of alpha-diversity difficult to decipher. Here, we observed a strong relationship

between alpha-diversity and the relative abundance of *Prochlorococcus*. Since the 99% OTU level was used to define alpha-diversity, relative abundance was a composite measurement of the microdiversity within the *Prochlorococcus* genus. In the eastern Indian Ocean, the abundance of four *Prochlorococcus* haplotypes varied latitudinally, and the variation within each haplotype was attributed to different combinations of environmental factors (Larkin et al. 2020). Therefore, the strong niche partitioning of microdiverse *Prochlorococcus* lineages may indicate why it was difficult to identify environmental drivers of alpha-diversity across larger geographic gradients. Overall, the Indian Ocean had moderate gradients in biodiversity, especially on the eastern side, but the environmental drivers could not be identified.

Third, we asked what lineages were endemic to these bioregions and what were the major ubiquitous lineages across the Indian Ocean? We found that there were no endemic taxa and that the Indian Ocean microbiome was primarily composed of a core set of taxa including *Candidatus* Actinomarina, *Prochlorococcus*, SAR11 Clade 1a and Clade 1b, and the NS2b, NS4, and NS5 marine groups. Here, *Prochlorococcus* was the most dominant genera along both transects matching previous estimates of cell counts (Baer et al. 2019). Estimates of nutrient limitation derived from *Prochlorococcus* genes, indicate that the Indian Ocean experiences a wide range of nutrient limitation types including N-, P-, and Fe-limitation as well as P/N- and N/Fe-colimitation with particularly high variability in nutrient stress type occurring throughout the Bay of Bengal (Ustick et al. 2021). These variations in nutrient limitation type approximately align with several of our bioregions. Additionally, we see taxonomic groups with traits such as nitrogen fixation, sulfur oxidation, and alkane degradation in shifting abundances across bioregions, indicating variations in biogeochemical processes. Thus, these bioregions may also differ in ecological functions, but metagenomic analysis is required to confirm this.

592 Fourth, we asked how did geochemical and physical dynamics vary across the bioregions?
593 We found that bioregions identified from bacterial biodiversity suggested a complex interplay
594 between geochemistry and physical dynamics across the oligotrophic Indian Ocean. Regions of
595 Fe-stress emerge in the western tropical region due to low dust deposition and mild upwelling
596 (Wiggert et al. 2006; Behrenfeld et al. 2009), while dust deposition from Australia (McGowan
597 and Clark 2008) may relieve Fe-stress in the eastern subtropical region. These shifts in Fe-stress
598 mirror shifts in bioregions, particularly off the eastern coast of Madagascar, where four
599 bioregions partitioned along a gradient of Fe-stress. A large portion of the Indian Ocean
600 experiences N-stress (Twining et al. 2019; Ustick et al. 2021), which can be relieved by
601 mesoscale physical processes (e.g., eddies). While eddies may be thought to provide short-term
602 influxes of nutrients, their dominance throughout the Bay of Bengal (Cui et al. 2016) and the
603 Arabian Sea (Sevsu and Al-Jufaili 2013), appear to have a larger scale impact. Indeed, we see
604 changes in community structure and sometimes entire bioregions are associated with eddy-
605 impacted areas across the northern Indian Ocean (Supplemental Figure 16). We also detect
606 different bioregions south versus north of Seychelles Island in the western Indian Ocean,
607 suggesting that microbial communities respond to subtle changes in nutrient supply due to island
608 proximity. During the intermonsoon season, there is a gradient of N-stress and co-limitation by P
609 and Fe in the eastern basin (Twining et al. 2019) coupled with a positive latitudinal trend in POM
610 concentrations (Garcia et al. 2018). We observed that changes in POM, and likely plankton
611 biomass, corresponded with shifting bioregions along the eastern transect, and thus is an
612 important factor to consider when delineating ecological regions across the Indian Ocean.
613 Gradients of nutrient availability, POM concentrations, and multi-dimensional physical factors

were defining environmental features of our bioregions, suggesting that they should be included as factors for defining marine ecological regions.

Results presented here have implications beyond addressing fundamental microbial ecology questions across large spatial scales. Many of the environmental variables that we identified as shaping bacterial communities (e.g., POC and physical drivers) are measurable from satellites. Thus, incorporating these additional factors into marine ecological regions that are defined by remote sensing may lead to improved delineation across warm, oligotrophic regions. Additionally, the Indian Ocean is historically under-sampled compared to other oceans, so the regional ecological consequences of warming on biodiversity are unknown. GO-SHIP sections are re-sampled approximately every ten years, enabling the re-evaluation of bioregions. While the analyses presented here are the first step towards defining operational bioregions, bioregion delineations are currently dependent on the collected sample set and thus are not stable. However, what the presented analyses successfully do is identify key genera and their distributions across these dynamic bioregions. Future definitions of bioregions can be anchored in such keystone lineages and their distributions to delineate regions independently of the exact sample set. Re-evaluation of bioregions under this framework can identify possible spatial shifts as well as long-term changes due to anthropogenic forcings. Thus, we propose that a combination of in-situ ‘omics analysis of microbial communities, detailed hydrography, and remote sensing can greatly aid in identifying regions that are most vulnerable to anthropogenic impacts.

References

Aronesty, E. 2013. Comparison of sequencing utility programs. Open Bioinforma. J. **7**: 1–8. doi:10.2174/1875036201307010001

Al Azhar, M., Z. Lachkar, M. Lévy, and S. Smith. 2017. Oxygen minimum zone contrasts between the Arabian Sea and the Bay of Bengal implied by differences in remineralization depth. Geophys. Res. Lett. **44**: 11,106-11,114. doi:10.1002/2017GL075157

Baer, S. E., S. Rauschenberg, C. A. Garcia, N. S. Garcia, A. C. Martiny, B. S. Twining, and M. W. Lomas. 2019. Carbon and nitrogen productivity during spring in the oligotrophic Indian Ocean along the GO-SHIP IO9N transect. Deep. Res. Part II Top. Stud. Oceanogr. **161**: 81–91. doi:10.1016/j.dsr2.2018.11.008

Behrenfeld, M. J., and P. G. Falkowski. 1997. Photosynthetic rates derived from satellite-based chlorophyll concentration. Limnol. Oceanogr. **42**: 1–20. doi:10.4319/lo.1997.42.1.0001

Behrenfeld, M. J., T. K. Westberry, E. S. Boss, and others. 2009. Satellite-detected fluorescence reveals global physiology of ocean phytoplankton. Biogeosciences **6**: 779–794. doi:10.5194/bg-6-779-2009

Bolyen, E., J. R. Rideout, M. R. Dillon, and others. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. **37**: 852–857. doi:10.1038/s41587-019-0209-9

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat. Methods **13**: 581–583. doi:10.1038/nmeth.3869

Cermeño, P., S. Dutkiewicz, R. P. Harris, M. Follows, O. Schofield, and P. G. Falkowski. 2008. The role of nutricline depth in regulating the ocean carbon cycle. Proc. Natl. Acad. Sci. U.

S. A. **105**: 20344–20349. doi:10.1073/pnas.0811302106

Cui, W., J. Yang, and Y. Ma. 2016. A statistical analysis of mesoscale eddies in the Bay of Bengal from 22–year altimetry data. Acta Oceanol. Sin. **35**: 16–27. doi:10.1007/S13131-016-0945-3

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26**: 2460–2461. doi:10.1093/bioinformatics/btq461

Fuhrman, J. A., J. A. Steele, I. Hewson, M. S. Schwalbach, M. V. Brown, J. L. Green, and J. H. Brown. 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc. Natl. Acad. Sci. **105**: 7774–7778. doi:10.1073/pnas.0803070105

Garcia, C. A., S. E. Baer, N. S. Garcia, S. Rauschenberg, B. S. Twining, M. W. Lomas, and A. C. Martiny. 2018. Nutrient supply controls particulate elemental concentrations and ratios in the low latitude eastern Indian Ocean. Nat. Commun. **9**: 1–10. doi:10.1038/s41467-018-06892-w

Hood, R. R., S. Wajih, A. Naqvi, J. D. Wiggert, and A. Subramaniam. 2007. Biogeochemical and ecological research in the Indian Ocean. Eos, Trans. Am. Geophys. Union **88**: 144. doi:10.1029/2007EO120007

Hörstmann, C., E. J. Raes, P. L. Buttigieg, C. Lo Monaco, U. John, and A. M. Waite. 2021. Hydrographic fronts shape productivity, nitrogen fixation, and microbial community composition in the southern Indian Ocean and the Southern Ocean. Biogeosciences **18**: 3733–3749. doi:10.5194/BG-18-3733-2021

Jayakumar, A., M. M. D. Al-Rshaidat, B. B. Ward, and M. R. Mulholland. 2012. Diversity, distribution, and expression of diazotroph *nifH* genes in oxygen-deficient waters of the Arabian Sea. FEMS Microbiol. Ecol. **82**: 597–606. doi:10.1111/j.1574-6941.2012.01430.x

Jeffries, T. C., M. Ostrowski, R. B. Williams, and others. 2015. Spatially extensive microbial biogeography of the Indian Ocean provides insights into the unique community structure of a pristine coral atoll. Sci. Rep. **5**: 15383. doi:10.1038/srep15383

Kavanaugh, M. T., B. Hales, M. Saraceno, Y. H. Spitz, A. E. White, and R. M. Letelier. 2014. Hierarchical and dynamic seascapes: A quantitative framework for scaling pelagic biogeochemistry and ecology. Prog. Oceanogr. **120**: 291–304. doi:10.1016/j.pocean.2013.10.013

Larkin, A. A., C. A. Garcia, K. A. Ingoglia, N. S. Garcia, S. E. Baer, B. S. Twining, M. W. Lomas, and A. C. Martiny. 2020. Subtle biogeochemical regimes in the Indian Ocean revealed by spatial and diel frequency of *Prochlorococcus* haplotypes. Limnol. Oceanogr. **65**: S220–S232. doi:10.1002/lno.11251

Longhurst, A. 2010. Ecological geography of the sea, Elsevier.

Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**: 1–21. doi:10.1186/S13059-014-0550-8

Marie, D., F. Partensky, S. Jacquet, and D. Vaulot. 1997. Enumeration and cell cycle analysis of natural populations of marine picoplankton by flow cytometry using the nucleic acid stain SYBR Green I. Appl. Environ. Microbiol. **63**: 186–193. doi:10.1128/AEM.63.1.186-193.1997

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal **17**: 10. doi:10.14806/ej.17.1.200

McCreary, J. P., Z. Yu, R. R. Hood, P. N. Vinaychandran, R. Furue, A. Ishida, and K. J. Richards. 2013. Dynamics of the Indian-Ocean oxygen minimum zones. Prog. Oceanogr. **112**–**113**: 15–37. doi:10.1016/j.pocean.2013.03.002

McGowan, H., and A. Clark. 2008. Identification of dust transport pathways from Lake Eyre, Australia using Hysplit. Atmos. Environ. **42**: 6915–6925. doi:10.1016/j.atmosenv.2008.05.053

McMurdie, P. J., and S. Holmes. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data M. Watson [ed.]. PLoS One **8**: e61217. doi:10.1371/journal.pone.0061217

Mühlenbruch, M., H. Grossart, F. Eigemann, and M. Voss. 2018. Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. Environ. Microbiol. **20**: 8. doi:10.1111/1462-2920.14302

Needham, D. M., and J. A. Fuhrman. 2016. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. Nat. Microbiol. **1**: 16005. doi:10.1038/nmicrobiol.2016.5

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. **41**: D590–D596. doi:10.1093/NAR/GKS1219

Raes, E. J., L. Bodrossy, J. van de Kamp, A. Bissett, M. Ostrowski, M. V. Brown, S. L. S. Sow, B. Sloyan, and A. M. Waite. 2018a. Oceanographic boundaries constrain microbial diversity gradients in the South Pacific Ocean. Proc. Natl. Acad. Sci. **115**: 35. doi:10.1594/PANGAEA

Raes, E. J., L. Bodrossy, J. van de Kamp, A. Bissett, and A. M. Waite. 2018b. Marine bacterial richness increases towards higher latitudes in the eastern Indian Ocean. Limnol. Oceanogr. Lett. **3**: 10–19. doi:10.1002/lol2.10058

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: A versatile open

source tool for metagenomics. PeerJ **2016**: e2584. doi:10.7717/peerj.2584

Roxy, M. K., K. Ritika, P. Terray, and S. Masson. 2014. The curious case of Indian Ocean warming. J. Clim. **27**: 8501–8509. doi:10.1175/JCLI-D-14-00471.1

Sayre, R. G., D. J. Wright, S. P. Breyer, and others. 2017. A three-dimensional mapping of the ocean based on environmental data. Oceanography **30**: 90–103. doi:10.5670/oceanog.2017.116

Sebastián, M., E. Ortega-Retuerta, L. Gómez-Consarnau, M. Zamanillo, M. Álvarez, J. Arístegui, and J. M. Gasol. 2021. Environmental gradients and physical barriers drive the basin-wide spatial structuring of Mediterranean Sea and adjacent eastern Atlantic Ocean prokaryotic communities. Limnol. Ocean. **66**: 4077–4095. doi:10.1002/lno.11944

Sevsu, S., and S. Al-Jufaili. 2013. Coastal upwellings and mesoscale eddies of the western Arabian Sea: some biological implications. Int. J. Ocean. Oceanogr. **7**: 93–115.

Team, R. C. 2019. R: A language and environment for statistical computing.

Twining, B. S., S. Rauschenberg, S. E. Baer, M. W. Lomas, A. C. Martiny, and O. Antipova. 2019. A nutrient limitation mosaic in the eastern tropical Indian Ocean. Deep. Res. Part II Top. Stud. Oceanogr. **166**: 125–140. doi:10.1016/j.dsr2.2019.05.001

Ustick, L. J., A. A. Larkin, C. A. Garcia, and others. 2021. Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. Science (80-.). **372**: 287–291. doi:10.1126/science.abe6301

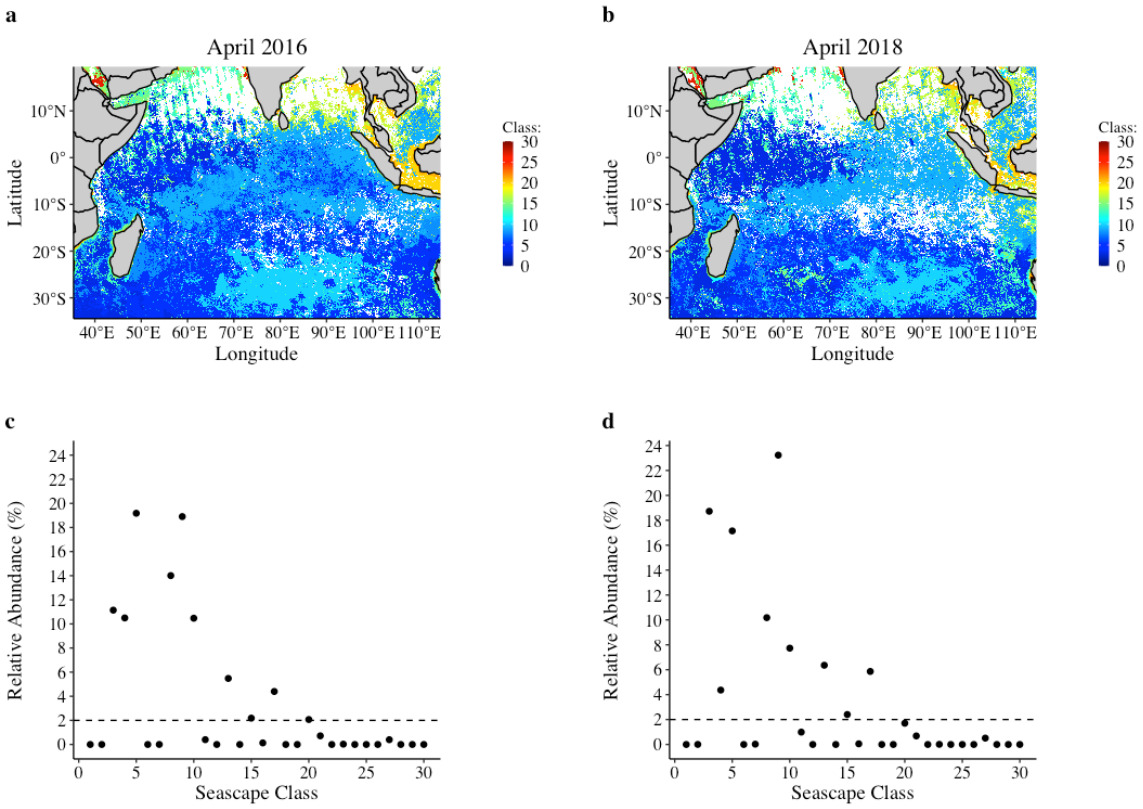
Vinayachandran, P. N. M., Y. Masumoto, M. J. Roberts, and others. 2021. Reviews and syntheses: Physical and biogeochemical processes associated with upwelling in the Indian Ocean. Biogeosciences **18**: 5967–6029. doi:10.5194/BG-18-5967-2021

Wiggert, J., R. Murtugudde, and J. Christian. 2006. Annual ecosystem variability in the tropical

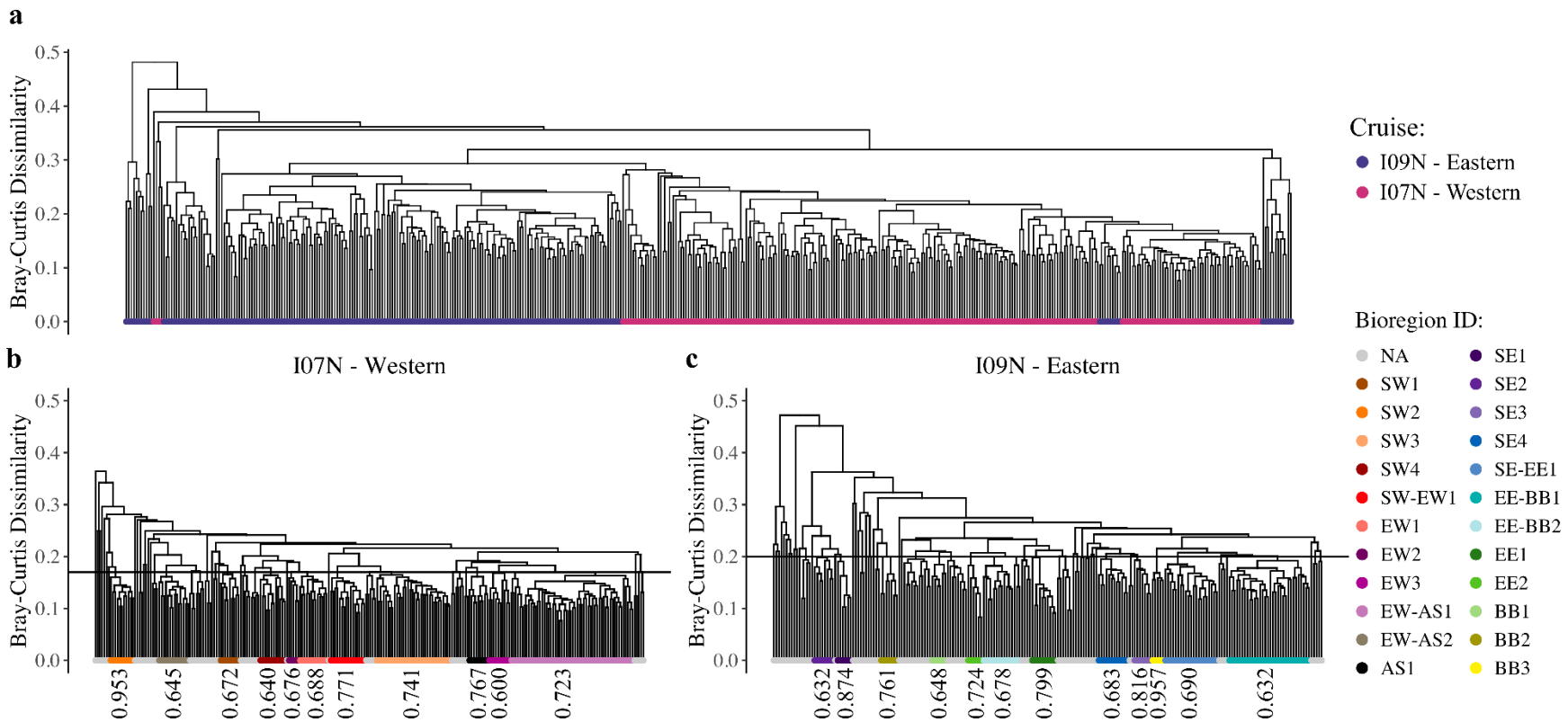
749 Indian Ocean: Results of a coupled bio-physical ocean general circulation model. Deep Sea
750 Res. Part II Top. Stud. Oceanogr. **53**: 644–676. doi:10.1016/j.dsr2.2006.01.027
751 Zheng, X., X. Dai, and L. Huang. 2016. Spatial variations of prokaryotic communities in surface
752 water from India Ocean to chinese marginal seas and their underlining environmental
753 determinants. Front. Mar. Sci. **3**: 17. doi:10.3389/fmars.2016.00017
754

Acknowledgements

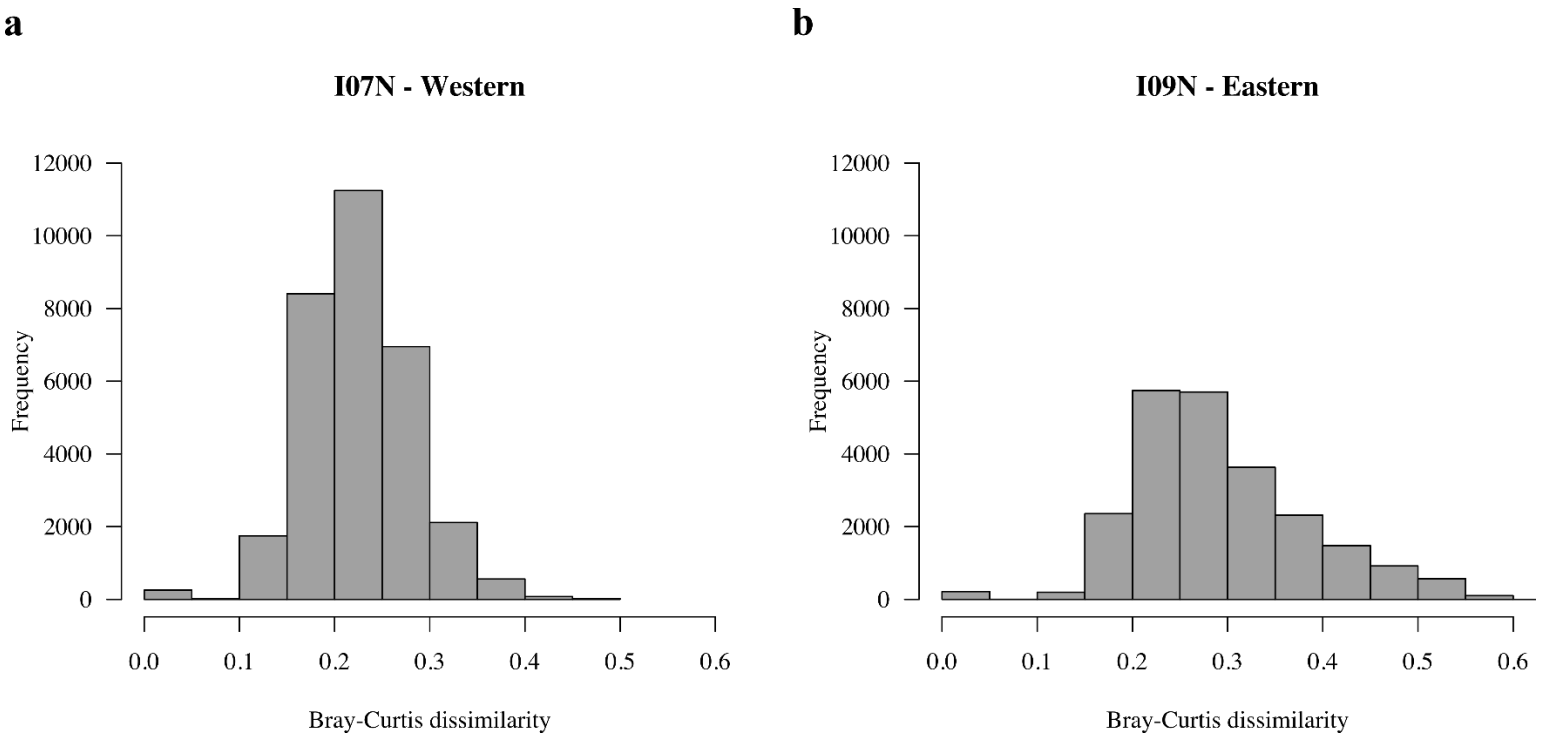
We thank many lab members for help with collecting and analyzing the samples and the GO-SHIP community for enabling cruise participation and sampling. The authors declare that they have no competing interests. **Funding:** For support, we thank the National Science Foundation (1559002, 1848576, and 1948842 to ACM), the National Oceanic and Atmospheric Administration (101813-Z7554214 to ACM), the National Aeronautics and Space Administration (80NSSC21K1654 to ACM), and the National Institutes of Health (T32AI141346 to MLB). **Data and materials available:** Raw sequence reads are available via NCBI SRA (BioProject ID: PRJNA656268). GO-SHIP cruise metadata are available via cchdo.ucsd.edu, and POM data are available via BCO-DMO (https://www.bco-dmo.org/).



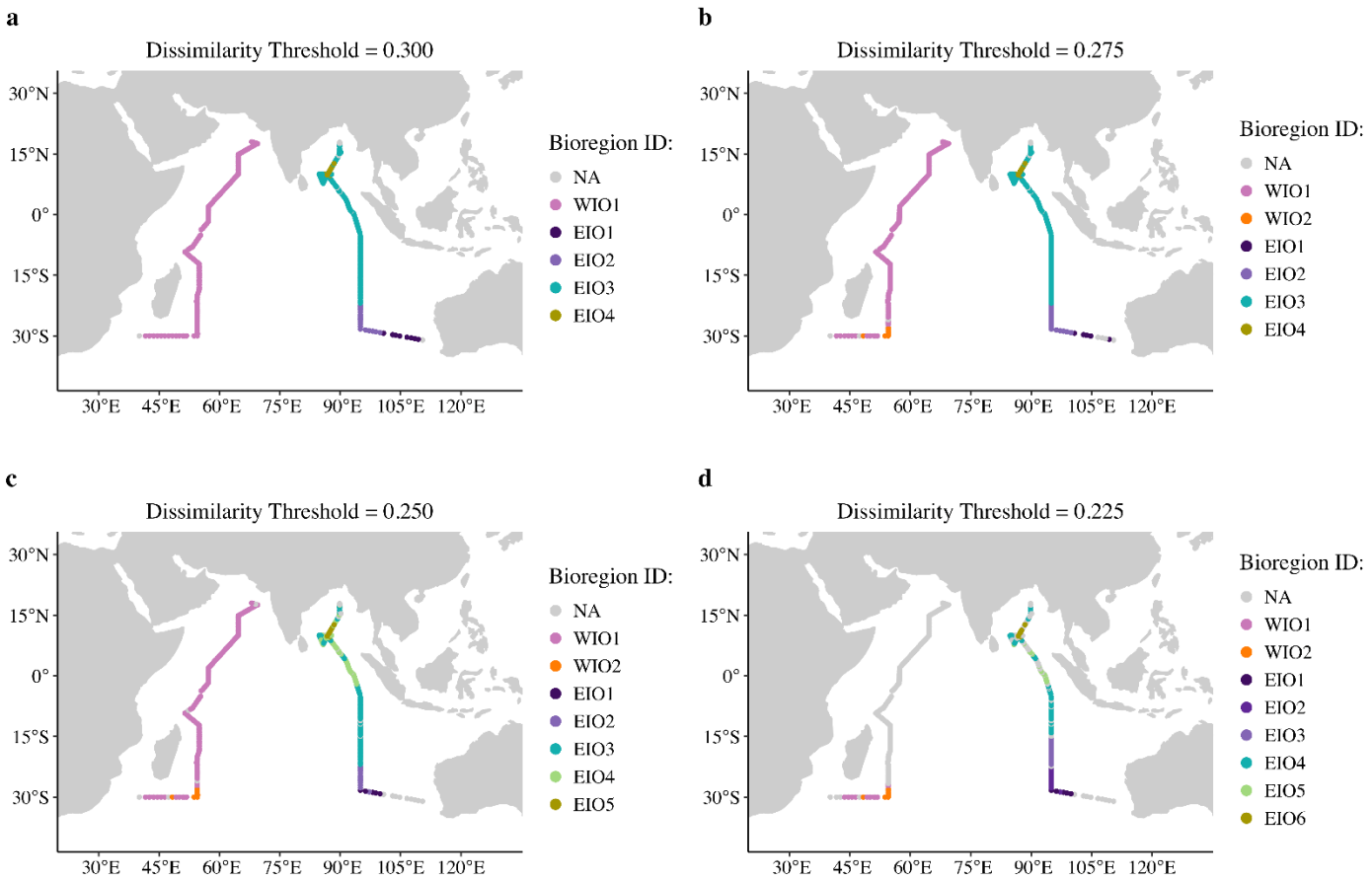
Supplemental Figure 1: Dynamic seascapes of the Indian Ocean. Monthly average distribution of seascape classes from (**a**) April 2016 and (**b**) April 2018. Monthly average relative abundance of seascape classes from (**c**) April 2016 and (**d**) April 2018.



Supplemental Figure 2: Bacterial community structure of the Indian Ocean. (**a**) Longitudinal separation of communities from the western and eastern transects. (**b**) Latitudinal separation of communities along the western transect. (**c**) Latitudinal separation of communities along the eastern transect. Hierarchical clustering was performed on the Bray-Curtis dissimilarity matrix using the average linkage method (UPGMA) and was visualized with dendrograms. The black lines represent the dissimilarity thresholds used to cut the dendrograms and define bioregions. Black labels indicate the clusterwise mean of each cluster.



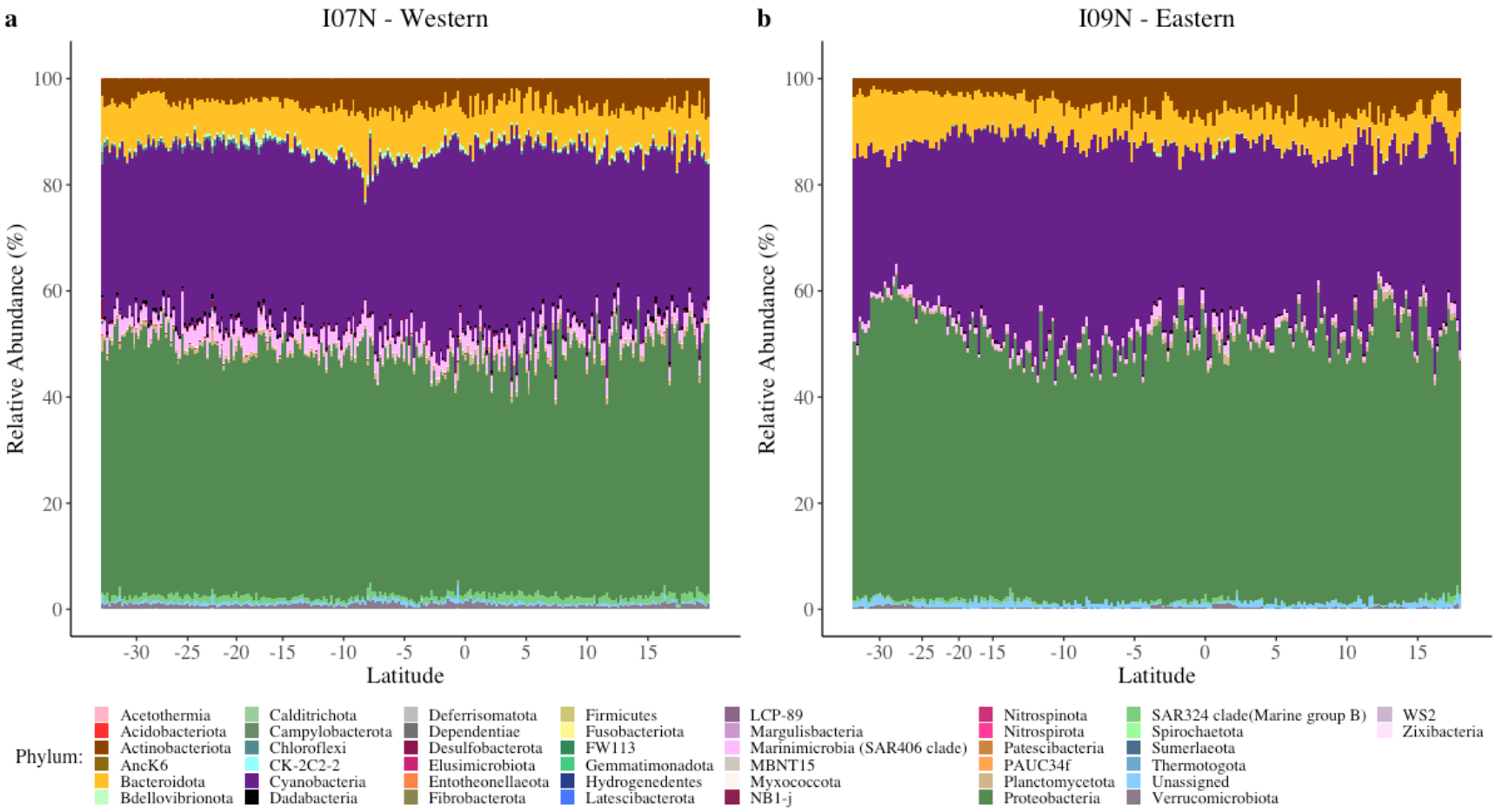
Supplemental Figure 3: Distribution of Bray-Curtis dissimilarities. (**a**) Symmetric distribution of dissimilarities along the western transect. (**b**) Right-skewed distribution of dissimilarities along the eastern transect. Histograms were constructed from the Bray-Curtis dissimilarities in the upper-half of the dissimilarity matrices of each transect.



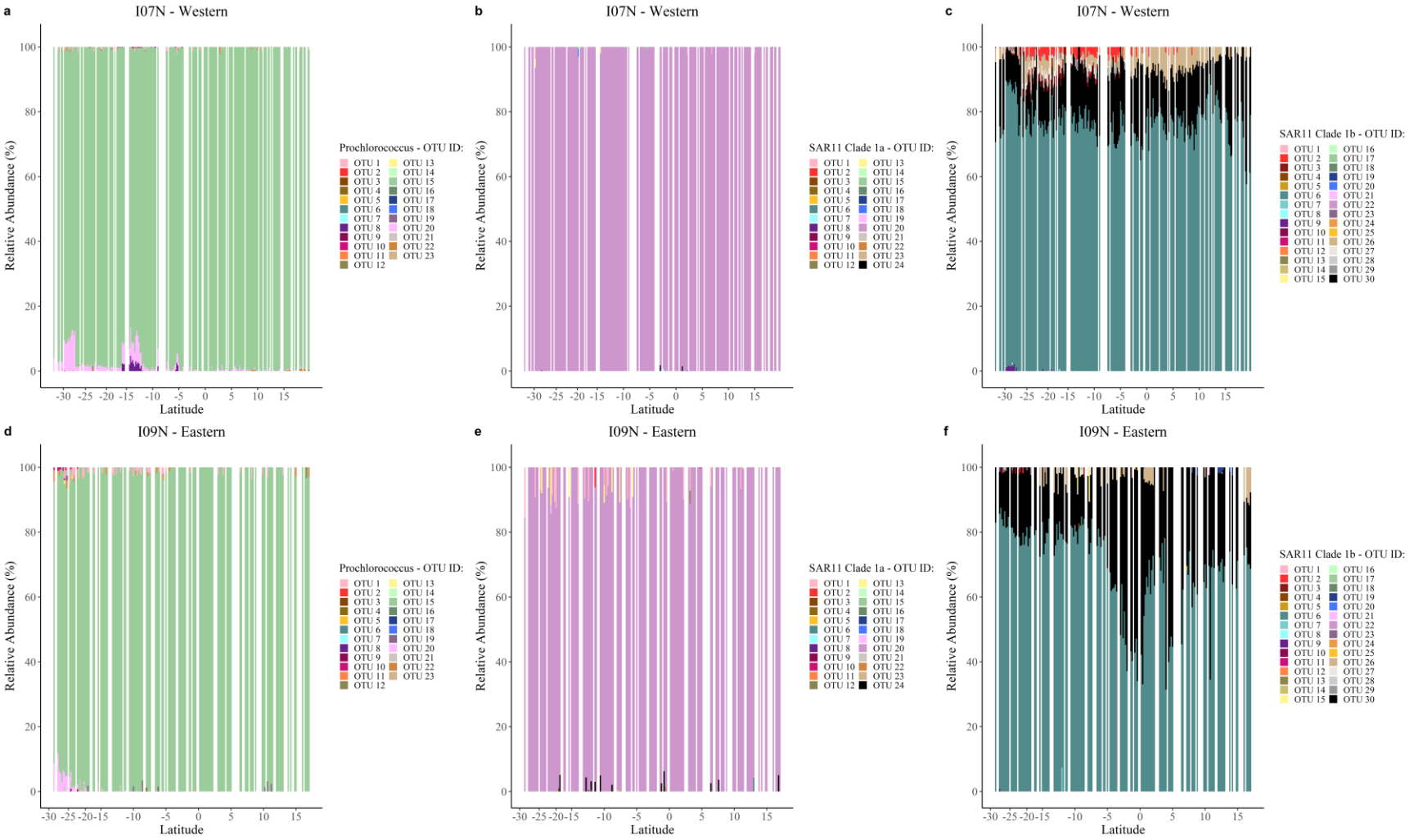
Supplemental Figure 4: Comparison of dissimilarity thresholds. Clustering analysis using different dissimilarity thresholds showed that higher dissimilarity thresholds resulted in unstable clusters or in poor geographic separation of regions. Each color represents a distinct cluster, and grey dots represent samples in unstable clusters or samples that did not cluster. WIO = Western Indian Ocean and EIO = Eastern Indian Ocean.

Supplemental Table 4: Comparison of clustering results using higher dissimilarity thresholds.

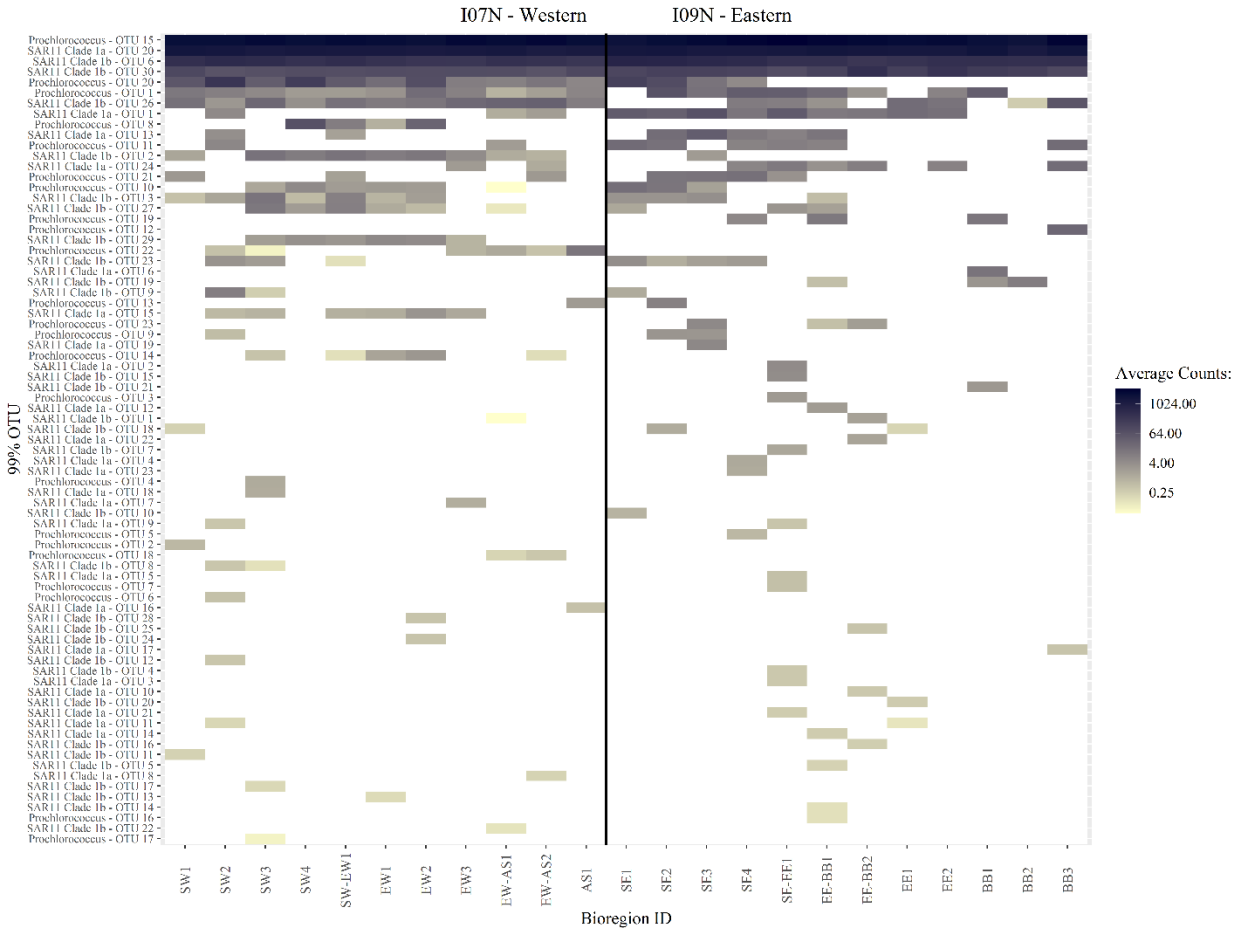
		Dissimilarity Threshold				
		0.300	0.275	0.250	0.225	0.200
I09N	Initial Number of Clusters	8	11	18	30	54
	Number of Clusters with < 5 Samples	3	8	13	22	40
	Number of Unstable Clusters	1	0	0	2	2
	Final Number of Clusters	**4**	**4**	**5**	**6**	**12**
I07N	Initial Number of Clusters	3	6	9	16	25
	Number of Clusters with < 5 Samples	2	4	7	12	18
	Number of Unstable Clusters	0	0	0	2	0
	Final Number of Clusters	**1**	**2**	**2**	**2**	**7**



Supplemental Figure 5: Taxonomic trends. Relative abundances of phyla across the (**a**) western and (**b**) eastern Indian Ocean revealed that Actinobacteria, Bacteroidetes, Cyanobacteria, Marinimicrobia (SAR406 clade), Proteobacteria, and Verrucomicrobia are ubiquitous across the Indian Ocean.



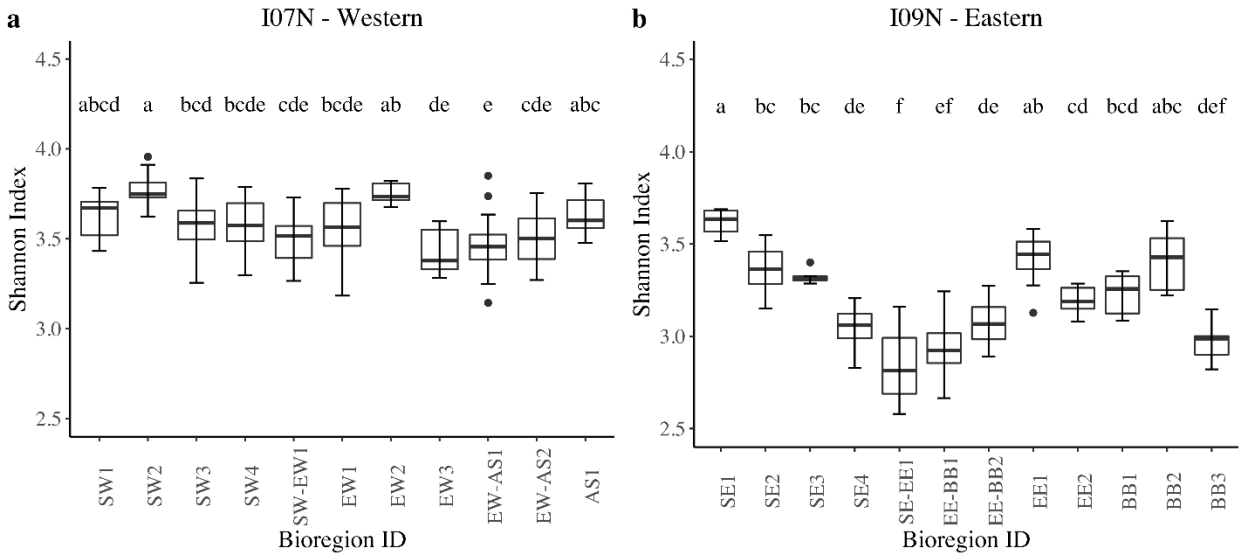
Supplemental Figure 6: Microdiversity trends. Relative abundances of 99% OTUs from the genera (**a** and **d**) *Prochlorococcus*, (**b** and **e**) SAR11 Clade 1a, and (**c** and **f**) SAR11 Clade 1b across samples assigned to bioregions in the western and eastern Indian Ocean revealed that these genera are primarily composed of a single OTU.



Supplemental Figure 7: Heatmap of 99% OTUs according to bioregion. Counts of 99% OTUs from the genera *Prochlorococcus*, SAR11 Clade 1a, and SAR11 Clade 1b showed that bioregions had different microdiverse communities.

Supplemental Table 5: Differential abundance analysis of genera with known geochemical importance.

Genus	Base mean	log2 fold change	adj. p-value
Southwestern Gyre vs. Equatorial Western			
UCYN-A	1.33	5.23	< 0.001
Thiomicrorhabdus	3.04	6.52	0.001
Southwestern Gyre vs. Arabian Sea			
UCYN-A	1.33	5.01	0.002
Thiomicrorhabdus	3.04	6.47	0.002
Sulfitobacter	0.82	-4.75	0.010
Equatorial Western vs. Arabian Sea			
Trichodesmium	3.25	-2.99	0.005
Southeastern Gyre vs. Equatorial Eastern			
Alcanivorax	1.60	-4.32	0.034
Oleibacter	2.27	-4.79	0.010
Southeastern Gyre vs. Bay of Bengal			
Alcanivorax	1.60	-4.75	0.027
Oleibacter	2.27	-5.43	0.005



38

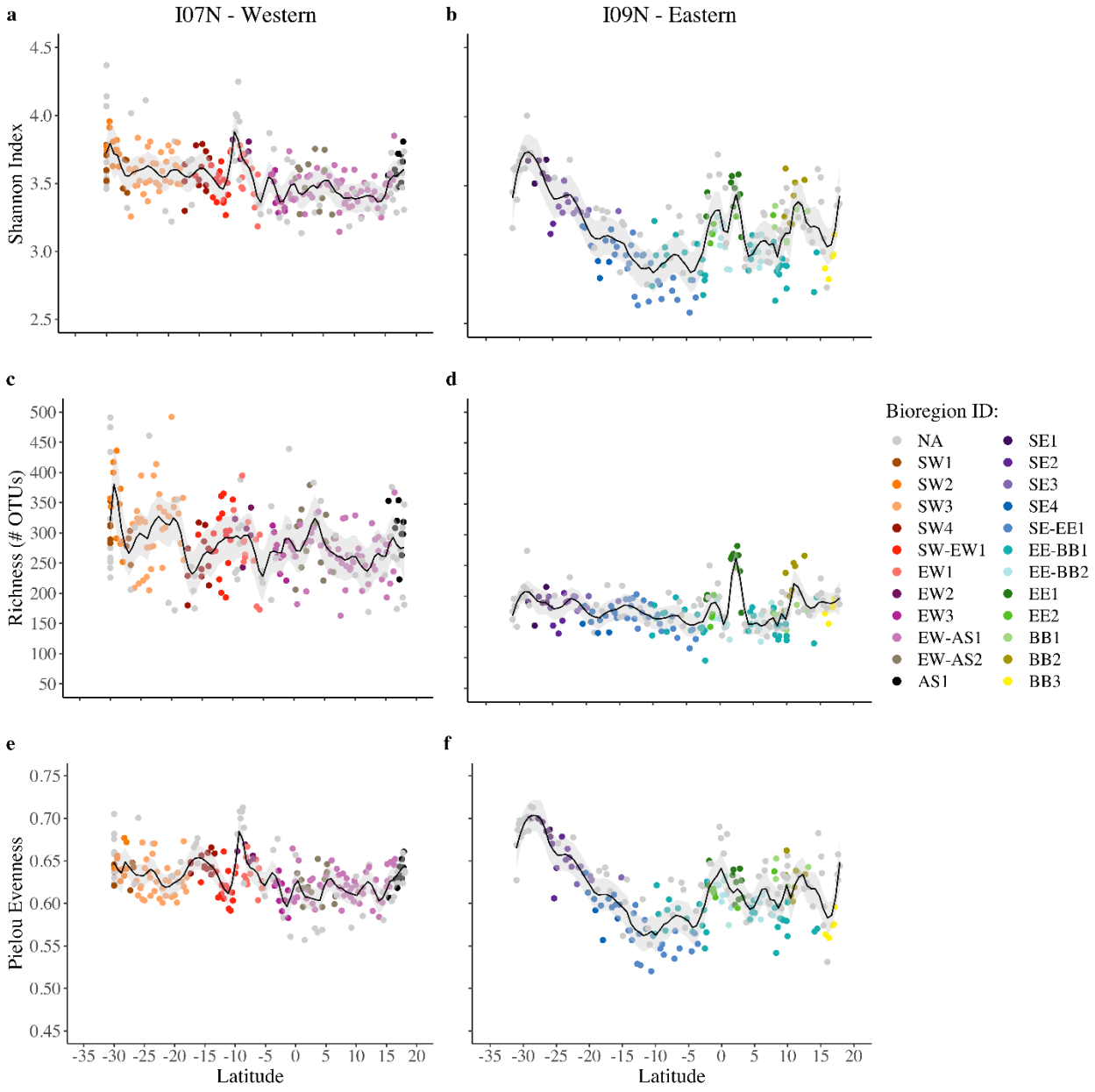
39 **Supplemental Figure 8: Alpha-diversity trends of the bioregions.** Average Shannon Index

40 according to bioregion for the (**a**) western and (**b**) eastern Indian Ocean showed minimal

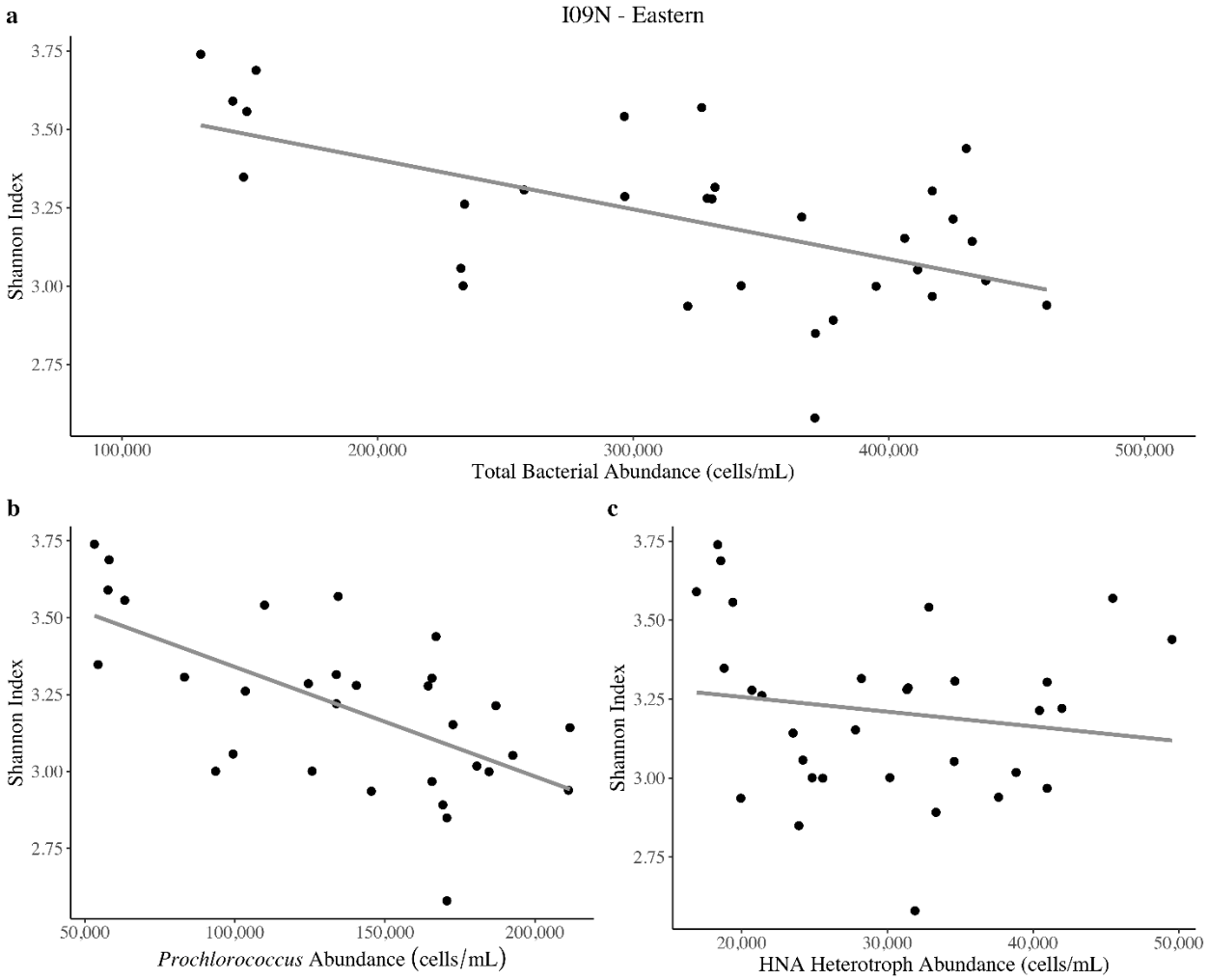
41 variation among the western bioregions and large variation among the eastern bioregions. Letters

42 represent a *post hoc* Tukey test ($p < 0.05$), where means not sharing any letters are significantly

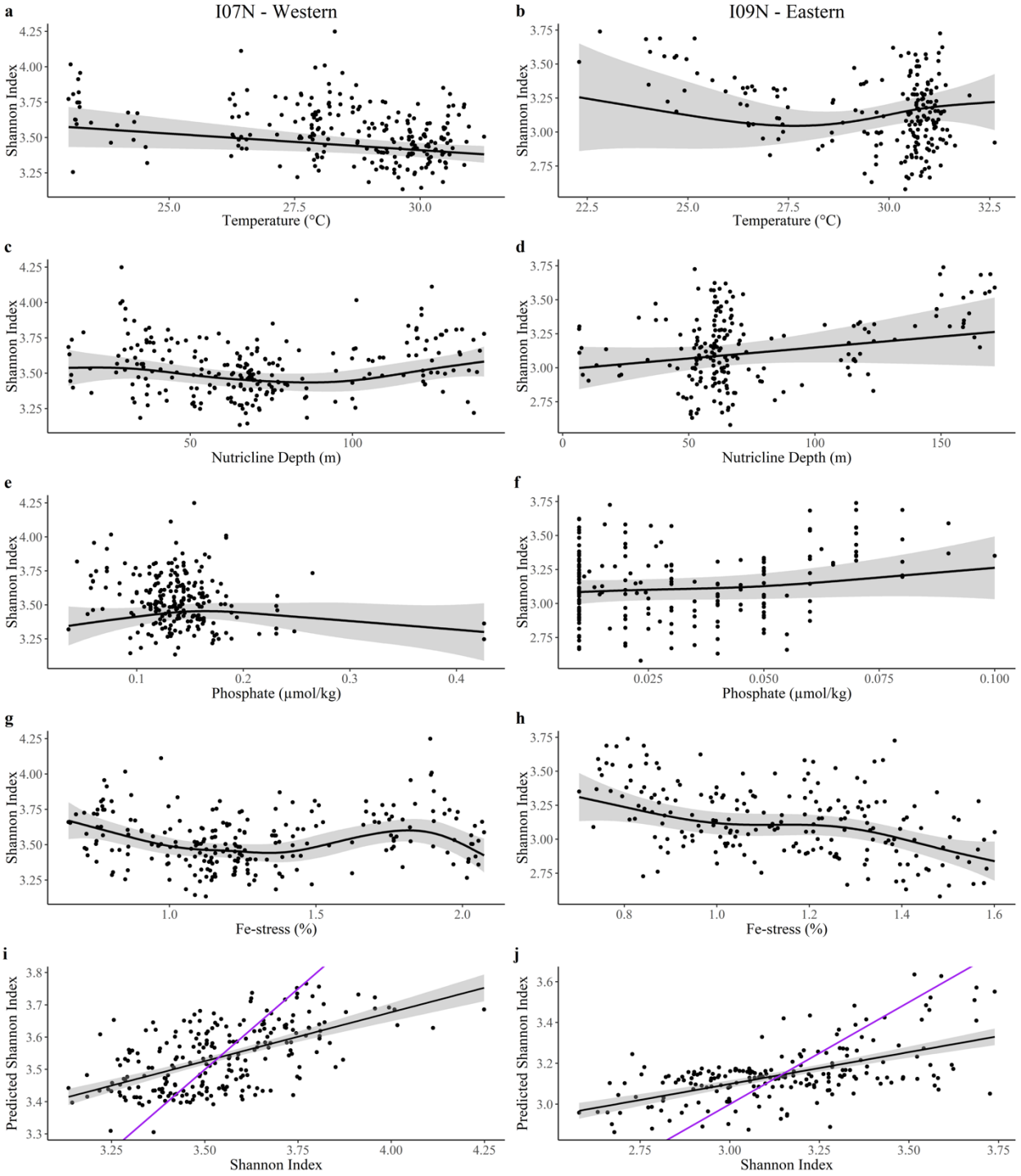
43 different.



Supplemental Figure 9: Trends in richness and evenness across the bioregions. Trends in richness (**c** and **d**) and evenness (**e** and **f**) showed how these two components contributed to changes in alpha-diversity (**a** and **b**) along the transects. Black lines represent smoothing curves fit with a polynomial regression (span = 0.1), and grey area represents the 95% confidence intervals of the smoothed curves.



Supplemental Figure 10: Relationship between alpha-diversity and absolute bacterial abundance. Correlations of Shannon Index with absolute abundances of (**a**) total bacterial abundance (**b**) *Prochlorococcus* abundance and (**c**) HNA heterotroph abundance within the eastern Indian Ocean showed a significant, negative relationship with total bacterial abundance and *Prochlorococcus* abundance and a non-significant relationship with HNA heterotroph abundance.



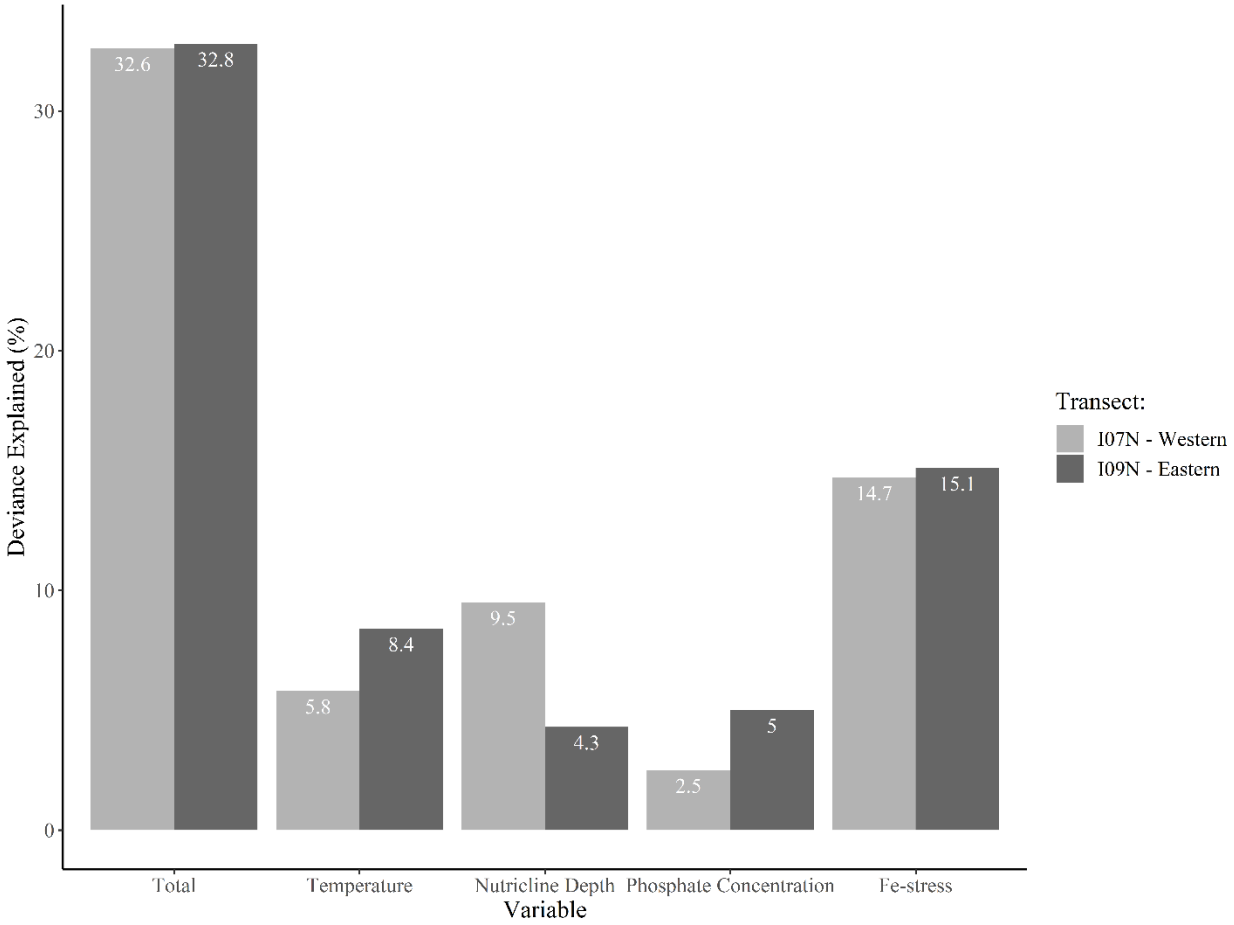
57

58 **Supplemental Figure 11: Best fit general additive models (GAMs) of alpha-diversity.** Total

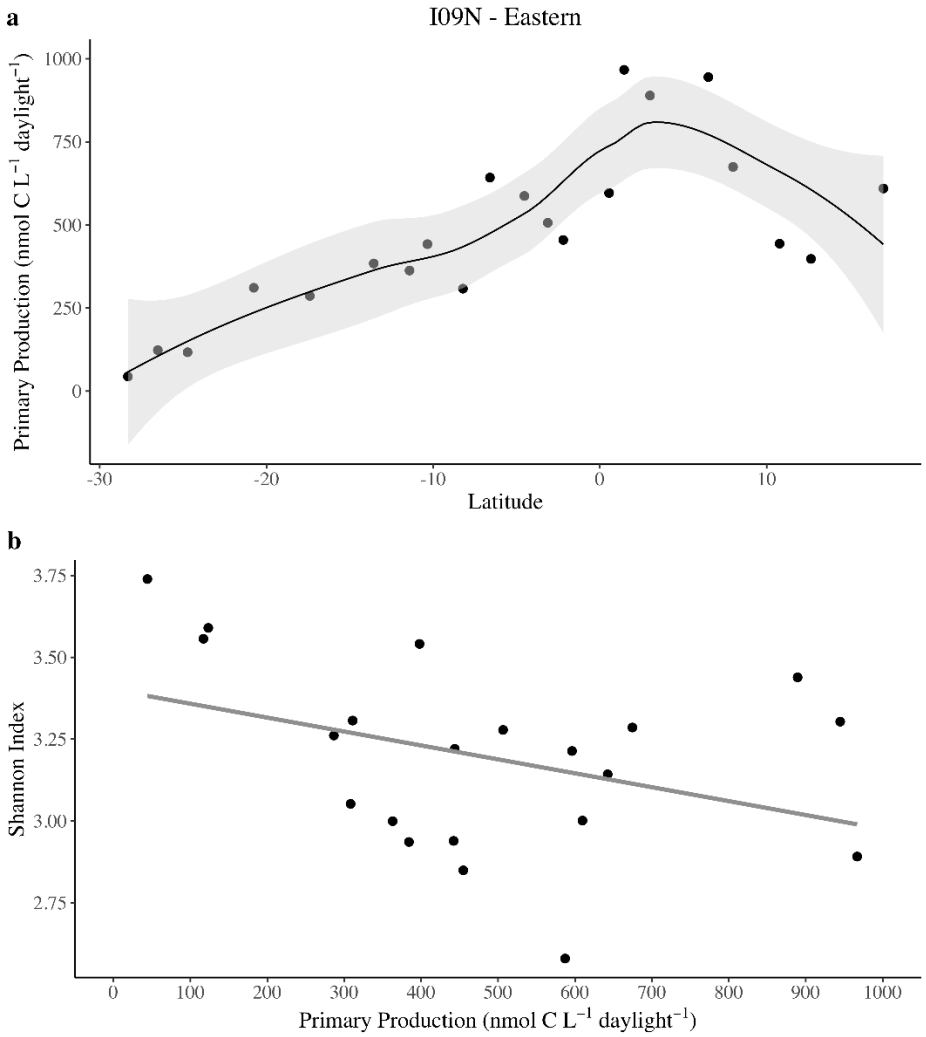
59 deviance explained for I07N was 32.6% ($p < 0.05$, *adj.* $R^2 = 0.291$, $n = 234$) and for I09N was

60 32.8% ($p < 0.05$, *adj.* $R^2 = 0.298$, $n = 199$). GAM-identified relationships between (**a** and **b**)

temperature, (**c** and **d**) nutricline depth, (**e** and **f**) phosphate concentrations, and (**g** and **h**) Fe-stress for alpha-diversity along the western and eastern transects. Lines represent predicted GAM smooths with 95% confidence intervals in grey. (**i** and **j**) Depict the linear relationship between GAM-based predictions and observed alpha-diversity. Black lines represent linear regressions with 95% confidence intervals in grey. Purple lines represent the 1:1 line.



Supplemental Figure 12: Deviance explained by GAMs. Total deviance explained by all environmental factors and deviance explained by each environmental factor from GAM analysis of alpha-diversity for the western and eastern transect.



72

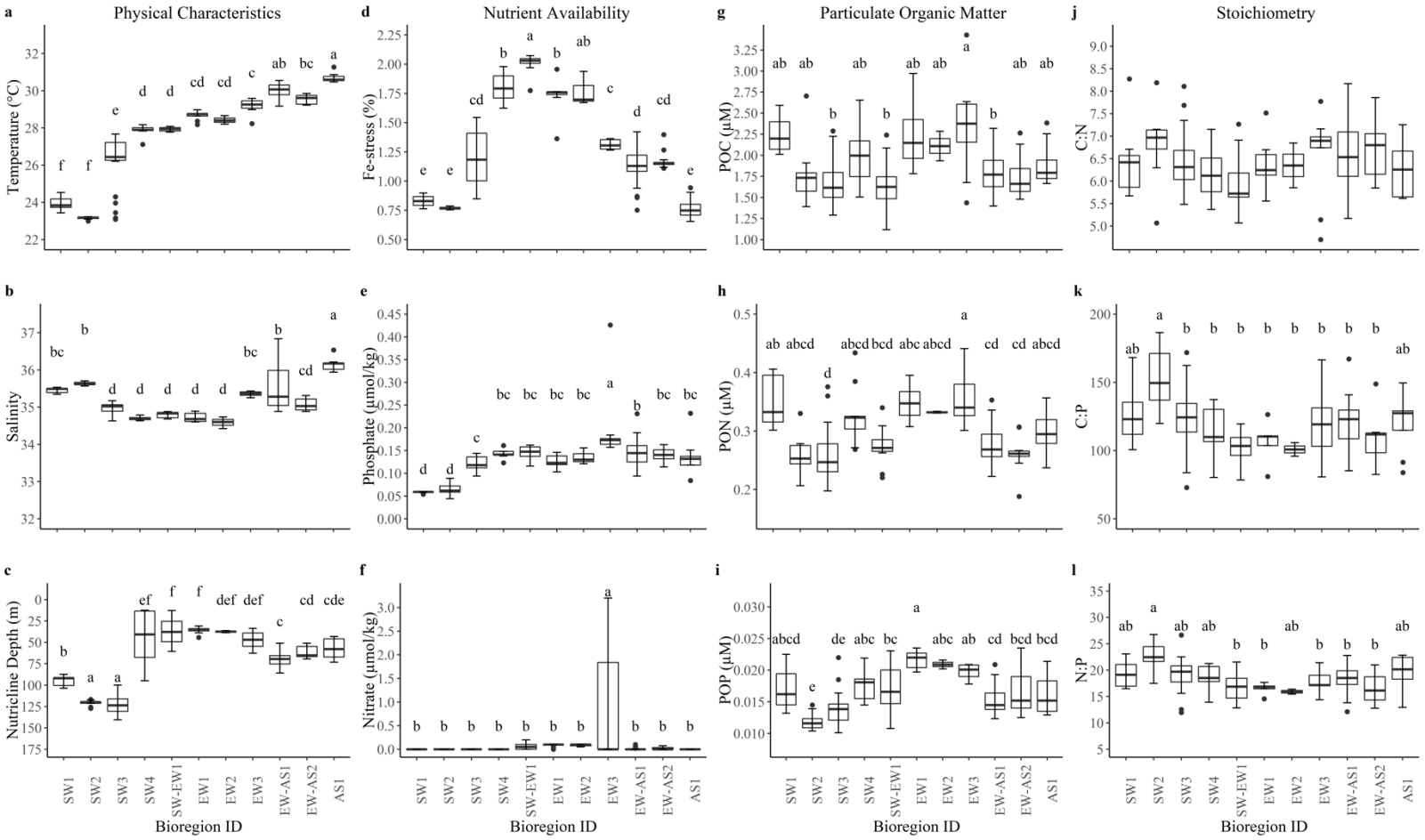
73 **Supplemental Figure 13: Relationship between alpha-diversity and primary production.** (**a**)

74 Primary production varied latitudinally (**b**) Correlation of Shannon Index with primary

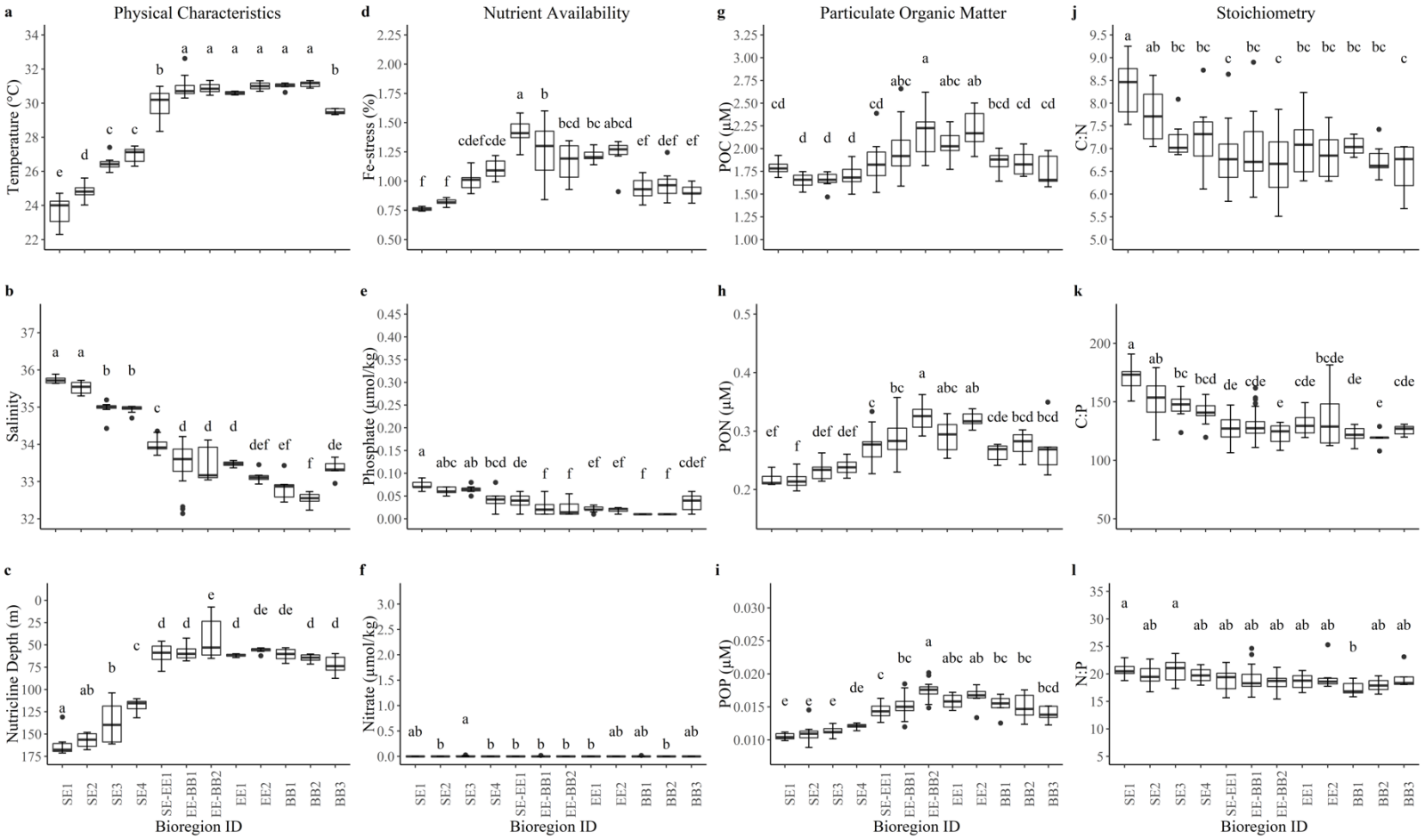
75 production showed a non-significant relationship.

Supplemental Table 6: Correlation of alpha-diversity with ubiquitous and cosmopolitan genera relative abundance. Phyla name is in parenthesis. *****denotes statistical significance (*adj. p* < 0.05)

Genus	I07N - Western		I09N - Eastern	
	adj. p-value	**correlation coefficient**	**adj. p-value**	**correlation coefficient**
(Actinobacteriota) *Candidatus* Actinomarina	0.380	-0.069	0.066	-0.132
(Actinobacteriota) Sva0996 marine group	0.008*	0.193	0.039*	0.148
(Bacteroidota) *Marinoscillum*	0.772	-0.027	0.015*	0.173
(Bacteroidota) NS2b marine group	0.063	0.141	< 0.001*	0.525
(Bacteroidota) NS4 marine group	0.380	0.071	< 0.001*	0.288
(Bacteroidota) NS5 marine group	< 0.001*	0.358	< 0.001*	0.526
(Bacteroidota) *Tenacibaculum*	0.823	-0.02	0.003*	-0.206
(Bdellovibrionota) OM27 clade	0.432	0.059	< 0.001*	0.262
(Cyanobacteria) *Prochlorococcus*	< 0.001*	-0.638	< 0.001*	-0.799
(Cyanobacteria) *Synechococcus*	0.862	-0.011	0.480	0.055
(Myxococcota) P3OB-42	0.102	0.126	0.002*	0.217
(Planctomycetota) *Pirellula*	0.009*	-0.188	0.480	-0.053
(Planctomycetota) Urania-1B-19 marine sediment group	0.837	0.016	< 0.001*	0.265
(Proteobacteria) OM60(NOR5) clade	0.297	-0.083	< 0.001*	0.464
(Proteobacteria) OM75 clade	0.425	0.062	< 0.001*	0.386
(Proteobacteria) *Pseudohongiella*	0.128	0.115	< 0.001*	0.281
(Proteobacteria) SAR 11 Clade Ia	0.049*	-0.149	0.807	-0.017
(Proteobacteria) SAR11 Clade Ib	0.106	0.122	< 0.001*	0.380
(Proteobacteria) SAR92 clade	< 0.001*	0.345	< 0.001*	0.498
(Proteobacteria) *Vibrio*	0.224	0.095	0.554	-0.043
(Verrucomicrobiota) *Lentimonas*	< 0.001*	0.304	< 0.001*	0.525
(Verrucomicrobiota) MB11C04 marine group	0.008*	0.192	< 0.001*	0.648



Supplemental Figure 14: Environmental variability of western Indian Ocean bioregions. (**a**) Temperature (°C). (**b**) Salinity. (**c**) Nutricline depth (m). (**d**) Fe-stress (%). (**e**) Phosphate (μmol/kg) (**f**) Nitrate (μmol/kg). (**g**) Particulate organic carbon (μM). (**h**) Particulate organic nitrogen (μM). (**i**) Particulate organic phosphorus (μM). (**j**) C:N. (**k**) C:P. (**l**) N:P. Letters represent a *post hoc* Tukey test ($p < 0.05$), where means not sharing any letters are significantly different.



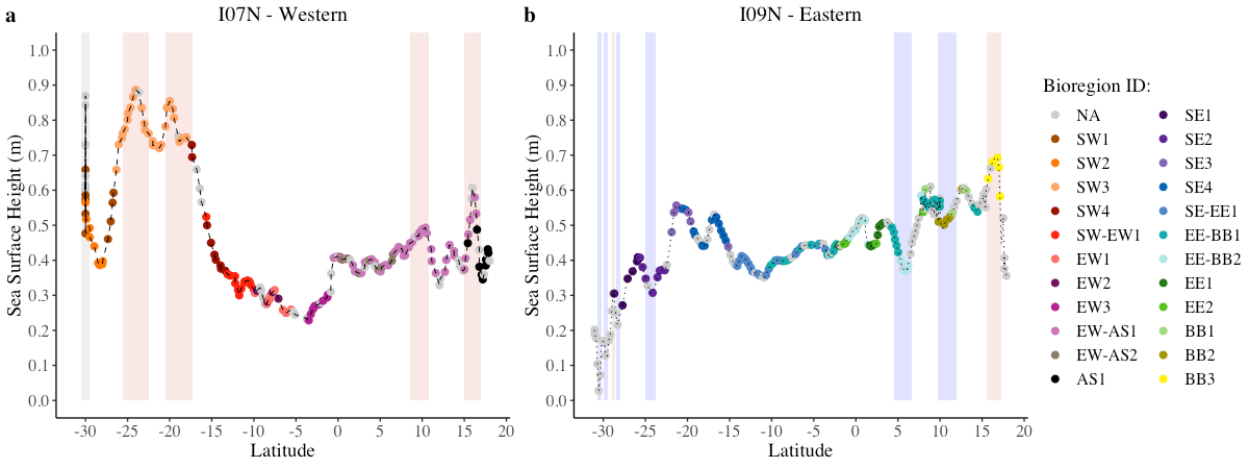
84

85 **Supplemental Figure 15: Environmental variability of eastern Indian Ocean bioregions.** (**a**) Temperature (°C). (**b**) Salinity. (**c**)

86 Nutricline depth (m). (**d**) Fe-stress (%). (**e**) Phosphate (μmol/kg) (**f**) Nitrate (μmol/kg). (**g**) Particulate organic carbon (μM). (**h**)

87 Particulate organic nitrogen (μM). (**i**) Particulate organic phosphorus (μM). (**j**) C:N. (**k**) C:P. (**l**) N:P. Letters represent a *post hoc*

88 Tukey test ($p < 0.05$), where means not sharing any letters are significantly different.

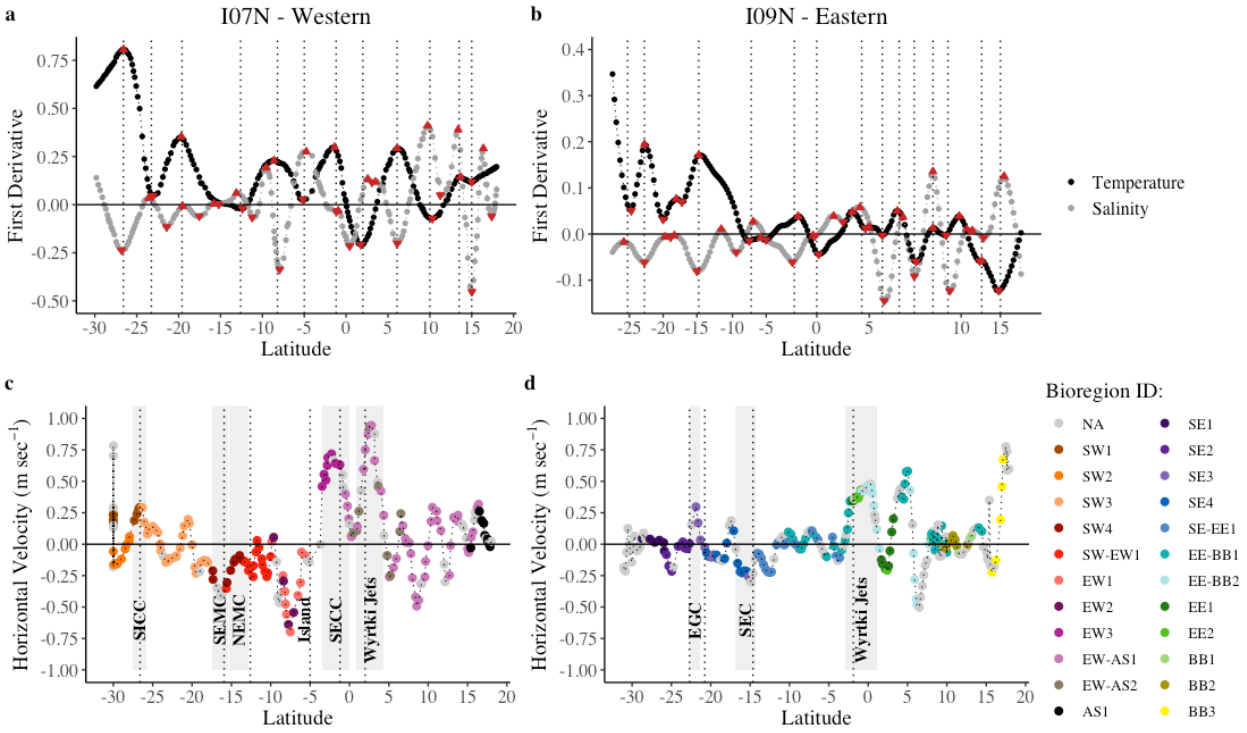


89

90 **Supplemental Figure 16: Location of eddies.** Daily mean sea surface height for the (**a**) western

91 and (**b**) eastern Indian Ocean. Blue, red, and grey shaded areas indicate locations of cold-core

92 eddies, warm-core eddies, and mixture of warm- and cold-core eddies, respectively.



Supplemental Figure 17: Location of fronts and currents. First derivative of temperature and salinity within the (**a**) western and (**b**) eastern Indian Ocean. Locations where a local minimum/maximum for both temperature and salinity occurred indicated a transition between water masses and are denoted with a vertical dashed line. Horizontal velocity within the (**c**) western and (**d**) eastern Indian Ocean. Positive values indicate eastward flowing currents and negative values indicate westward flowing currents. Known currents in the Indian Ocean are shaded grey and labeled: SICC = South Indian Countercurrent, SEMC = South East Madagascar Current, NEMC = North East Madagascar Current, SECC = South Equatorial Countercurrent, EGC = Eastern Gyral Current, SEC = South Equatorial Current. Dashed lines indicate changes in density derived from panels (**a**) and (**b**).