UNIVERSITY OF CALIFORNIA SAN DIEGO

# Marine microbiome structure, diversity, and function within a coastal upwelling region

A Dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy

in

Oceanography

by

Chase C. James

Committee in charge:

Professor Andrew E. Allen, Co-Chair
Professor Andrew D. Barton, Co-Chair
Professor Rachel J. Dutton
Professor Peter J.S. Franks
Professor Brice X. Semmens
Professor Jonathan B. Shurin

2022

The Dissertation of Chase C. James is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

To my parents, Judy & Chet

For their patience and love towards their sons

and instilling within us, the passion and determination to aim for our wildest dreams

whether they be amongst the stars or the seas

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

# LIST OF SUPPLEMENTARY FILES

Supplementary Data 1: Chapter 2 Supplementary Data 1, Differential Abundance Table

Supplementary Data 2: Chapter 2 Supplementary Data 2, Metadata Table

ACKNOWLEDGEMENTS

Below is a list of folks who, without their support, I would most certainly not have made it through the last six years. Know that these words will likely not do justice to the admiration and gratitude that I have for you all.

First, I would like to thank my advisors, Dr. Andrew Barton and Dr. Andrew Allen. Starting this program six years ago I thought that through great determination and perseverance you could get through any difficult task. Unfortunately, I learned this is not always the case and that there are some situations where, despite your best efforts, things will not go as planned. Three years ago, I reach a crossroads in my PhD that left me feeling rather hopeless. Three years ago, you both showed immense empathy and understanding and without hesitation let me join your labs. Words cannot express how thankful I am for the choice you both made. These last three years have been a bit of a whirlwind as I tried to make up for lost time, get acclimated in your labs, and well… do as much science as possible. Thank you for always challenging me to go the extra mile, for supporting my work, and for being an advocate for my career.

Thank you to my committee members: Rachel Dutton, Peter Franks, Brice Semmens, and Jonathan Shurin for pushing me to dig a little deeper, to step back and ask "Why?", and for your overall time and effort. I am so thankful for the interest you all showed in my work and for the kind words after a wave of constructive criticism, even if my exhaustion at times near the end masked my appreciation!

Thank you to my professors, mentors, and co-authors. I feel so fortunate to work in a field where everyone is so passionate about what they do. Thank you for sharing your passion with me and inspiring me to do my best science. Thank you to James Nieh for opening the door to academia and for all the amazing opportunities you provided me during my undergrad and

masters. Thank you to Hao Ye, for always trying to do the right thing and for helping me when I often felt alone. Thank you to Jose Alfredo “Benjamin” Giron-Nava and Andrew Johnson. You two in many ways shaped the scientist I am today, and I will be forever thankful for your mentorship and friendship. Thank you to the Barton and Allen labs, what a difference a community makes. Thank you for including me in your labs and making me feel welcome. From tackling science problems together, to a beer after a long week, you all made the day-in, day-out life of grad school a pleasure.

Thank you to my extend family, I love you all and always look forward to any chance we get to connect, no matter how much time passes in between. Thank you especially to Pete, Kathleen, Cheryl, and Jim, for always looking out for my family and for the love you’ve showed the youngest cousin (me!). Taking a page from a friend’s dissertation I’ll put a disclaimer here, feel absolutely zero need to read this thesis, it is far too long and boring.

Thank you to my friends here at SIO. To Will and Mike, while upon first glance it may appear that you ride alternative surfboard shapes, know that I couldn’t have asked for two better friends that mirror my relationship with Keric. JT said it best, “it’s like you’re my mirror, my mirror staring back at me”. Also, shoutout to your amazing partners Fiona and Allie, both phenomenal humans in their own regard. Thank you to Hugh and Kate for all the laughs and Irish goodbyes (looking at you Hugh). Thank you to my BO cohort, Elliot, Angie, and Rebecca, for being such an amazing collective of human beings and friends. To Allison, Kelly, and Nina (Mr. Chalet!) for many late-night laughs. To the Semmens crew, Erica, Jordan, and Kayla, thanks for all the countless hours chatting about science and life and for always making me feel like I had a second home at Hubbs when things got rough. Kayla, thank you for being a pillar of support through my PhD. I loved our science discussions that always bled into life discussions, and I can’t wait for

the many more conversations we'll have throughout our lives. Thank you to Trevor, I will always feel so lucky to have so many shared experiences with you, from week one of undergrad, to being in your wedding, to both of us defending at SIO, what a journey and what a phenomenal friend. There are so many more folks I wish I could name here but I would likely end up with an acknowledgements page longer than my thesis and so I will just say a HUGE thank you to the SIO community in general. You all are what makes this place magical, and I can't thank each of you enough for all the thoughtful conversations, laughs, and beers we shared, thank you.

Thank you to my immediate family. I dedicated this thesis to my folks, Chet and Judy. I feel so lucky that I get to call my parents my heroes. This last month has been a whirlwind of life and it's such as joy I've gotten to share those moments with you both. I will never tire of people coming up to me and telling me how amazing and wonderful you two are, no matter the occasion. Your love, patience, and self-sacrifice for Karsten and me is something I will never take for granted. Thanks for teaching us to fight for what is right and for the determination to achieve our goals no matter what challenges lay ahead. I love you both with all my heart. Thank you also to Karsten, as my older brother you always showed me what was possible when you poured your heart and soul into the things you loved. While it was often my role as younger brother to zig where you zagged, I know that I would not be where I am today without you setting an example for how to go after your dreams. Let these sentences be the literary equivalent of a big Chase hug, love you!

To the Jordans, thank you for three decades of memories, I cherish every moment we get with each other. Thanks also for keeping an eye on Chet and Judy. Thank you to my dance friends and dance teammates for many late nights, brews, and swing outs. Thank you to all the rock climbing friends I've made in the past few years and the adventures we've had, both on real

rocks and plastic holds. Thank you to my extended, extended families, the Moore's and Sykes' for always welcoming me into your homes and for all the great meals and memories.

To Bonnie, Keric, and Will, thank you for years of moments big and small. To life, shared and enjoyed over a glass of wine and a home cooked meal. To accusations of betrayal shouted over a small landscape of towns and roads. To laughing so hard that we can't even look at one another without breaking into laughter again. Thank you for making the best moments of my life shine even brighter and for helping carry the load when life gets hard. You guys really are my family, and I love you all so much. This year represents an end of an era, but I know that what lies ahead will be phenomenal so long as you are all there alongside me. To Travis, though we never lived together, I feel so lucky to call you my friend. Whew! Did we butt heads at first or what! Thank you for your unbridled passion and willingness to sign up for any adventure and for always being there. You represent what true commitment towards a friendship means.

To Kayla. I could not have imagined that in less than two years my life would change so completely. Every day I feel so lucky that I get to experience life with you. Thank you for making me feel loved for who I am. For letting me be the nerdy, whale-fact spouting, Emperor's New Groove-quoting person that, after two hours of nervous blubbering, finally got up the courage to kiss you that night looking out over the ocean. You inspire me every day to do that little bit more for someone, to speak up when it's easier to stay quiet, and to love without restraint. After six years at SIO and twelve years in San Diego, change is scary, but it's also exciting. I can't wait to take this next step with you (and the future pet that has been contractually promised, ha!). I love you with all my heart and look forward to all that life brings in these upcoming years. Also, to Patty and Bernie, for raising this phenomenal human being and for welcoming me into your home and treating me like family, I can't thank you enough.

Thank you again to all who are mentioned and for the countless others that were not mentioned but for whom I am incredibility thankful of. I'll end this section with a quote that summarizes my experience in my PhD and captures the extend of my nerdiness. Thank you all. "*Some people charged toward the goal, running for all they had. Others stumbled. But it wasn't the speed that mattered. It was the direction they were going*" – Sanderson.

Chapter 2, in full, is a reprint of the material as it appears in James, C. C., Barton, A. D., Allen, L. Z., Lampe, R. H., Rabines, A., Schulberg, A., Zheng, H., Goericke, R., Goodwin, K. D., Allen, A. E. (2022). Influence of nutrient supply on plankton microbiome biodiversity and distribution in a coastal upwelling region. Nature Communications. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in part is currently being prepared for submission for publication of the material. James, C.C., Allen, A. E., Lampe, R. H., Barton, A. D. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in part is currently being prepared for submission for publication of the material. James, C. C., Barton, A. D., Allen, L. Z., Smith, S. R., Venepally, P., Lampe, R. H., Rabines, A., Schulberg, A., Zheng, H., Goericke, R., Goodwin, K. D., Allen, A. E. The dissertation author was the primary investigator and author of this paper.

# VITA

2014 Bachelor of Science in Biology, University of California San Diego

2016 Master of Science in Biology, University of California San Diego

2022 Doctor of Philosophy in Oceanography, University of California San Diego

## PUBLICATIONS

**James CC**, Barton AD, Zeigler LA, Lampe RH, Rabines A, Schulberg A, Zheng H, Goericke R, Goodwin KD, Allen AE. Influences of nutrient supply on plankton microbiome biodiversity and distributions in a coastal upwelling region. Nature Communications. 13, 2448 (2022).

**James CC**, Sánchez D, Cruz-López L, Nieh J. Fighting ability and the toxicity of raiding pheromone in an obligate kleptoparasite, the stingless bee Lestrimelitta niitkib. Behavioral Ecology and Sociobiology. 2022 Mar;76(3):1-2.

Agarwal V, **James CC**, Widdicombe CE, Barton AD. Intraseasonal predictability of natural phytoplankton population dynamics. Ecology and evolution. 2021 Nov;11(22):15720-39.

Giron-Nava A, Munch SB, Johnson AF, Deyle E, **James CC**, Saberski E, Pao GM, Aburto-Oropeza O, Sugihara G. Circularity in fisheries data weakens real world prediction. Scientific reports. 2020 Apr 24;10(1):1-6.

Lee, SW, Dong KY, **James CC**, Lee S, Koh HY, Sheen YH, Oh JW, Han MY, Sugihara G. Short-term effects of multiple outdoor environmental factors on risk of asthma exacerbations: Age-stratified time-series analysis. Journal of Allergy and Clinical Immunology 144, no. 6 (2019): 1542-1550.

Sugihara G, Criddle KR, McQuown M, Giron-Nava A, Deyle E, **James CC**, Lee A, Pao GM, Saberski E, and Ye H. "Comprehensive incentives for reducing Chinook salmon bycatch in the Bering Sea walleye Pollock fishery: Individual tradable encounter credits." Regional Studies in Marine Science 22 (2018): 70-81.

Cohen RE, **James CC**, Lee A, Martinelli MM, Muraoka WT, Ortega M, Sadowski R, Starkey L, Szesciorka AR, Timko SE, Weiss EL, Franks PJ. Marine host-pathogen dynamics: influences of global climate change. Oceanography. 2018 Jun 1;31(2):182-93.

Giron-Nava A, **James CC**, Johnson AF, Dannecker D, Kolody B, Lee A, Nagarkar M, Pao GM, Ye H, Johns DG, Sugihara G. "Quantitative argument for long-term ecological monitoring." Marine Ecology Progress Series 572 (2017): 269-274.

ABSTRACT OF THE DISSERTATION

Marine microbiome structure, diversity, and function within a coastal upwelling region

by

Chase C. James

Doctor of Philosophy in Oceanography

University of California San Diego, 2022

Andrew E. Allen, Co-Chair
Andrew D. Barton, Co-Chair

In the pelagic environment, microbes act as the base of the food web (photosynthetic autotrophs), recycle nutrients (microbial loop), and perform other crucial ecosystem processes and services (such as carbon sequestration). The relative scale of these different process is driven by changes in marine microbiome community structure, diversity, and function. Over the last two decades, meta-omic sampling has provided a pathway forward with which to observe the

community structure and function of the marine microbiome at a previously inaccessible resolution. However, with this increase in data complexity (large numbers of identified species and genes), it can be challenging to synthesize results across the multitude of observed taxonomic and functional groups. The goal of this thesis is to provide a general framework for understanding marine microbiome community responses (structure, diversity, and function) to environmental perturbations at previously unresolvable scales.

The first study (Chapter 2) identifies the mechanisms that shape patterns in marine microbiome community structure and diversity across space and time within a coastal upwelling region. While traditional methods (such as microscopy and flow cytometry) have highlighted general patterns for broad taxonomic groups and or conspicuous taxa, this study represents a comprehensive examination of the mechanisms that shape all types of marine microbial groups, and in particular, highlights cryptic groups that could not be identified through more traditional means.

The second study (Chapter 3) takes a more species-centric approach and asks, what is the rate of habitat specificity within marine microbes? Terrestrial systems often contain many species that are endemic to habitats or locales. Within the marine environment, habitats are constantly in motion, moving dynamically across space in time. The dynamic marine environment, coupled with the fast generation times of most microbes is thought by many to lead to less habitat specificity and more cosmopolitan (universally distributed) species. By identifying water mases (with internally consistent physical and chemical environments) we present a view of habitat specificity within the marine microbiome in a way that is comparable to terrestrial studies.

The third study (Chapter 4) shifts to look at regional metatranscriptomic data and asks what are the mechanisms that shape the function and distribution of active marine microbes. Metatranscriptomics provides a framework to identify which taxa and their associated functions are active within a community in response to changing environmental conditions. In targeting the active community, we identify how environmental conditions can lead to in-situ functional traits within the microbial community—a crucial next step to better understanding the links between environmental conditions and the local to global magnitude of key ecological functions such as primary productivity, nutrient recycling, and carbon sequestration in the pelagic ocean.

# Chapter 1 - Introduction

Marine primary productivity accounts from roughly half of global productivity and has been responsible for massive shifts in the geochemistry of the Earth for over 3 billion years (Field et al. 1998; Falkowski, Barber, and Smetacek 1998; Falkowski, Fenchel, and Delong 2008). Within the pelagic ocean, these effects are driven by both photoautotrophs and their associated heterotrophic bacterial and archaeal communities (Azam et al. 1983; Fenchel 2008; Not et al. 2012). Particular groups, like diatoms are known to be key players in both the local and global magnitude of environmental processes like primary production and carbon sequestration (Goldman 1993; Sommer et al. 2002; Taylor et al. 2015; Abrantes et al. 2016). While groups like diatoms are both notable and clearly important in determining the magnitude of numerous ecological services, it is safe to say the biogeographies and functional traits of most marine microbes remain poorly understood.

Until recently, methods of sampling the marine microbiome have been at a relatively coarse resolution. Traditional light microscopy relies on identifying physiological differences between cells which, while possible for conspicuous species can be challenging for the majority of single celled prokaryotes and eukaryotes. Other methods, such as chlorophyll-α measurements provide a broad image of the photoautotrophic community but do little to parse out variations in community structure and diversity (Foukal and Thomas 2014; Taylor et al. 2015). Meta-omic sampling of the marine environment has provided a revolutionary new lens with which to assess the immense diversity and dynamic ecological community structure and function present within the marine microbiome (Rusch et al. 2007; J. A. Fuhrman et al. 2008; Sunagawa et al. 2015; de Vargas et al. 2015; Needham and Fuhrman 2016; Villarino et al. 2018; Kolody et al. 2019).

Major marine environmental meta-omics studies up to this point have fallen into one of two sampling regimes: 1) global surveys (Rusch et al. 2007; de Vargas et al. 2015; Sunagawa et al. 2015; Villarino et al. 2018) or 2) local (single station) time series (Needham and Fuhrman 2016; Gilbert et al. 2012). These studies have already provided an unprecedented step forward with which to observe the local-to-global patterns and processes that shape marine microbiomes (Jed A. Fuhrman, Cram, and Needham 2015; Ibarbalz et al. 2019), however, due to their focus on the extremes of sampling (a single global snapshot or a continuous time series in one location), they are restricted in their ecological interpretations.

Within the marine environment, observed patterns and processes are often the result of combined spatio-temporal processes. Water masses, which have conserved properties such as temperature and salinity, and other properties such as nutrient concentrations which can be internally variable but relatively different between water masses, represent the available habitat of pelagic microbiomes (D'Ovidio et al. 2010; Bowman et al. 2018; Bograd, Schroeder, and Jacox 2019). However, unlike terrestrial systems, these habitats are in constant motion, shifting across space and time. Time series collected at a single point may appear to highlight a succession between different communities that, when viewed in a regional context may be the result of multiple distinct water parcels passing through a static point. Similarly, a global snapshot of the marine microbiome may capture one pattern, which can vary significantly given seasonal or interannual environmental processes (Haury, McGowan, and Wiebe 1978). For marine meta-omics sampling to take the next step, collection of samples must occur across suitable spatial and temporal scales to reasonably capture the processes that shape the community structure and function of the pelagic microbiome.

The California Cooperative Oceanic Fisheries Investigations (CalCOFI) represents one of the longest and most comprehensive long-term ecological monitoring programs within the marine environment (Mantyla, Venrick, and Hayward 1995; Hayward and Venrick 1998; Bograd and Lynn 2003; Hsieh et al. 2005). Over 70 years of sampling, the CalCOFI program has constantly evolved to generate the best available science towards improving our understanding of marine systems (Taylor et al. 2015; Powell and Ohman 2015; Rudnick et al. 2017). In 2014, the NOAA CalCOFI Ocean Genomics (NCOG) project began collecting quarterly metabarcoding and metatranscriptomic samples from across the region (winter, spring, summer, and fall). The data analyzed in this thesis represents seven years of NCOG sampling (2014-2020) across the highly variable region. Samples are collected from San Diego to Point Conception and from nearshore (0-10km) to offshore (roughly 400km) stations. This region includes wide environmental gradients ranging from meso/eutrophic conditions in the nearshore as the result of coastal upwelling (Checkley and Barth 2009), to oligotrophic conditions similar to the North Pacific subtropical gyre (Bograd, Schroeder, and Jacox 2019). Conditions within the region also vary inter-annually as the result of El Niño/La Niña cycles and other regional anomalies (Bograd and Lynn 2003; Kim et al. 2009; Zaba and Rudnick 2016; Lilly and Ohman 2018). Combined, this study represents a major milestone in marine environmental genomic sampling, combining both spatial and temporal sampling at an unprecedented scale.

Chapters 2, 3, and 4 of this thesis show analyses of meta-omics samples across space and time to identify connections between environmental conditions and community structure, diversity, and function within the marine microbiome at a previously inaccessible resolution. Chapter 2 captures how regional conditions drive spatial and temporal patterns in community structure and diversity within the marine microbiome, identifying overall microbiome patterns as

well as relationships across various taxonomic groups, many of which would be impossible to assess through traditional means. Chapter 3 explores regional rates of endemism, cosmopolitanism, and habitat affinity within the region. Water masses, representative of marine microbial habitats, are identified in a way that is comparable to terrestrial systems and used to ask whether the high dispersal potential within the marine environment truly leads to cosmopolitan microbial distributions. Finally, Chapter 4 examines the regional metatranscriptome and asks: 1) which environmental gradients lead to the greatest niche partitioning amongst active microbial members within the region and 2) how does the functional composition of the microbial community change as a result of environmental conditions and community structure. Combined, these questions aim to identify the relationship between environmental conditions and the resulting ecological function within the marine microbiome—a crucial step towards understanding the processes that shape local to global magnitude of microbiome functions like primary productivity, nutrient recycling, and carbon sequestration. Overall, the data collected from NCOG, and the resulting analyses presented in these three chapters represent the potential for high-resolution ecological insight when meta-omics data is collected at appropriate spatial and temporal scales for the system in question, in this case the marine environment.

# Chapter 2 - Influence of nutrient supply on plankton microbiome biodiversity and distribution in a coastal upwelling region


## Abstract

The ecological and oceanographic processes that drive the response of pelagic ocean microbiomes to environmental changes remain poorly understood, particularly in coastal upwelling ecosystems. Here we show that seasonal and interannual variability in coastal upwelling predicts pelagic ocean microbiome diversity and community structure in the Southern California Current region. Ribosomal RNA gene sequencing, targeting prokaryotic and eukaryotic microbes, from samples collected seasonally during 2014-2020 indicate that nitracline depth is the most robust predictor of spatial microbial community structure and biodiversity in this region. Striking ecological changes occurred due to the transition from a warm anomaly during 2014-2016, characterized by intense stratification, to cooler conditions in 2017-2018, representative of more typical upwelling conditions, with photosynthetic eukaryotes, especially diatoms, changing most strongly. The regional slope of nitracline depth exerts strong control on the relative proportion of highly diverse offshore communities and low biodiversity, but highly productive nearshore communities.


## 2.1 Introduction

Coastal regions disproportionally contribute to marine global primary productivity and are thus important both ecologically and economically (Ryther 1969a). The Southern California Current (SCC) region encompasses spatial and temporal gradients ranging from the eutrophic nearshore to the oligotrophic offshore and provides ideal conditions for quantifying variation in microbial community structure and biodiversity in response to dynamics associated with physical, chemical, and biological gradients.

Spatial patterns in marine microbial communities are strongly shaped by dispersal, environmental selection (Follows et al. 2007; Edwards, Litchman, and Klausmeier 2013; Nemergut et al. 2013; Villarino et al. 2018), and, on longer timescales, evolution (Collins, Rost, and Rynearson 2014). Global-scale surveys, such as Tara Oceans and Malaspina (Rusch et al. 2007; de Vargas et al. 2015; Sunagawa et al. 2015; Villarino et al. 2018) suggest that temperature gradients most strongly shape marine microbial community structure and biodiversity (J. A. Fuhrman et al. 2008; Sunagawa et al. 2015; Righetti et al. 2019). Other environmental conditions, such as nutrient and light availability can also provide strong bottom-up constraints in plankton communities (Cermeño et al. 2008; Barton et al. 2010) and are particularly important along coastal boundaries (Ryther 1969b). Within the SCC, coastal upwelling creates strong spatial gradients in temperature, nutrients, and light (Mantyla, Venrick, and Hayward 1995; Hayward and Venrick 1998) (Fig. 2.1S). Previous studies have shown that phytoplankton and zooplankton communities vary along these gradients (Venrick 2009; Powell and Ohman 2015b; Taylor et al. 2015). Furthermore, changes in seasonal nearshore upwelling are thought to drive distinct differences in phytoplankton and zooplankton assemblages across the region with variation occurring on seasonal, interannual (El Niño/La Niña), and multidecadal (Pacific Decadal Oscillation) time frames (Catlett et al. 2021; Lilly and Ohman 2018a). Within the microbial community however, the bulk of knowledge exists at a broad level of taxonomic and or functional groups, masking the effects of environmental perturbation within these broad groups and completely missing "cryptic" groups that cannot be identified with more traditional methods (such as bacterial and archaeal groups).

Metabarcoding and metagenomic datasets provide a crucial next step with which to explore the patterns and processes of marine microbial communities at a far higher resolution

and in doing so, illuminate the key processes that structure the base of the marine food web. However, our current understanding of the high taxonomic resolution spatial patterns in microbial community structure and biodiversity are limited by the spatial and or temporal scale of sampling. Studies often focus on changes across space or time but rarely both (Richardson et al. 2006; Z. Wang et al. 2019; Y. Wang et al. 2020). Global datasets of marine microbiome data capture spatially extensive physical and ecological domains (de Vargas et al. 2015; Ibarbalz et al. 2019; Villarino et al. 2018) and can identify the large environmental gradients such as temperature that appear to shape communities across large ocean basins. In contrast, investigations conducted at singular stations identify changes in the marine microbiome through time (Jed A. Fuhrman, Cram, and Needham 2015; Gilbert et al. 2012; Karl and Lukas 1996; Steinberg et al. 2015), exploring questions such as how succession within one group (such as phytoplankton) can drive changes in the overall community composition (Needham and Fuhrman 2016). However, the biotic and abiotic mechanisms that shape biodiversity and community composition patterns often remain uncertain (Nemergut et al. 2013). Combined spatial and temporal metagenomic and metabarcoding sampling of marine microbial communities is necessary to illuminate the gaps in spatially or temporally explicit microbiome studies, such as whether trends happening in one location occur elsewhere or whether observed spatial patterns are conserved or vary across time.

Here we leverage 995 microbial community composition observations from quarterly CalCOFI surveys from 2014-2020, hereafter referred to as the NOAA CalCOFI Ocean Genomics (NCOG) data. The CalCOFI surveys spans from highly productive coastal upwelling waters to oligotrophic offshore waters with NCOG sampling at both the surface and deep chlorophyll maximum (DCM, Fig. 2.1). With these data, we identify spatial patterns in community structure

and biodiversity and highlight the environmental factors that correlate with these ecological parameters. Next, we explore how biodiversity and community structure responded to the 2014-2016 warm anomaly period, followed by the return of cooler conditions in 2017-2018. Ecological changes as a result of this shift included harmful algal blooms (Zhu et al. 2017), possible poleward displacements of planktonic organisms (Mcclatchie et al. 2016), and the occurrence of novel fish species (Walker Jr et al. 2020). Within the SCC, it has been shown that mesoplankton communities tend to recover from other warming events (El Niño) within one year (Lilly and Ohman 2018b). However, beyond trends in total chlorophyll (Kahru, Jacox, and Ohman 2018), little is known about the response of microbial communities to the warm events in 2014-2016. Conditions were also distinct in 2019-2020 when the region experienced a smaller spring pulse of upwelling (similar to 2014-2016) that persisted from spring to early fall. To better understand the patterns and processes that shape the pelagic ocean microbiome our analyses focus on five key functional groups based on their consequential roles in marine food webs and biogeochemical cycles (Azam et al. 1983; Calbet and Landry 2004; Buchan et al. 2014): heterotrophic bacteria, cyanobacteria, Archaea, and heterotrophic and photosynthetic eukaryotic protists. These functional groups comprise many smaller subgroups and amplicon sequence variants, or ASVs.

Within all groups, we find strong cross-shore patterns in community structure and diversity that align with gradients in nutrient supply to the surface ocean. Across both seasonal and interannual timescales, we find that the intensity of regional nutrient supply can alter cross-shore patterns in community structure varying the availability of habitat for highly productive nearshore communities. These results confirm previously observed patterns in well-studied taxonomic groups and suggest that similar environmental forcings shape the community

structure and diversity of cryptic groups that were not possible to resolve through traditional techniques. Our study represents a synthesis of how both temporal and spatial environmental gradients influence microbial community assembly in a coastal upwelling biome, providing fundamental knowledge about the structure and diversity at the base of a highly productive and economically valuable ecosystem.

## 2.2 Results

Across 995 samples, small subunit ribosomal RNA gene sequencing was performed on the V4-V5 region of the 16S rRNA gene for prokaryotes and the V9 region of the 18S rRNA gene for eukaryotes. Within these samples, we identified 19,204 16Sv4-5 ASVs and 34,454 18Sv9 ASVs (Table 2.1S). Compared to the number of 18Sv9 ASVs observed in Tara Oceans (207,827) or Tara Polar (65,655) (de Vargas et al. 2015), the number of ASVs found in the Southern California Current region was lower (Fig. 2.1e). However, of the 18Sv9 ASVs identified within NCOG, 43% were not found in either Tara survey, highlighting both the undersampling of coastal ecosystems in global datasets and the added value of repeat monitoring through time towards uncovering novel regional diversity. A large proportion of the ASVs that were only found in NCOG are dinoflagellates, though many others belonged to a diverse set of taxonomic groups (Fig. 2.2S).

### 2.2.1 Spatial gradients in community structure and diversity

Nearshore to offshore gradients in community structure were an emergent property found in our self-organizing maps (SOMs; see Methods) and occurred within all five key functional groups: heterotrophic bacteria, cyanobacteria, Archaea, and heterotrophic and photosynthetic eukaryotic protists (Fig. 2.2). SOMs are a neural-network, data reduction technique which we used to convert the highly dimensional ASV tables (995 samples x 1,000s of ASVs) into a 2-

dimensional map (Kohonen 1997). Both surface (10m) and deep chlorophyll maximum (DCM) samples were included in this analysis. Community clusters identified by SOMs have been subsequently labeled as “nearshore” or “offshore” based upon whether they were found more frequently in nearshore or offshore stations (weighted centroid). For the five key functional groups, these clusters aligned with waters of contrasting trophic status. On average, stations found in the northeast, nearshore corner of the sampling grid experienced mesotrophic (2.5-8 µg Chl-a L-1) and eutrophic conditions (> 8 µg Chl-a L-1) (Istvánovics 2009). This contrasted strongly with the oligotrophic conditions found in most of the stations further offshore, where chlorophyll was typically low (< 2.5 µg Chl-a L-1) (Fig. 2.1Si).

Differences in community structure, as classified by SOM clusters, were driven by the differential relative abundance of ASVs within each of the five main groups. Within each of the five groups, there were finer-grained subgroups (e.g., SAR 11 clade and diatoms) that exhibited differences in mean relative abundance between SOM clusters. SAR 11 ASVs were abundant in both the nearshore and offshore clusters (Fig. 2.3Sa). However, what initially appeared to be a homogenous distribution of SAR 11 across the region was driven by three distinct SAR 11 Clade 1a ASVs: one that dominated the nearshore and two that dominated the offshore (Supplementary Data 1). One previously identified relationship within cyanobacteria (Partensky, Blanchot, and Vaulot 1999) was observed where Procholoroccus ASVs had a higher relative abundance in the offshore and Synechococcus ASVs had a higher relative abundance in the nearshore (Fig. 2.3Sc). Within the eukaryotic phytoplankton, diatoms were abundant in the nearshore but not the offshore SOM cluster (Fig. 2.3Se). Dominant nearshore diatom genera/species included: Thalassiosira, Chaetoceros, and Pseudo-nitzchia. In contrast, dinoflagellates dominated the offshore SOM cluster (Fig. 2.3Se). Dominant offshore dinoflagellates included: Karlodinium

veneficum, Warnowia, and Prorocentrum. The ASVs that show the greatest differential abundance (> 99th percentile) between nearshore and offshore clusters are provided in Supplementary Data 1.

The export rate of primary production (ef-ratio) also varied in relation to SOM clusters (Fig. 2.4S). Here ef-ratio is defined as new production/total production = export production/total production (Laws et al. 2000), where higher ef-ratio values indicate increased export of surface primary productivity to depth (important for carbon sequestration within the ocean). This was particularly evident in both the cyanobacteria and photosynthetic eukaryotic protists SOM clusters (Fig. 2.4Sc-d), which both showed strong and significant relationships between the frequency with which their nearshore cluster was observed at a given station and the mean ef-ratio at that station over the seven years. The strong link between ef-ratio and proportion of nearshore and offshore communities highlights the connection between community structure and function, in this case the export of carbon from the ocean surface.

SOMs were also generated for eleven more finely resolved taxonomic groups (for a list of all groups see Table 2.1S). Seven out of the eleven groups showed a similar nearshore-offshore gradient in community structure, while other groups, such as Prochlorococcus and haptophytes showed little to no spatial patterns in community structure (Fig. 2.5S).

We extended the SOM analysis to examine the relationship between the frequency of observed community type (nearshore/offshore) against environmental covariates, using both the mean and coefficient of variation (coeff. var.) at each station across all seven years. In doing so, we identified the conditions across all seven years that best align with spatial patterns in the occurrence of nearshore or offshore microbial communities within the region. Coefficients of variation were included in this analysis as environmental variability is thought to promote

distinct life strategies and drive population dynamics in phytoplankton species (Grover 1990; Benincá et al. 2008). Nitracline depth (see Methods for definition) was a significant predictor of the nearshore-offshore gradient in community structure (lowest Akaike information criterion, AIC, Fig. 2.3), with the mean or coefficient of variation of nitracline depth being the most significant environmental predictor of community structure for eight out of the eleven taxonomic groups (Fig. 2.6S). Nitracline depth varies as the result of both abiotic and biotic factors, with upwelling bringing nutrients to the surface waters leading to a shallower nitracline and biological drawdown of nitrate within the surface ocean leading to a deepening of the nitracline. As such, nitracline depth is thought to be a critical indicator of nutrient supply into the surface ocean (Williams and Follows 2011) and can be seen as both a potential driver as well as a potential response to community changes. Mean chlorophyll a concentrations were also a significant predictor of the nearshore-offshore gradient in community structure (Fig. 2.3). However, this variable may not signify a mechanistic link, but instead reflect the ecosystem state, particularly for groups that comprise our chlorophyll a measurements (Lindegren et al. 2016).

Mean alpha (α) diversity across all ASVs, in this case calculated as the mean per station per cruise diversity, generally increased away from shore (Fig. 2.4a,b). For this analysis, Shannon index was used as the primary measure of diversity. The lowest mean alpha diversity was present in the northeast, nearshore subregion of the SCC and the highest mean alpha diversity was seen in the furthest offshore stations in the south. Across both surface and DCM samples (separately) we observed the same pattern of low diversity in the nearshore and high diversity offshore (Fig. 2.7S). Overall diversity was higher in the DCM compared to the surface, this was also true for archaea, bacteria, and cyanobacteria (Fig. 2.7Sa-d). In contrast, autotrophic and eukaryotic protist tended to have similar levels of diversity in both the surface and DCM

samples (Fig. 2.7Se-f). Similar increases in mean alpha diversity away from shore were found among most taxonomic subgroups (e.g., Prochlorococcus, SAR 11 Clade, and Syndiniales; Fig. 2.8S). However, the direction of the gradient was reversed (high diversity nearshore, low diversity offshore) for diatoms (Fig. 2.4d). Gamma diversity (ɣ; total diversity at a station over all time points) also increased away from shore (Fig. 2.4b), but certain groups were distinct from the pattern across all ASVs. For instance, within diatoms, mean alpha diversity was greatest nearshore, but there was little to no gradient in gamma diversity (Fig. 2.4d).

Nitracline depth (mean/coeff. var.) was the best predictor of spatial gradients in mean alpha diversity for all major groups except Archaea (Fig. 2.4e) and four out of the eleven taxonomic subgroups (Fig. 2.9S). Three of the eleven subgroups were better predicted by the coefficient of variation in nitrate concentrations (Fig. 2.9S). For most groups, the relationship between nitracline depth and mean alpha diversity was positive, while, for certain groups such as diatoms, Synechococcus, and Flavobacteriales, this relationship was negative.

Previous studies have shown that diversity-productivity relationships can be unimodal (Vallina et al. 2014), or vary with scale (Chase and Leibold 2002). For the subset of our data where primary-productivity measurements are available (Supplementary Data 2), we found a wide variety of productivity-diversity relationships (Fig. 2.10S). Positive productivity-diversity relationships occurred within flavobacteria and diatoms and negative relationships occurred for the SAR 11 clades, Prochlorococcus, and Syndiniales. In some groups, the productivity-diversity relationship appeared consistent across all time periods (e.g. Prochlorococcus, SAR 11, Syndiniales), while others appeared to vary depending of the time period (Haptophytes, Chlorophytes, Dinoflagellates, Fig. 2.10S).

### 2.2.2 Temporal gradients in community structure and diversity

To better understand how community structure and diversity might be affected by temporal environmental variation, we first looked at how the environment changed over seasonal to interannual time scales in this region. Given the primary importance of nutrient supply in shaping spatial ecological gradients (Figs. 3-4), we focused on how coastal upwelling and nutrient availability in the surface ocean was affected across the seven-year study period. We examined three local indices of upwelling presented by Jacox et al. (2018): Coastal Upwelling Transport Index (CUTI), Biologically Effective Upwelling Transport Index (BEUTI), and Regionally Available Nitrate (Fig. 2.5a-c). In the SCC, physical upwelling (CUTI) and regionally available nitrate tend to be the lowest in late fall through winter and highest in the spring to early summer (Fig. 2.5 a,c). While physical upwelling (CUTI) was similar throughout the years of study (Fig. 2.5a), the biologically effective upwelling (BEUTI) was much lower during the first three years which were affected by the 2014-2015 warm anomaly and El Niño (Fig. 2.5b). Upwelling in 2019-2020 was unique compared to the other years, characterized by a spring period with relatively low CUTI and BEUTI but an overall expanded upwelling season (stronger upwelling into the summer and fall relative to all other years). During the anomalously warm years 2014-2016, nitrate concentrations were relatively low in the northeast, nearshore subregion of the Southern California Current region (Fig. 2.5d-f). In 2014-2016, phosphate and silicate concentrations were also lower close to the coast in the northeast subregion, while concentrations of these nutrients were higher everywhere else (Fig. 2.11S). Mixed layer and nitracline depths across the region were similar between nearshore and offshore stations from 2014-2016—likely the result of intense stratification within the surface ocean (Zaba and Rudnick 2016) (Fig. 2.11S).

Interannual changes in microbial community composition across contrasting warm and cool periods were pronounced, with the largest changes occurring within eukaryotic groups (Fig. 2.5g-k). We compared the warm period in 2014-2016 with the relatively cool period that followed in 2017-2018, as these two periods had strongly contrasting environmental conditions. The conditions in 2019-2020, which we discuss below, were intermediate between the warm and cool phases—the offshore experienced a warm anomaly similar to 2014-2015 (Weber et al. 2021), while the nearshore experienced an expanded, though moderate, upwelling season. We calculated the average community similarity (Bray-Curtis) between surface samples across the warm and cool phases for each station across our five major groups (Fig. 2.5g-k). Archaea, photosynthetic eukaryotic protists, and heterotrophic eukaryotic protists, showed large shifts in community structure between the warm and cool phases (low Bray-Curtis Similarity, Fig. 2.5g,j,k). Cyanobacterial communities appeared to change less between the two phases than the other groups, particularly in the offshore (Fig. 2.5i). Changes within the samples collected at the deep chlorophyll maximum (DCM) between the warm and cool phases were less pronounced, though photosynthetic eukaryotic protist communities within the DCM were quite different between the two phases (Fig. 2.12S). Overall, eukaryotic groups exhibited far greater region-wide shifts in community structure between the warm and cool phases (Fig. 2.13Sf-k). Prokaryotes, such as those ASVs assigned to the SAR 11 clade, had little to no change in community composition between the two phases (Fig. 2.13Sa-e). Groups like Prochlorococcus showed almost no change in community composition in the offshore between the two phases, while simultaneously exhibiting drastic shifts in community structure in the nearshore environment (Fig. 2.13Sa).

The 2014-2016 warm anomaly, which was localized to the upper 50 meters of the water column (Zaba and Rudnick 2016), had a clear influence on the effectiveness of physical upwelling to deliver nutrients to the surface ocean relative to 2017-2018 (Closset et al. 2021) (Fig. 2.5, and Fig. 2.11S). This intense stratification may have shaped where, when, and how communities changed across the region. To test the hypothesis that temporal changes to regional stratification drove microbial community structure, we examined the relationship between the regional, cross-shore slope of nitracline depth and the proportion of samples that were identified as the nearshore (per taxonomic group via our SOMs) on a cruise-by-cruise basis. The regional slope of nitracline depth was calculated for each cruise by first flattening the sampling grid into a two-dimensional plane where the x-axis was distance to the coast (km), and the y-axis was the nitracline depth (m) for each station. A regional slope of the nitracline depth for each cruise was then calculated as the best linear fit through the points in this two-dimensional plane (Fig. 2.14S). Under normal upwelling conditions we expect the nitracline depth to be shallowest in the nearshore, coastal upwelling region, and deepest in the offshore, leading to a steep regional slope in the nitracline depth. Conversely, intense stratification of the surface ocean would promote a deeper nitracline depth in the nearshore and a shallower nitracline depth in the offshore, flattening the regional slope of nitracline depth.

We found that during the warm and cool periods, when the regional slope of nitracline depth was steeper (shallow in the nearshore and deep in the offshore), a higher proportion of samples were identified as the nearshore community type for both photosynthetic groups (cyanobacteria and photosynthetic eukaryotes) as well as bacteria. Conversely, when the regional slope of the nitracline depth was relatively flat, fewer samples were identified as nearshore (Fig. 2.6). Across all years (2014-2020), cruises in the spring and summer tend to have the steepest

regional nitracline slopes (for an illustrative example see Fig. 2.6b). Fall and winter tended to have shallower regional slopes in nitracline depth and also tended to have a lower proportion of observed nearshore communities (for an illustrative example see Fig. 2.6a). Winter 2019 appeared to be quite distinct for this dataset, as the cruise data suggested that the region was experiencing the flattest regional slope in nitracline depth observed in all seven years, yet the proportion of nearshore communities was relatively high. However, sampling during this cruise was abnormally compressed (8-days across fewer stations) due to ship malfunction, making interpretation difficult.

Most groups tended to have a seasonal pattern in the relative dominance of nearshore/offshore communities (Fig. 2.15S). SAR 11 nearshore communities were more common in the spring and summer (Flavobacteriales, Rhodobacterales, metazoans showed similar trends). Other groups such as Prochlorococcus and diatoms showed peaks in the winter, though the presence of an increased nearshore diatom community tended to last through the spring as well (Fig. 2.15S). While seasonal patterns in community structure were common across all groups, the pattern was not always consistent across all years.

The 2019-2020 time period was characterized by two major anomalies, a warm, stratified layer of surface water (similar to 2014-2016) but localized to the offshore (Weber et al. 2021), and prolonged biologically effective upwelling from spring through early fall (Fig. 2.5b). These events combined to decrease the interseasonal variability of nutrient supply to the surface ocean within the SCC from 2019-2020. As a result, relationships between the nitracline slope and spatial extent of the nearshore communities were uncoupled in 2019-2020 (Fig. 2.6). This was particularly evident in diatoms and dinoflagellates, two groups that respond strongly to changes

in nutrient supply (Cermeño et al. 2008; Kenitz et al. 2020), where seasonal patterns in the relative abundance of nearshore communities disappeared in 2019-2020 (Fig. 2.15S).

Temporal changes to mean alpha diversity occurred across both seasonal and interannual time scales. In contrast with the findings related to community structure, mean alpha diversity tended to be highest when the regional slope of nitracline depth was most flat, although, certain groups such as diatoms exhibited the reverse pattern though the relationship was not significant (2017-2018, Fig. 2.16S). Like community structure, relationships between diversity and regional nitracline slope were far more frequent in the earlier years of sampling (2014-2018), when interseasonal variability in the regional nutrient supply was higher (Fig. 2.5b). Metazoans were the only group that showed a relationship between the regional nitracline slope and mean alpha diversity in 2019-2020 (Fig. 2.16Sk).

## 2.3 Discussion

The depth of the nitracline was a robust predictor of community structure in the SCC (Fig. 2.3b). In this region, the nitracline tends to be deeper in offshore waters and shallower in nearshore waters (Mullin 1998), creating strongly contrasting habitats. The depth of the nitracline is shaped to a great degree by the strength of upwelling; when upwelling is stronger, the nitracline is closer to the surface, and the supply of nutrients to the surface is higher, if not the actual concentration of nutrients in the surface (Mullin 1998; Rykaczewski and Checkley 2007). Nitrate limitation, as the result of variable nutrient supply, can exhibit a strong selective pressure on marine microbial communities, forcing organisms into metabolic tradeoffs in order to survive (Grzymski and Dussaq 2012). Thus, the strongly contrasting environments in the nearshore and offshore within the SCC select for very different communities. Because nutrients are rapidly consumed by microbes in the ocean surface, the concentrations of nutrients measured

represent the residual not consumed by microbes, and are in many cases not as good of a predictor of community composition when compared to the nitracline depth (Hayward and Venrick 1998; Kenitz et al. 2020).

Nitracline depth was also more strongly correlated with community structure changes than temperature (Fig. 2.3b, Fig. 2.6S). On local to global scales, nutrient availability strongly shapes primary productivity and community structure (Margalef 1978; Falkowski and Oliver 2007; Taylor et al. 2015; Mende et al. 2017; Phoma and Makhalanyane 2021). Yet in a range of recent studies, temperature has been shown to be a key correlate of global patterns of bacterial (J. A. Fuhrman et al. 2008; Sunagawa et al. 2015; Ibarbalz et al. 2019) (16S) and protistan (Ibarbalz et al. 2019) (18S) biodiversity and community structure as well as changes in the functional community composition of marine bacteria (Sunagawa et al. 2015). Surprisingly, these studies found little to no relationship between biodiversity, community structure, functional community composition and nitracline depth. A possible explanation is that global surveys of microbial communities have, thus far, focused their sampling effort within the open ocean, failing to capture strong coastal-open ocean physical and ecological gradients. The relative importance of environmental factors in shaping marine microbial community structure is likely to vary between regions (Z. Wang et al. 2019) and across different spatial scales (local to global). This is likely the result of both the overall selective pressure of a variable and its relative range within the observable spatiotemporal scope of the study. Yet here within the SCC, large spatial gradients in nutrient availability, compared with temperature variability, occur with both seasonal and interannual variability, providing a testing ground to explore the selective pressure of nutrient availability in a coastal upwelling region.

Previous studies have highlighted the strong cross-shore gradients in community structure in the SCC, primarily through the use of general indices (Kahru and Mitchell 2001; Taylor et al. 2015) (such as the ratio of autotrophic carbon to chlorophyll a) or select groups of bacteria (Taylor et al. 2015), phytoplankton (Hayward and Venrick 1998; Venrick 2009; Taylor et al. 2015; Barth et al. 2020; Catlett et al. 2021) and zooplankton (Powell and Ohman 2015a). The results generated from this study support and expand upon many of the findings from these previous studies. Taylor et al. 2015 (Taylor et al. 2015) found that the ratio of autotrophic carbon (AC) to Chl-a increased with increasing nitracline depth within the SCC and that the relatively low ratios of AC:Chl-a near the coast were a result of the dominant nearshore diatom communities which have low AC: Chl-a ratios. In turn, these diatom-dominated communities can lead to an "enhanced" microbial loop, with higher flows and heterotrophic bacteria standing stock biomass (Taylor and Landry 2018). We find similar evidence that gradients in nitracline depth structure community composition in both phytoplankton and bacterial groups. Given the level of taxonomic resolution provided by ASVs, we were able to expand upon these prior studies to identify that these gradients also shape the taxonomic composition within groups (such as diatoms, dinoflagellates, rhodobacteria, and SAR 11 clade Fig. 2.5S), highlighting spatio-temporal variability in community structure at a previously inaccessible resolution. These results suggest that selection across gradients such as nutrient limitation can drive not only dominance between taxonomic groups with contrasting ecological niches and functions (diatoms vs cyanobacteria) but also drive selection within groups that are traditionally "lumped" into singular functional and or taxonomic groups (Fig. 2.5S, Fig. 2.8S). Furthermore, ASVs allow for the examination of "cryptic" groups that cannot be identified through traditional approaches (microscopy, flow cytometry, chl-a) such as various heterotrophic bacteria (rhodobacteria,

flavobacteria, SAR 11 clade) and Archaea. We found that groups such as SAR 11, which are often thought to have cosmopolitan distributions, are comprised of distinct strains with varying oligotrophic to eutrophic preferences. The patterns and processes identified within this study confirm the relationship between nutrient availability microbial community structure in the SCC while further highlighting that these selective processes not only drive preferences between large functional and taxonomic groups, but also within groups.

Across most groups, mean alpha diversity was lower in the nearshore and higher offshore (Fig. 2.4e). The nearshore environment had relatively high nutrient concentrations and temporally variable habitats (Fig. 2.1S), factors which favor the competitive dominance of fast-growing, opportunistic phytoplankton such as diatoms at the expense of other species, and likely leading to lower diversity nearshore (Dutkiewicz, Follows, and Bragg 2009; Barton et al. 2010; Vallina et al. 2014). In some cases, the coefficient of variation of nitrate was a good predictor of spatial biodiversity patterns (Fig. 2.9S), highlighting that the nearshore environment, with its high variability and episodic pulses of nutrients, may exhibit a strong selective pressure for organisms adapted to this variable environment. An additional explanation could be that the offshore subregion of the CalCOFI grid represents a mixing zone, or ecotone, combining subtropical and coastal communities with consequently relatively high diversity (Barton et al. 2010; D'Ovidio et al. 2010; Clayton et al. 2013; Moisan et al. 2017). The CalCOFI grid does not, however, include stations spanning deep into the subtropical North Pacific, so we cannot assess this possibility.

Diatoms presented a notable exception to the observed diversity patterns, as they showed an opposite trend in mean alpha diversity, with higher mean alpha diversity in the more productive nearshore region (Fig. 2.4c). While diatoms are found in subtropical waters globally,

they are generally more abundant in regions and seasons with higher nutrient availability (Margalef 1978), and this may underpin the greater alpha diversity observed within this coastal zone. In contrast, we find no evidence of a nearshore-offshore gradient in diatom gamma diversity (Fig. 2.4d). This suggests that over the seven years, diatom community turnover was higher in the offshore subregion of the SCC. One possible explanation for the high overturn in diatoms but not other microbial assemblages stems from the intermittent presence of eddies and fronts in offshore waters that mediate vertical motions and nutrient supply (Combes et al. 2013; Chenillat et al. 2013; Chenillat, Franks, and Combes 2016). Because diatoms as a group are faster-growing than other microbial groups (Edwards et al. 2012), their populations respond faster to episodic pulses in nutrients than other groups. The intermittent passage of eddies and fronts in offshore waters may therefore drive an overturn of diatom ASVS while not creating a similar overturn in other groups.

Variation in the intensity of coastal upwelling across seasonal to interannual time periods controlled the relative dominance of offshore vs. nearshore community types and diversity observed within the region. During periods of strong upwelling, coastal communities were more dominant and mean alpha diversity was lower (Fig. 2.6, Fig. 2.15S, Fig. 2.16S). Conversely, when the regional, cross-shore slope in nitracline depth was flat, most samples resembled the "offshore" community type in both structure and diversity. The 2014-2015 warm anomaly and subsequent 2015-2016 El Niño drastically reduced the extent of coastal upwelling and nutrient availability in surface waters within the region, converting nearly all available habitat into an environment that favored offshore communities. From 2014-2016, within fall and winter cruises, the majority of samples were identified as resembling an "offshore" ecotype, suggesting a drastic departure from the typical ecological gradients that exist in the region (Fig. 2.6). In particular,

the eukaryotic assemblage changed substantially between the warm and cool phases (Fig. 2.5j,k). Many of the taxonomic groups showed region-wide shifts in community composition between the two phases (for example: diatoms and Syndiniales, Fig. 2.13S). 2019-2020 brought the return of the marine heatwave, though unlike 2014-2016, its effects were primarily observed offshore (Weber et al. 2021). BEUTI measurements from the region suggest that spring upwelling for 2019-2020 had been closer to 2014-2016, however, this upwelling persisted to some degree through summer and early fall (Fig. 2.5b). This may have led to our observation that for certain groups such as diatoms and dinoflagellates, seasonal shifts in community structure were less pronounced (Fig. 2.15S). These temporal changes in the marine microbial community have implications for higher trophic levels. For example, anchovies tend to predominate in more nutrient rich coastal waters while sardines are more abundant in oligotrophic conditions offshore (Rykaczewski and Checkley 2007). Consistent with this paradigm, following the 2014-2016 warm anomaly, anchovy egg counts in Southern California reached high levels in 2017 and 2018 that had not been seen since the mid 1990s (Wells et al. 2017; Thompson et al. 2018).

While previous metabarcoding studies have explored how community structure and diversity changes over time at one location (Gilbert et al. 2012; Jed A. Fuhrman, Cram, and Needham 2015; Ward et al. 2017), here we provide a comprehensive metabarcoding exploration of seasonal to interannual community variation at the regional scale. The unique lens afforded by this dataset suggests that community variability can occur across space and time, though their relative influence may vary depending on the spatial extent of temporal perturbations. We find that the depth of the nitracline is a robust predictor of both microbial community structure and biodiversity and that globally important variables such as temperature are far less predictive in the Southern California Current. Furthermore, we found that changes in community composition

can be found not only between large functional groups, but also within groups that are often considered functionally similar. Metabarcoding also allows for the investigation of "cryptic" groups whose patterns and processes have previously been inaccessible. Across the seven years we show that changes to the spatial patterns of community structure and biodiversity coincide with seasonal and interannual changes to the steepness of cross-shore physical gradients (nitracline depth). Physical differences within the region between the warm (2014-2016) and cool (2017-2018) phases brought drastic changes in community composition, whereas reductions in the interseasonal variability of nutrient supply from 2019-2020 led to a more "static" community structure across the region. Combined, these results highlight the clear benefits of genomic surveys that sample across both space and time. Provided that there is adequate support and infrastructure to do so, future studies should be conducted in a similar manner if we are to better understand the linkages between the physical environment and microbial community structure and biodiversity.

## 2.4 Methods

### 2.4.1 Study location and sample collection

The Southern California Current ecoregion is a component of one of the world's most productive eastern boundary currents. Productivity in the region is largely driven by seasonal upwelling—triggering the dominance of bloom forming eukaryotic phytoplankton (like diatoms) in the spring that serve as the base of a food web supporting a diverse ecosystem and many economically important fisheries (Hayward and Venrick 1998; Venrick 2009; Bograd, Schroeder, and Jacox 2019).

Molecular and environmental data were collected on quarterly CalCOFI cruises (winter, spring, summer, and fall). At each station, seawater was collected near the surface (10 m) and the

depth of the chlorophyll maximum, which varies in time and space. The chlorophyll maximum is identified on the downcast of the CTD and subsequently sampled on the upcast of the CTD. If these two depths coincided with one another then only one seawater sample was collected. Two types of stations were sampled during this study: cardinal stations and productivity stations. Cardinal stations were sampled every cruise and occur on lines 80 (stations 55.0, 70.0, 80.0, 100.0), 81.8 (station 46.9) and 90 (stations 37.0, 53.0, 70.0, 90.0, 120.0) (Fig. 2.1a). Productivity stations, which measure 14C primary production at approximately local noon were also sampled. The locations of productivity stations vary from cruise to cruise depending on where the ship is located each day at approximately local noon. Productivity stations can overlap with cardinal stations during a given cruise if the ship is located at a cardinal station at local noon.

Both molecular and environmental data were collected from a CTD rosette. Temperature and salinity were measured with a Seabird 911+ CTD. CTD salinity is validated against bottle samples which were analyzed via a Guildline Portasal Salinometer model 8410A. Nitrate, phosphate and silicate measurements were analyzed with a QuAAtro continuous segmented flow autoanalyzer (SEAL Analytical). For chlorophyll a, seawater was filtered onto GF/F filters and then measured with the acidification method. Full methods for environmental data collection and analysis can be found at: https://calcofi.org/references/methods. At primary productivity stations, 14C half-day incubations were started at local noon and measured as mg of carbon per m3 per half day. Integrated primary production in the euphotic zone was then calculated as the average primary production across six light levels. For a complete procedural walkthrough of productivity incubations see: https://calcofi.org/references/methods/25-primary-productivity.html. For this study, primary productivity measurements were doubled to estimate the total production per full light day. The nitracline depth is a derived variable and is calculated

as the depth where nitrate concentrations exceed or reach 1 μM via a linear interpolation based on discrete depth measurements. Metadata for all samples can be found in Table 2.2S.

### 2.4.2 DNA collection and extraction

Approximately 0.5 – 2 L of seawater was filtered through a 0.22 μm Sterivex-GP filter unit (MilliporeSigma, Burlignton, MA, USA) for all DNA samples. Samples were immediately sealed with a sterile luer-lock plug and hematocrit sealant, wrapped in aluminum foil, and flash frozen in liquid nitrogen. DNA was extracted with the NucleoMag Plant Kit for DNA purification (Macherey-Nagel, Düren, Germany) on an epMotion 5057TMX (Eppendorf, Hamburg, Germany) as described here: https://dx.doi.org/10.17504/protocols.io.bc2hiyb6. DNA was assessed on a 1.8% agarose gel after extraction.

### 2.4.3 Amplicon sequencing and analysis

Amplicon libraries targeting the V4-V5 region of the 16S rRNA gene and V9 region of the 18S rRNA gene were generated as described here: https://www.protocols.io/view/amplicon-library-preparation-bmuck6sw. Briefly, DNA was amplified via a one-step PCR using the TruFi DNA Polymerase PCR kit (Azura, Raynham, MA, USA). For 16S, the 515F (GTGYCAGCMGCCGCGGTAA) and 926R (CCGYCAATTYMTTTRAGTTT) primer set was used (Parada, Needham, and Fuhrman 2016). For 18S, the 1389F (TTGTACACACCGCCC) and 1510R (CCTTCYGCAGGTTCACCTAC) primer set was used (Amaral-Zettler et al. 2009). Each reaction was performed with an initial denaturing step at 95°C for 1 minute followed by 30 cycles of 95°C for 15 seconds, 56°C for 15 seconds, and 72°C for 30 seconds. Custom mock communities (Parada, Needham, and Fuhrman 2016) were included in the sequencing runs (Fig. 2.17S). 2.5 μL of each PCR reaction was ran on a 1.8% agarose gel confirm amplification. PCR products were purified using Beckman Coulter AMPure XP beads following the standard 1x

PCR clean-up protocol. PCR quantification was performed in duplicate using Invitrogen Quant-iT PicoGreen dsDNA Assay kit. Samples were then pooled in equal proportions into seven pools for the 16s data and five pools for the 18s data followed by another 0.8x AMPure XP bead purification. Pools was evaluated on an Agilent 2200 TapeStation and quantified with Qubit HS dsDNA. Each pool was sequenced at the University of California, Davis Sequencing Core on a single Illumina MiSeq lane (2 x 300 bp for 16S, 2 x 150bp for 18S) with a 15% PhiX spike-in. For the 2014-2016 data, the 18s pool was sequenced on an Illumina NextSeq (2 x 150 bp).

Amplicons were analyzed with QIIME2 v2019.104 (Bolyen et al. 2019). Briefly, demultiplexed paired-end reads were trimmed to remove adapter and primer sequences with cutadapt (Martin 2011). Trimmed reads were then denoised with DADA2 to produce amplicon sequence variants (ASVs) (Callahan et al. 2016). Each pool was denoised with DADA2 individually to account for different error profiles in each run. Taxonomic annotation of ASVs was conducted with the q2-feature-classifier classify-sklearn naïve-bayes classifier (Bokulich et al. 2018; Pedregosa et al. 2011) against SILVA (Release 138) (Pruesse et al. 2007) for 16S amplicons or PR2 v4.13.0 (Guillou et al. 2013) for 18S amplicons.

Tara Oceans and Tara Polar data were downloaded from the European Nucleotide Archive under the project accessions PRJEB6610 [https://www.ebi.ac.uk/ena/browser/view/PRJEB6610] and PRJEB9737 [https://www.ebi.ac.uk/ena/browser/view/PRJEB9737]. Raw sequences were analyzed in the QIIME2 environment with DADA2 as described above. As run information was not available, each sample was analyzed with DADA2 individually; however, on average each sample contains enough reads to accurately estimate the error rates (>1 million reads).

For this study, we rarefied our libraries to 17,000 reads, maintaining 99% of our samples (11 were removed due to small library sizes). While there have been arguments on either side concerning rarefaction in microbiome datasets (McMurdie and Holmes 2014; Gloor et al. 2016; Cameron et al. 2021), we believe that the wide variability in our library sizes, ranging from thousands of reads to hundreds of thousands of reads, justifies our decision to rarefy—large differences in library size can drastically alter biodiversity estimates (Cameron et al. 2021).

### 2.4.4 Biodiversity metrics

The Shannon Index was used in our measures of both alpha and gamma diversity. Mean alpha diversity was calculated per station (Fig. 2.4, Fig. 2.9S) or per cruise (Fig. 2.6, Fig. 2.16S). Gamma diversity was calculated by summing together all observed reads per station before calculating a Shannon Index to get the total gamma diversity per station across all seven years of sampling (Fig.4). Beta diversity was calculated as a Bray-Curtis Similarity (Fig. 2.5, Fig. 2.12S, Fig. 2.13S). Both Shannon Index and Bray-Curtis similarity were calculated using the vegan package in R (Oksanen et al. 2020).

### 2.4.5 Self-organizing maps (SOMs)

Self-Organizing Maps (SOMs) are a data reduction technique capable of reducing highly variable data into a two-dimensional map while retaining properties of the original highly dimensional data. Consequently, SOMs are suitable for identifying distinct ecological communities with amplicon data, as they can reduce the complexity of tens of thousands of unique species (ASVs) to a small set of discrete communities (Bowman et al. 2017). For these data we generated the SOMs on a 6x6 neuronal map using the SOMbrero package in R (Boelaert et al. 2014). SOMs included all 984 individual samples. For each taxonomic group, once a SOM was generated, hierarchical clustering was used to cluster neurons (nodes of the map) together,

identifying the two most distinct community clusters present on the maps (see Table 2.1S for a list of taxonomic groups). See Fig. 2.2 and Fig. 2.5S for station maps representing the relative dominance between the two SOM clusters for all taxonomic groups.

### 2.4.6 Generalized linear models (GLM)

Generalized linear models (GLMs) were used to test the relative importance of environmental conditions on plankton community structure in the California Current. For the first set of models (Fig. 2.3 and Fig. 2.6S), the response variable was the frequency at which a specific community (nearshore or offshore), as defined by the SOMs, was found at a given station. A binomial fit was used as the range of possible values was between 0 and 1. For the second set of models (Fig. 2.4 and Fig. 2.9S), the response variable was mean alpha diversity. In this case, the fit was normal as the distribution of mean alpha diversity values was close to normal. GLM's only considered stations with at least four data points (one year). Single parameter models were compared to one another using the Akaike Information Criterion (AIC) to identify the most suitable model (Johnson and Omland 2004).

### 2.4.7 Generalized additive models (GAM)

Generalized additive models were used to fit Shannon index-distance to coast (Fig. 2.4a,d) and productivity-diversity relationships (Fig. 2.10S). Here, GAMs were used as they provide a flexible and simple means of identifying relationships between variables without the need to specify a specific type of relationship (linear, exponential, logistic) per fit.

### 2.4.8 Data availability

The 16S rDNA raw reads have been deposited at NCBI under Bioproject IDs PRJNA555783 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA555783], PRJNA665326 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA665326] and PRJNA804265

[https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA804265] and Biosample accession nos. SAMN25705811-SAMN25706151, SAMN16250568-SAMN16251083, and SAMN25756929-SAMN25757078 and for the 2014-2016, 2017-2019, and 2020 periods respectively. The 18S rDNA raw reads have been deposited at NCBI under Bioproject IDs PRJNA555783 [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA555783], PRJNA665326 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA665326], and PRJNA804265 [https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA804265] and Biosample accession nos. SAMN25710021-SAMN25710361, SAMN16251281-SAMN16251796, and SAMN25757352-SAMN25757501 for the 2014-2016, 2017-2019, and 2020 periods respectively.
Tara Oceans and Tara Polar 18Sv9 sequences can be found at the European Nucleotide Archive under the project accession IDs PRJEB6610 [https://www.ebi.ac.uk/ena/browser/view/PRJEB6610] and PRJEB9737 [https://www.ebi.ac.uk/ena/browser/view/PRJEB9737] respectively.

Associated sample metadata are provided in the Supplementary Data 2 file.

### 2.4.8 Code availability

The code for this study is located at https://github.com/ChaseCJames/NCOG_Spatial_Environ (James et al. 2022).

DOI: 10.5281/zenodo.6359865

## 2.5 Figures

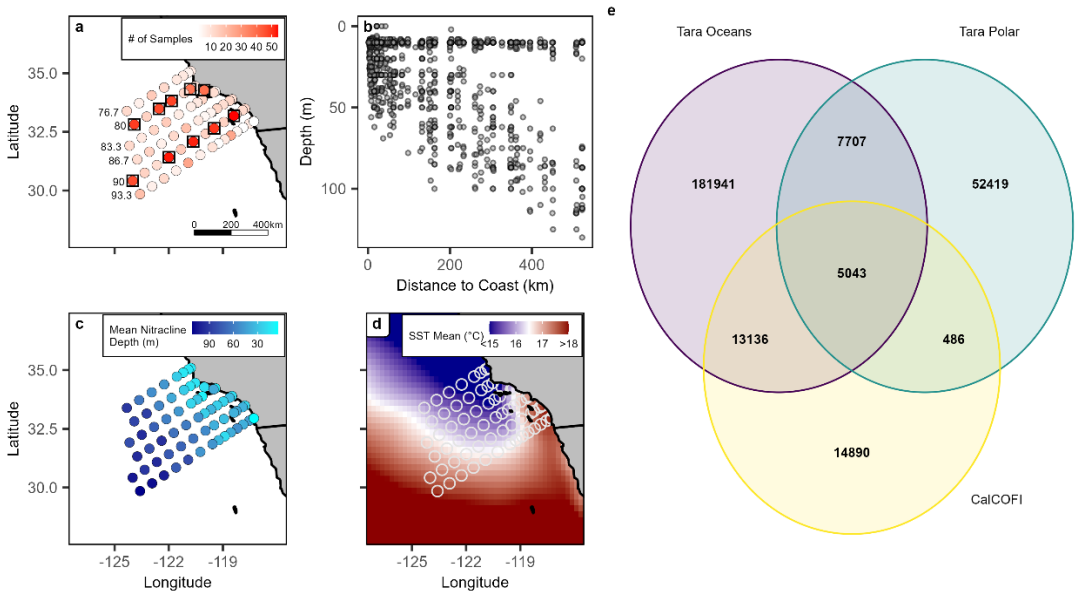


**Figure 2.1:** NCOG sampling and the physical environment of the Southern California Current region. a, the number of samples collected at each CalCOFI station from 2014-2020. Stations highlighted by squares are the Cardinal Stations (sampled every cruise) b, location of all samples by distance to coast (X-axis) and depth (Y-axis) c, mean nitracline depth (m) measured at each station across all seven years. d, mean SST (°C) from NOAA's OI SST V2 Dataset. White open circles represent the location of CalCOFI stations e, overlap between NCOG 18Sv9 ASVs with those present in Tara Oceans and Tara Polar.

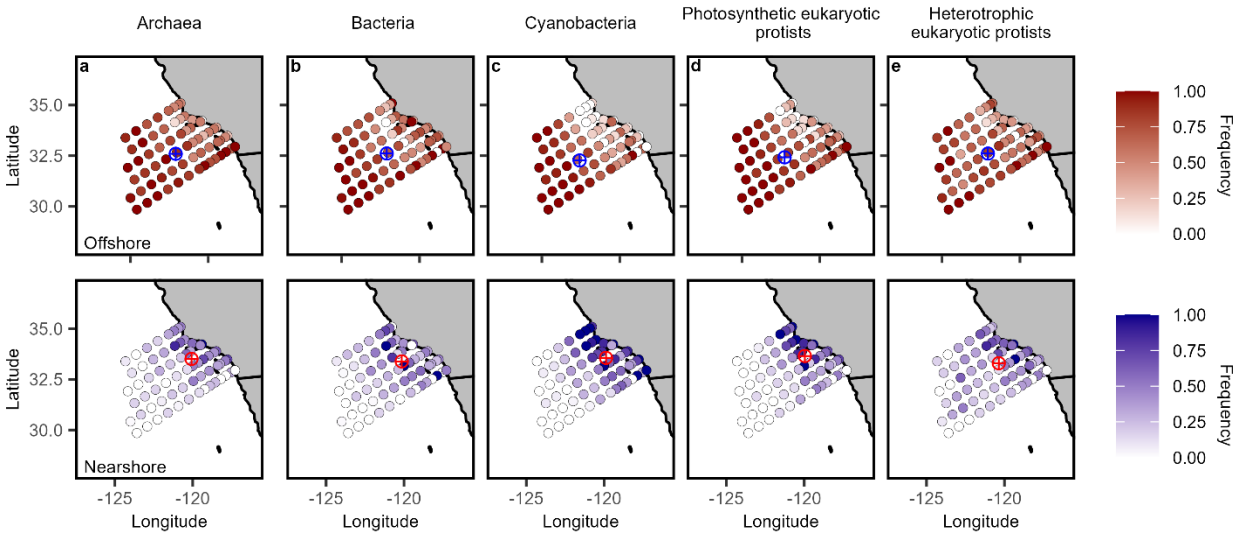


**Figure 2.2:** Nearshore and offshore gradients in community structure within five major microbial groups. (a-e) Colors indicate the frequency that the community at each location is offshore (red, top row) or nearshore (blue, bottom row) in character. The designation of nearshore vs. offshore community is determined by the cluster whose weighted centroid is closer to the coast. The weighted centroid for each cluster is shown as a circled plus symbol in the opposite color.

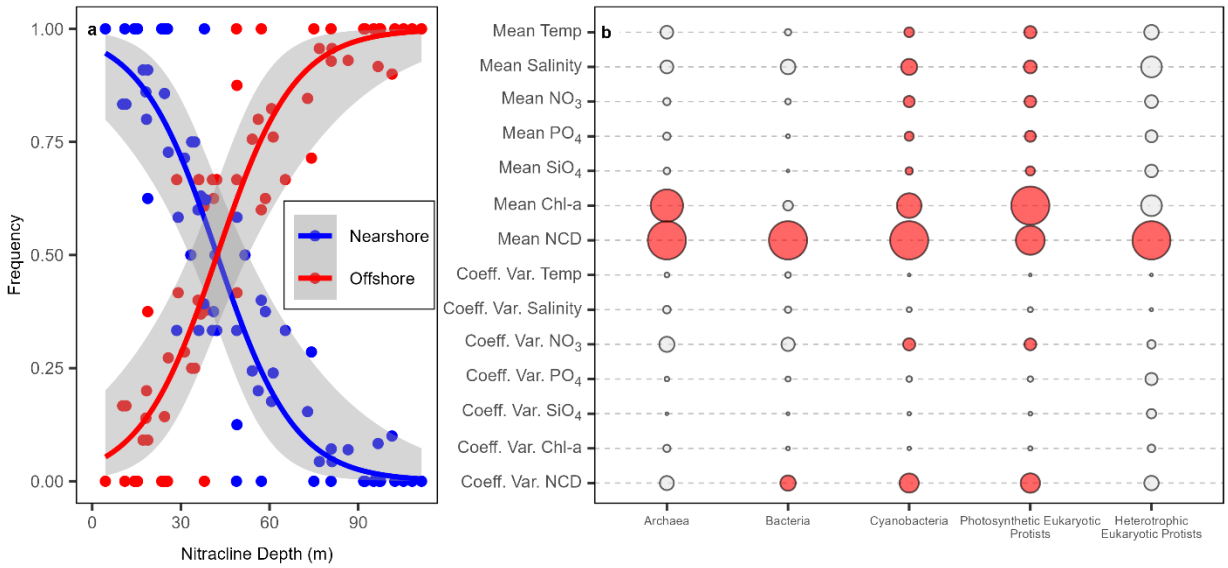


**Figure 2.3:** Environmental drivers of community structure. a, example relationship between mean nitracline depth (m) and the frequency of observed community clusters ("Nearshore" in blue or "Offshore" in red) per station for cyanobacteria. Lines represent a generalized linear model with a binomial fit. Shading represents a 95% confidence interval around the model fit b, relative importance of all explanatory variables (mean and coefficient of variation) used to predict the frequency of SOM clusters at a given station. As in the example (a), relationships were assessed via a generalized linear model with a binomial fit. Larger circles represent lower AIC values within a column; in other words, variables with larger circles are likely to be more important drivers than variables with smaller circles. Relationships that are not significant ($p > 0.05$) are colored in gray. Circles and their associated AIC values should not be compared across columns, only within columns, as AIC values are specific to each response variable. Relationships were analyzed between the frequency of observed community clusters and the mean and coefficient of variation (Coeff. Var.) of environmental variables. Environmental variables included: temperature (Temp), salinity, NO3, PO4, SiO4, chlorophyll a (Chl-a), and nitracline depth (NCD).

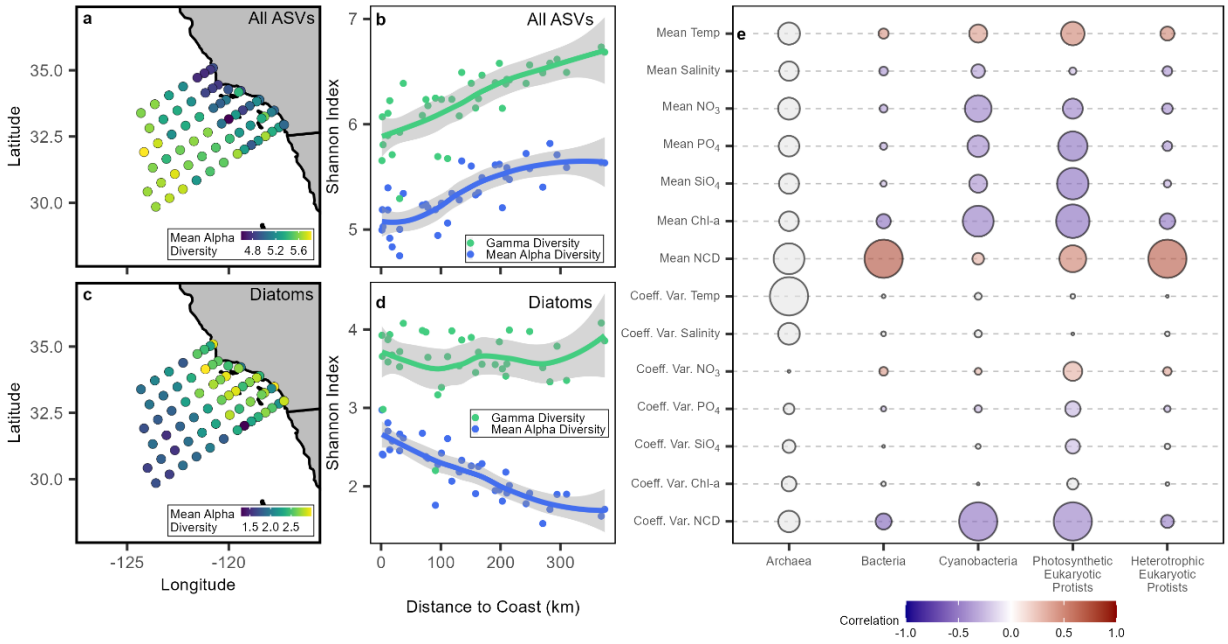


**Figure 2.4:** Spatial patterns and drivers of diversity. a, map showing the mean alpha diversity for all ASVs for each station. b, mean alpha (blue) and gamma diversity (green) per station for all ASVs as a function of distance to shore (km). Shannon index was used as the primary measure of diversity and was calculated as the mean per station per cruise for this analysis. Relationships are fit as a generalized additive model (GAM) with a 95% confidence interval. c, map showing the mean alpha diversity for diatoms for each station. d, mean alpha (blue) and gamma diversity (green) per station for diatoms as a function of distance to shore (km). Relationships are fit as a generalized additive model (GAM) with a 95% confidence interval (shading). e, relative importance of all explanatory variables (mean and coefficient of variation) used to predict mean alpha diversity at a given station. Relationships between environmental variables and diversity were assessed via a generalized linear model with a gaussian fit. Larger circles represent lower AIC values within a column. Circles and their associated AIC values should not be compared across columns. Color represents the correlation coefficient between each explanatory variable and mean alpha diversity. Gray circles represent relationships that are not significant ($p > 0.05$). Relationships were analyzed between diversity and the mean and coefficient of variation (Coeff. Var.) of environmental variables. Environmental variables included: temperature (Temp), salinity, $NO_3$, $PO_4$, $SiO_4$, chlorophyll a (Chl-a), and nitracline depth (NCD).

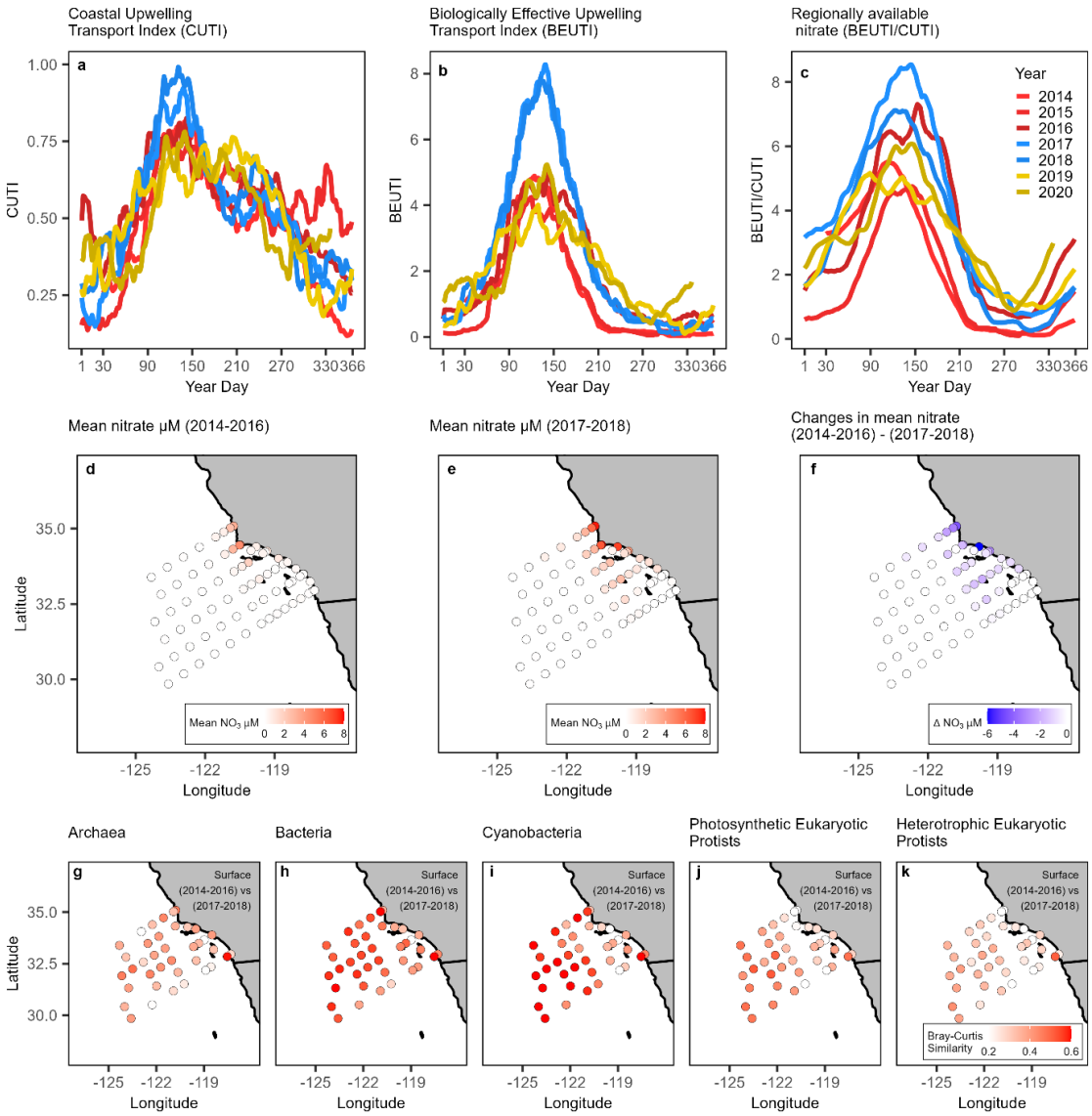


**Figure 2.5:** Physical and ecological changes in the region across time. a, annual cycle of Coastal Upwelling Transport Index (CUTI). b, Biologically Effective Upwelling Transport Index (BEUTI). c, regionally available nitrate time for the studied time period (2014-2020). Lines are 2-month moving averages and the colors palettes represent three distinct time periods, 2014-2016 (red), 2017-2018 (blue), and 2019-2020 (gold). CUTI (m2 s-1) is a regionally integrated rate of vertical volume transport. BEUTI (µM m-1 s-1) is an estimate of nitrate flux into the surface mixed layer. Regionally available nitrate (µM) is the concentration of nitrate at the base of the mixed layer and can be calculated by dividing BEUTI by CUTI (see Jacox et al. 2018 for a full explanation). d, mean nitrate (µM) concentrations at 10 m depth at each CalCOFI station during the warm period (2014-2016). e, mean nitrate concentrations during the cool period. f, the difference in nitrate concentrations between the two phases (2014-2016) – (2017-2018). g-k, maps of the mean Bray-Curtis similarity (Legendre and Legendre 2012) between samples from the warm (2014-2016) and cool (2017-2018) phase for each station. Maps show surface samples for our five main groups (g Archaea, h Bacteria, i Cyanobacteria, j Photosynthetic Eukaryotic Protists, and k Heterotrophic Eukaryotic Protists). For DCM samples, see Fig. 11S.

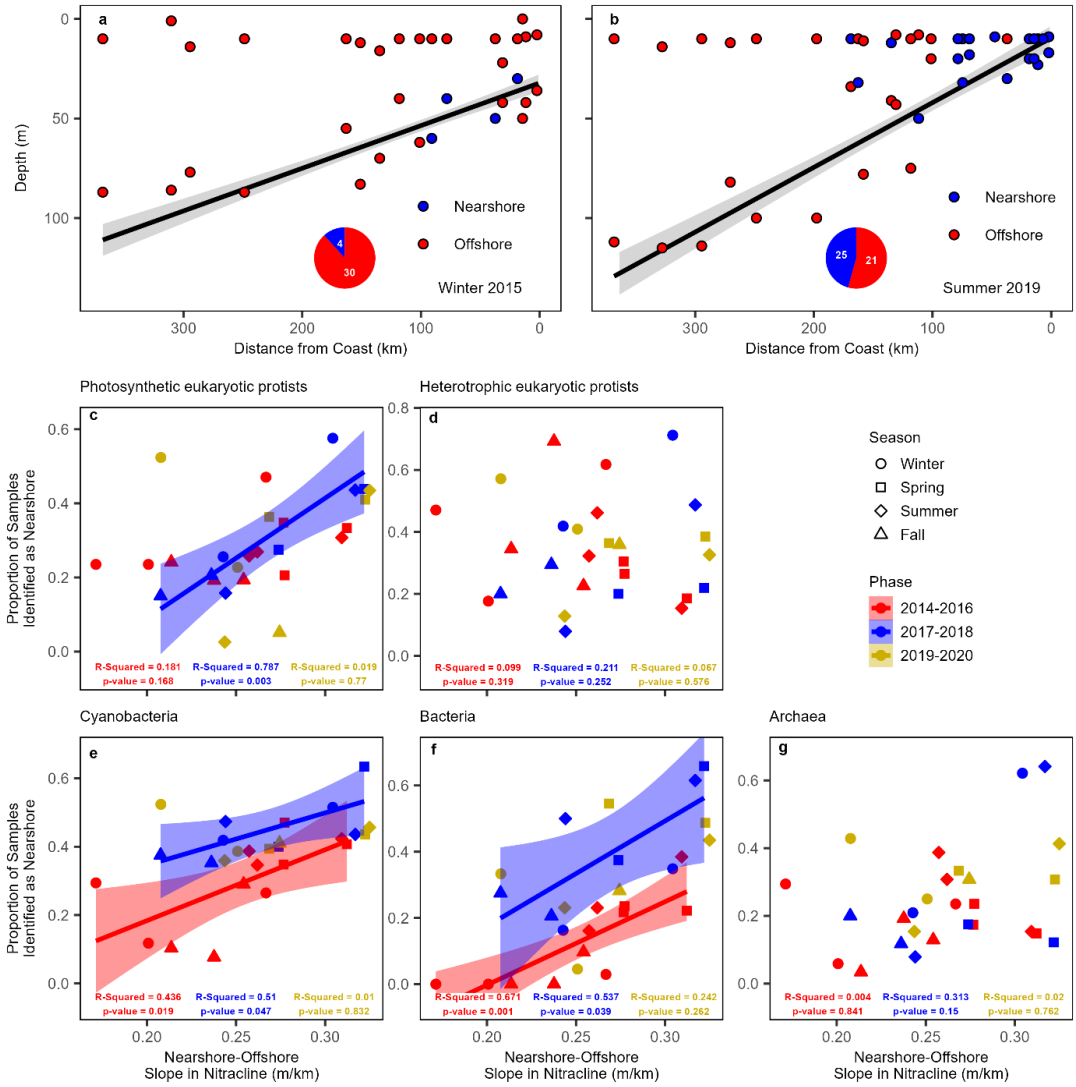


**Figure 2.6:** Temporal shifts in regional nitracline gradients align with relative community dominance. a, illustrative example highlighting a cruise (Winter 2015) where the regional slope in nitracline depth is relatively low. The black line indicates the regional slope in the nitracline depth with a 95% confidence interval around the model fit (glm). Points indicate individual samples taken during this cruise. The color of the points indicates cyanobacteria communities that were identified by SOMs as either "nearshore" (blue) or "offshore" (red). b, illustrative example highlighting a cruise (Summer 2019) where the regional slope in nitracline depth is much greater. The black line indicates the regional slope in the nitracline depth with a 95% confidence interval around the model fit (glm). Points indicate individual samples taken during this cruise. The color of the points indicates cyanobacteria communities that were identified by SOMs as either "nearshore" (blue) or "offshore" (red). c-g, proportion of samples per cruise that were identified by SOMs as "nearshore" communities relative to the slope in the nitracline across the entire region. Shapes represent the different seasons during which cruises took place (circle = winter, square = spring, diamond = summer, triangle = fall) and the colors represent samples that were collected from 2014-2016 (red), 2017-2018 (blue), or 2019-2020 (gold). Data were fitted as separate linear models per phase, shading represents the 95% confidence interval around the model fit.

## 2.6 Supplementary Information

**Table 2.1S:** List of groups used in analysis. Groups listed in bold and shaded in grey are the key functional groups used for the main analysis. Finer taxonomic groups are listed below the broad groups.

| Group | # of ASVs | Amplicon Region |
|---|---|---|
| **Archaea** | 621 | 16S |
| **Heterotrophic Bacteria** | 17142 | 16S |
| **Cyanobacteria** | 511 | 16S |
| **Photosynthetic Eukaryotic Protists** | 7770 | 18S |
| **Heterotrophic Eukaryotic Protists** | 24311 | 18S |
| *Prochlorococcus* | 224 | 16S |
| *Synechococcus* | 40 | 16S |
| *Flavobacteriales* | 1559 | 16S |
| *Rhodobacterales* | 476 | 16S |
| SAR 11 Clade | 873 | 16S |
| Diatoms | 620 | 18S |
| Dinoflagellates (without Syndiniales) | 4494 | 18S |
| Syndiniales | 5698 | 18S |
| Haptophytes | 483 | 18S |
| Chlorophytes | 1056 | 18S |
| Metazoans | 1943 | 18S |

**Table 2.2S:** Composition of 18Sv9 ASVs endemic to each dataset (columns 2-4) or found in multiple datasets (columns 5-8). Zeros represent groups where no ASVs are found within a given dataset or dataset overlap.

| Division | NCOG | Tara Oceans | Tara Polar | NCOG + Tara Oceans | NCOG + Tara Polar | Tara Oceans + Tara Polar | All Datasets |
|---|---|---|---|---|---|---|---|
| Alveolata | 3 | 12 | 1 | 4 | 0 | 2 | 3 |
| Apicomplexa | 128 | 1631 | 277 | 136 | 4 | 121 | 27 |
| Apusomonadidae | 2 | 26 | 25 | 3 | 0 | 5 | 2 |
| Breviatea | 0 | 3 | 11 | 0 | 0 | 3 | 1 |
| Centroheliozoa | 8 | 78 | 33 | 15 | 0 | 1 | 8 |
| Cercozoa | 104 | 988 | 682 | 191 | 25 | 99 | 120 |
| Chlorophyta | 528 | 2407 | 248 | 419 | 12 | 66 | 77 |
| Choanoflagellida | 19 | 136 | 76 | 25 | 4 | 5 | 29 |
| Ciliophora | 349 | 3447 | 1155 | 589 | 46 | 572 | 300 |
| Conosa | 0 | 26 | 22 | 0 | 0 | 2 | 2 |
| Cryptophyta | 8 | 145 | 72 | 18 | 1 | 7 | 15 |
| Dinoflagellata | 3491 | 21790 | 6626 | 5617 | 124 | 977 | 1993 |
| Discoba | 543 | 8304 | 9934 | 172 | 11 | 294 | 224 |
| Foraminifera | 13 | 949 | 265 | 14 | 2 | 48 | 8 |
| Fungi | 52 | 1145 | 403 | 46 | 7 | 97 | 65 |
| Haptophyta | 98 | 559 | 153 | 269 | 4 | 32 | 109 |
| Katablepharidophyta | 3 | 43 | 21 | 6 | 0 | 2 | 6 |
| Lobosa | 27 | 451 | 142 | 20 | 1 | 33 | 11 |
| Mesomycetozoa | 3 | 24 | 3 | 4 | 0 | 6 | 3 |
| Metazoa | 1113 | 26431 | 3103 | 560 | 41 | 274 | 188 |
| Ochrophyta | 420 | 5769 | 2211 | 528 | 30 | 161 | 288 |
| Opalozoa | 107 | 685 | 147 | 147 | 12 | 38 | 90 |
| Opisthokonta | 1 | 51 | 1 | 13 | 0 | 2 | 4 |
| Picozoa | 12 | 197 | 71 | 7 | 1 | 4 | 17 |
| Prasinodermophyta | 2 | 1 | 0 | 3 | 1 | 0 | 2 |
| Pseudofungi | 40 | 329 | 197 | 40 | 2 | 16 | 30 |
| Radiolaria | 324 | 11290 | 956 | 869 | 16 | 129 | 338 |
| Rhodophyta | 23 | 347 | 8 | 40 | 0 | 9 | 11 |
| Sagenista | 46 | 663 | 204 | 106 | 6 | 40 | 66 |
| Stramenopiles | 0 | 16 | 9 | 1 | 0 | 2 | 4 |
| Streptophyta | 46 | 249 | 79 | 28 | 1 | 11 | 9 |
| Telonemia | 24 | 189 | 86 | 39 | 2 | 11 | 46 |
| Eukaryota | 1 | 5 | 3 | 2 | 0 | 0 | 0 |
| Hilomonadea | 2 | 10 | 7 | 3 | 0 | 0 | 0 |
| Mantamonadidea | 0 | 2 | 0 | 0 | 0 | 0 | 0 |
| Other | 7350 | 93539 | 25186 | 3202 | 133 | 4637 | 947 |
| Total | 14890 | 181937 | 52417 | 13136 | 486 | 7706 | 5043 |

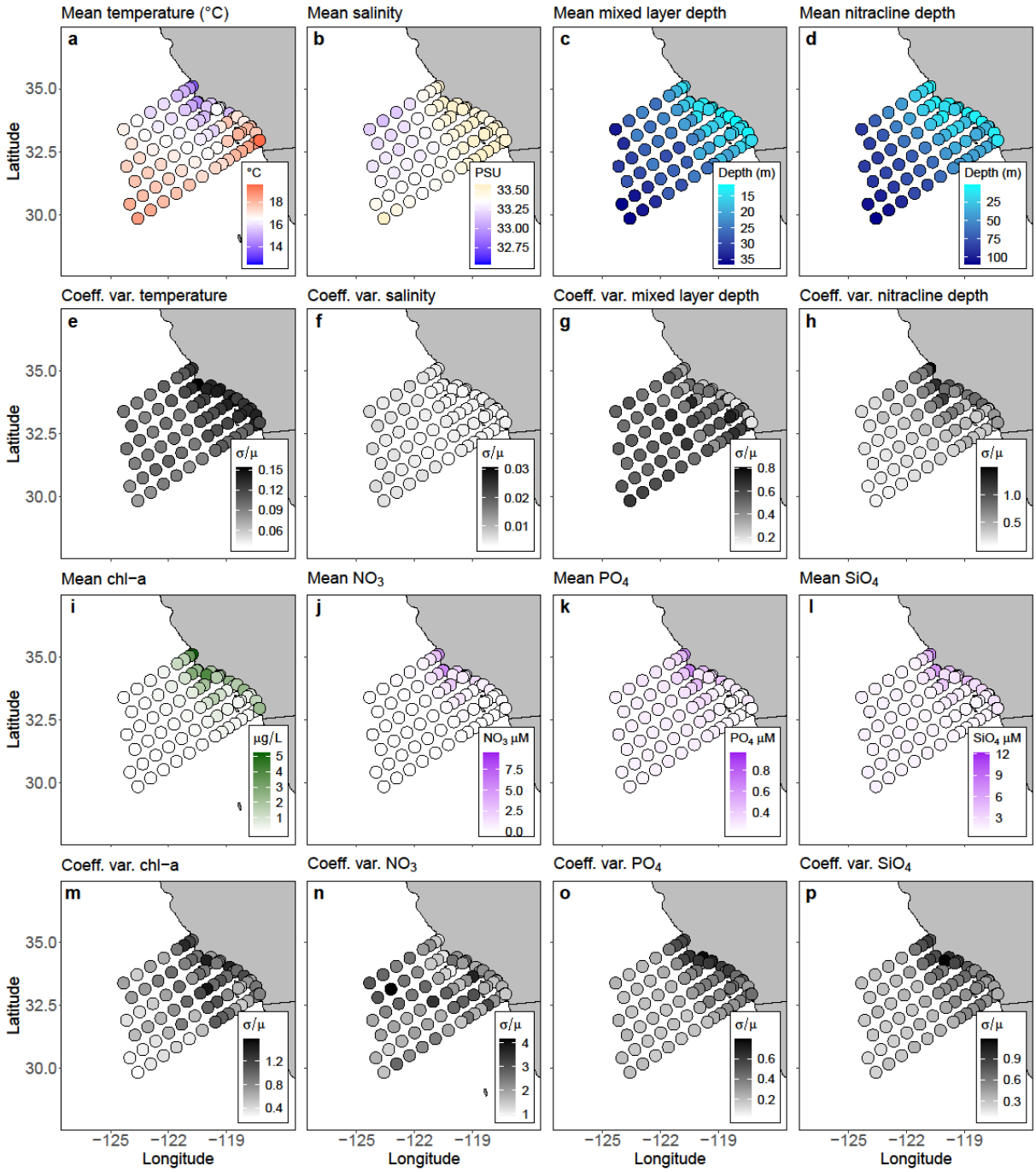


**Figure 2.1S:** Mean spatial gradients in physical and ecological variables shown in color: (a) temperature (°C), (b) salinity (PSU), (c) mixed layer depth (m), (d) nitracline depth (m), (i) chlorophyll a (µg/L), (j) nitrate (µM), (k) phosphate (µM), and (l) silicate (µM). Spatial gradients in the coefficient of variation (Coeff. var.) are shown in grayscale for: (e) temperature, (f) salinity, (g) mixed layer depth, (h) nitracline depth, (m) chlorophyll a, (n) nitrate, (o) phosphate, and (p) silicate.

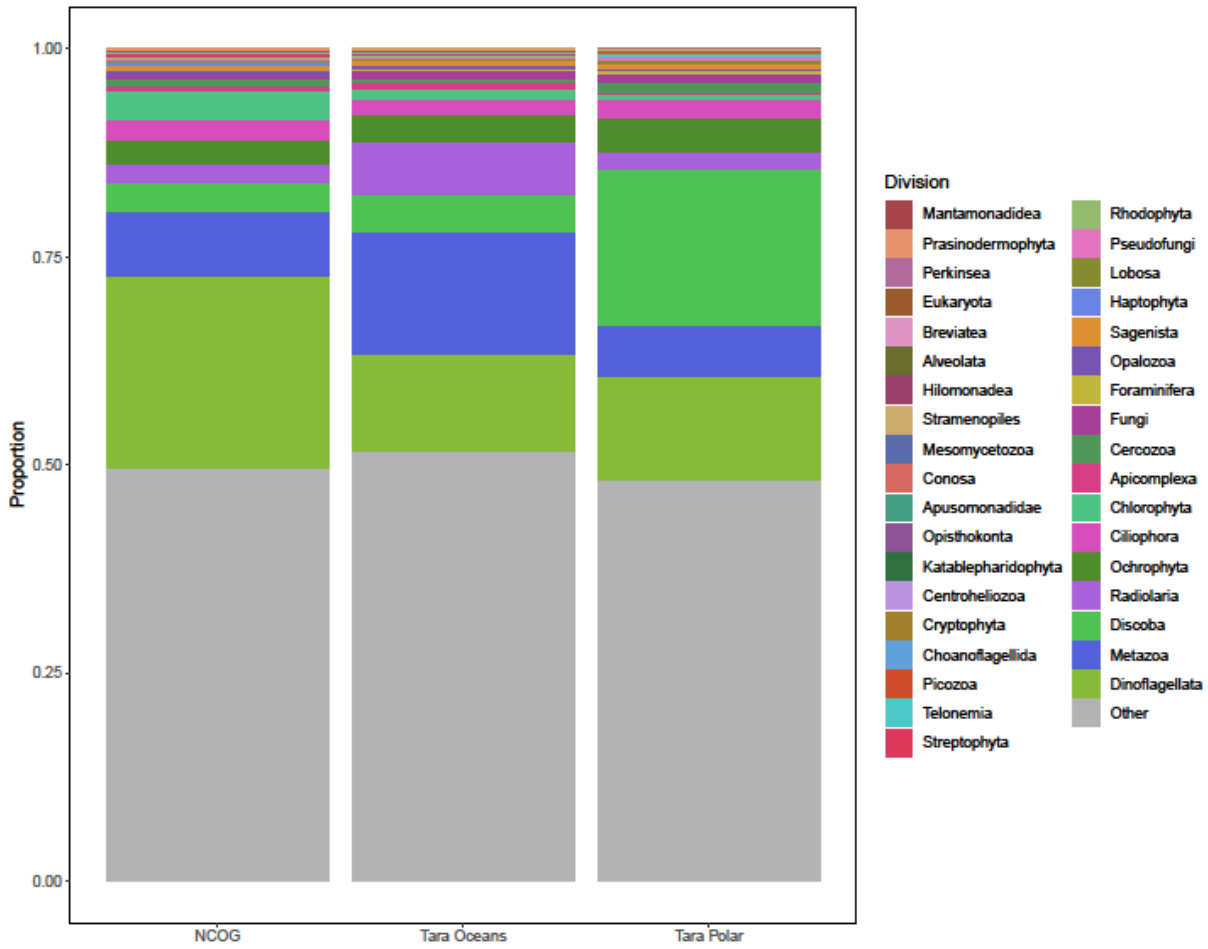


**Figure 2.2S:** Proportional taxonomic composition of ASVs that are endemic to NCOG, TARA Oceans, and TARA Polar. Colors represent the proportional dominance of broad taxonomic groups within the ASVs that are endemic to each dataset. Total number of endemic ASVs per dataset can be found in Fig. 1e.

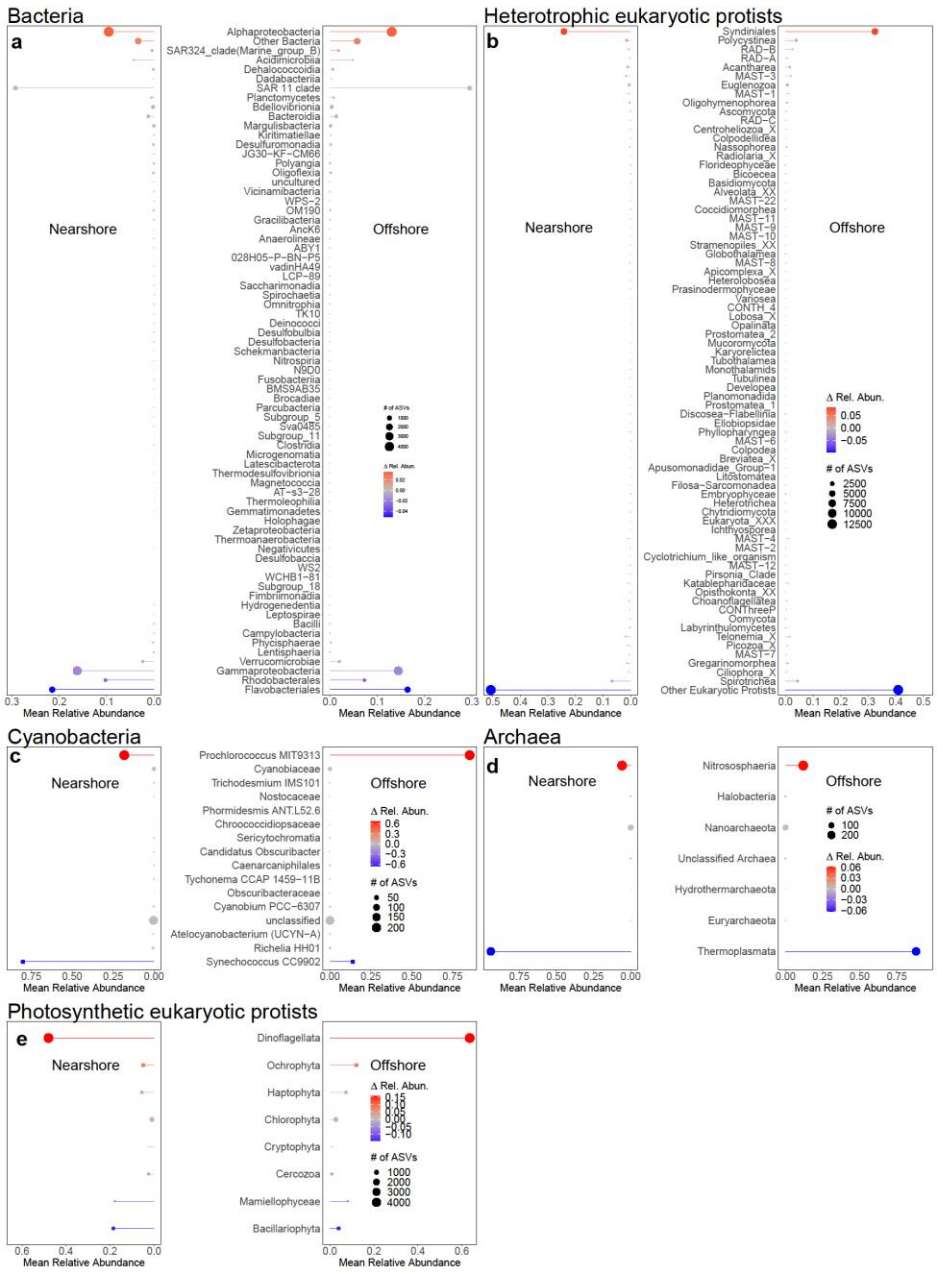


**Figure 2.3S:** Mean relative abundance of taxonomic groups in both the nearshore and offshore clusters for major groups: (a) bacteria, (b) heterotrophic eukaryotic protists, (c) cyanobacteria, (d) archaea, and (e) photosynthetic eukaryotic protists. Mean relative abundance is calculated as the mean abundance of all ASVs within a taxonomic group per cluster. Taxonomic groups are ordered, from top to bottom, by their difference in mean relative abundance within the offshore versus the nearshore: Δ Mean Relative Abundance = Mean Offshore Relative Abundance – Mean Nearshore Relative Abundance. A positive difference (red) indicates the mean relative abundance is greater in the offshore and a negative difference (blue) indicates the mean relative abundance is greater in the nearshore. The size of the circles represents the number of ASVs found within each taxonomic group. Listed taxonomic groups include those that were most abundant or are thought to be ecologically important. The rest of the ASVs are included in the "Other" categories found in each subplot.

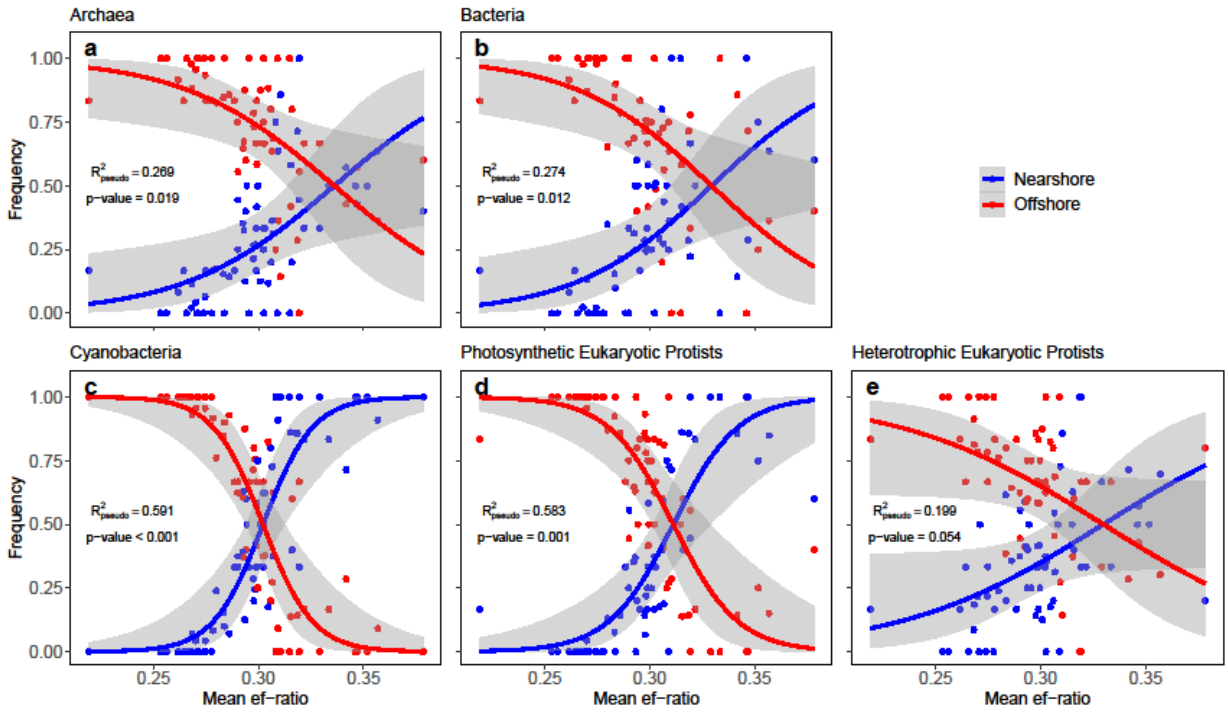


**Figure 2.4S:** Relationship between ef-ratio (calculated from Eq. 2 of Laws, D'Sa, and Naik 2011) and the frequency of SOM clusters per station from 2014-2020 for the five major taxonomic groups: (a) archaea, (b) bacteria, (c) cyanobacteria, (d) photosynthetic eukaryotic protists, and (e) heterotrophic eukaryotic protists. Cragg and Uhler's pseudo $R^2$ was used to assess the goodness of fit[95] between mean ef-ratio and frequency. Shading represents the 95% confidence interval around each model fit.

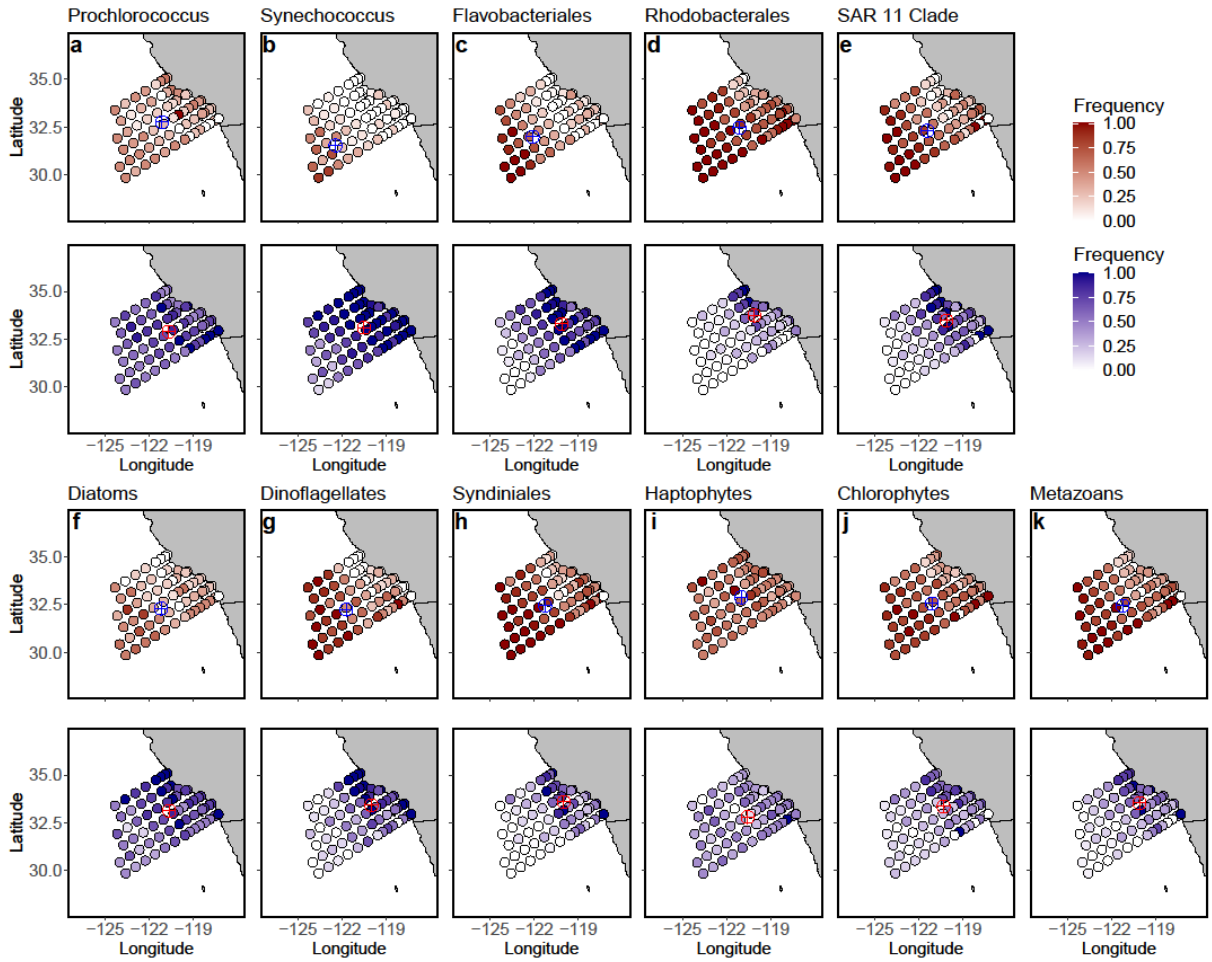


**Figure 2.5S:** Nearshore and offshore gradients in community structure within the eleven taxonomic groups: (a) Prochlorococcus, (b) Synechococcus, (c) Flavobacteriales, (d) Rhodobacteriales, (e) SAR 11 Clade, (f) Diatoms, (g) Dinoflagellates, (h) Syndiniales, (i) Haptophytes, (j) Chlorophytes, and (k) Metazoans. Colors indicate the frequency that the community at each location is offshore (red, top row) or nearshore (blue, bottom row) in character. The designation of nearshore vs. offshore community is determined by the cluster whose weighted centroid is closer to the coast. The weighted centroid for each cluster is shown as a target in the opposite color.


**Figure 2.6S:** Relative importance of all explanatory variables (mean and coefficient of variation) used to predict the frequency of offshore vs. nearshore clusters at a given station for taxonomic specific groups. Relationships were assessed via a generalized linear model with a binomial fit. Larger circles represent lower AIC values within a column. Relationships that are not significant ($p > 0.05$) are colored grey. Circles and their associated AIC values should not be compared across columns. Relationships were analyzed between the frequency of observed community clusters and the mean and coefficient of variation (Coeff. Var.) of environmental variables. Environmental variables included: temperature (Temp), salinity, $NO_3$, $PO_4$, $SiO_4$, chlorophyll a (Chl-a), and nitracline depth (NCD).

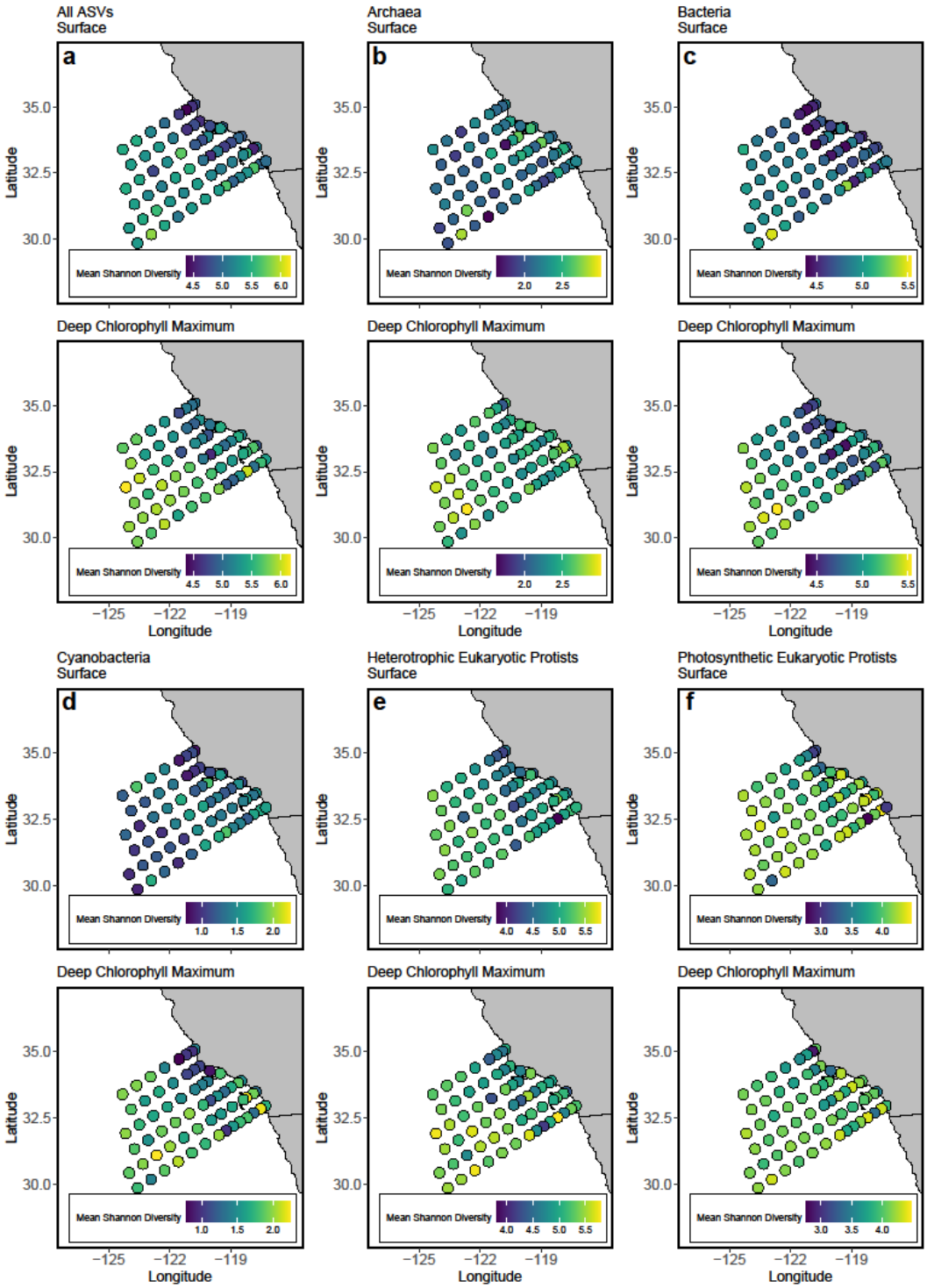


**Figure 2.7S:** Mean alpha diversity for (a) all ASVs and the five major groups: (b) archaea, (c) bacteria, (d) cyanobacteria, (e) heterotrophic eukaryotic protists, and (f) photosynthetic eukaryotic protists at each CalCOFI station in both surface samples (top panel) and deep chlorophyll maximum samples (bottom panel) across all years (2014-2020). Scale bars for each group are consistent across depths (surface and deep chlorophyll maximum).

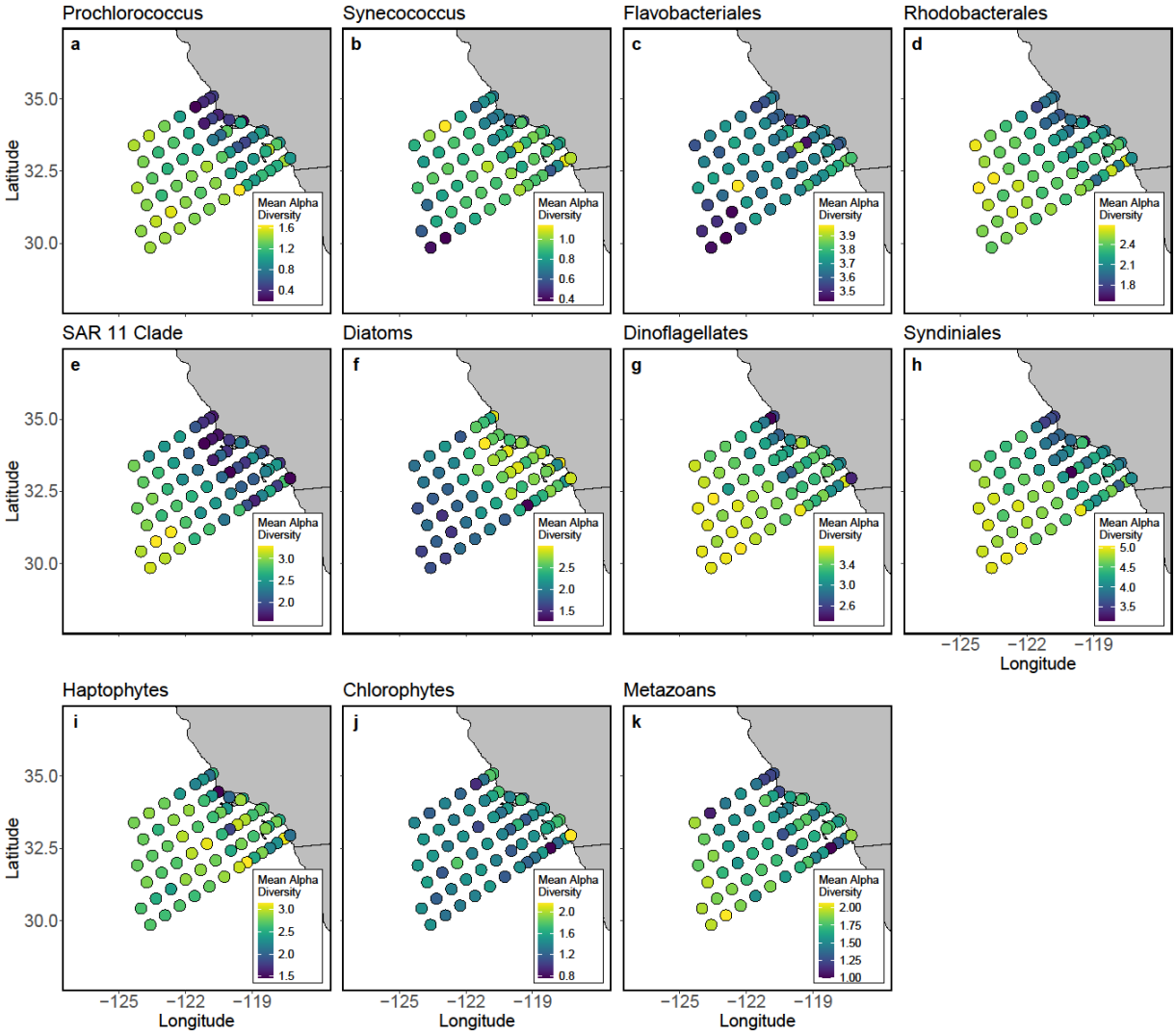


**Figure 2.8S:** Mean alpha diversity for the eleven taxonomic groups: (a) Prochlorococcus, (b) Synechococcus, (c) Flavobacteriales, (d) Rhodobacteriales, (e) SAR 11 Clade, (f) Diatoms, (g) Dinoflagellates, (h) Syndiniales, (i) Haptophytes, (j) Chlorophytes, and (k) Metazoans at each CalCOFI station across all samples and years (2014-2020).

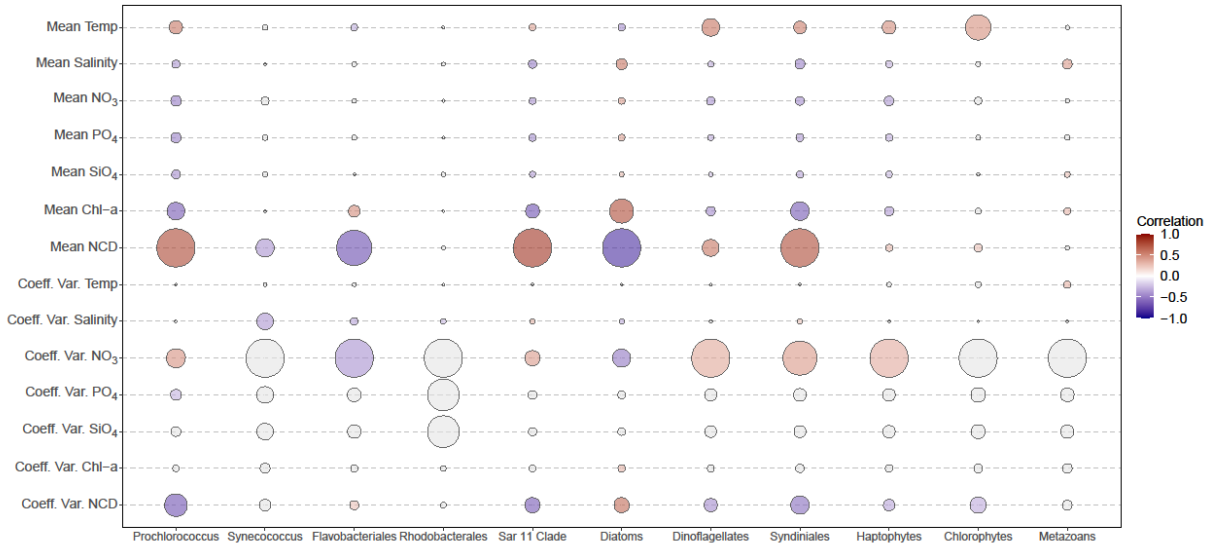


**Figure 2.9S:** Relative importance of all explanatory variables (mean and coefficient of variation) as used to predict mean alpha diversity at a given station for the finer taxonomic groups. Relationships were assessed via a generalized linear model with a gaussian fit. Larger circles represent lower AIC values within a column. Circles and their associated AIC values should not be compared across columns. Color represents the correlation coefficient between each explanatory variable and mean alpha diversity. Gray circles represent relationships that are not significant ($p > 0.05$). Shannon index was used as the primary measure of diversity and was calculated as the mean per station per cruise for this analysis. Relationships between environmental variables and diversity were assessed via a generalized linear model with a gaussian fit. Larger circles represent lower AIC values within a column. Circles and their associated AIC values should not be compared across columns. Color represents the correlation coefficient between each explanatory variable and mean alpha diversity. Gray circles represent relationships that are not significant ($p > 0.05$). Relationships were analyzed between diversity and the mean and coefficient of variation (Coeff. Var.) of environmental variables. Environmental variables included: temperature (Temp), salinity, $NO_3$, $PO_4$, $SiO_4$, chlorophyll a (Chl-a), and nitracline depth (NCD).

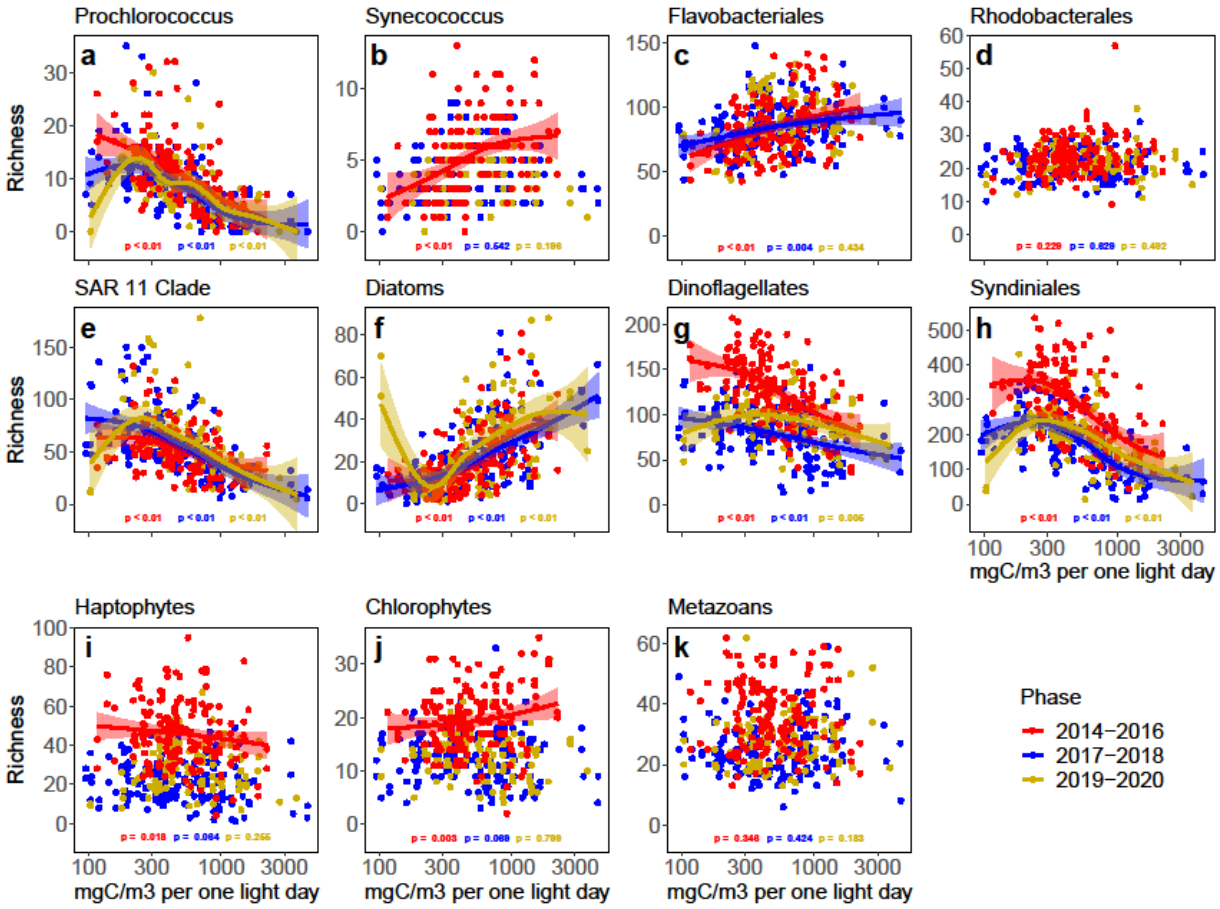


**Figure 2.10S**: Productivity-diversity relationship for all eleven taxonomic groups: (a) Prochlorococcus, (b) Synechococcus, (c) Flavobacteriales, (d) Rhodobacteriales, (e) SAR 11 Clade, (f) Diatoms, (g) Dinoflagellates, (h) Syndiniales, (i) Haptophytes, (j) Chlorophytes, and (k) Metazoans. The data was subset to include only productivity station samples where $^{14}C$ was measured. Productivity-diversity relationships were fit with a generalized additive model (GAM). Significant relationships are denoted by a red line (2014-2016), blue line (2017-2018), or gold line (2019-2020). Lines represent the best GAM fit with shading representing a 99% confidence interval. Richness (total number of ASVs) rather than Shannon Index is used in this figure as this is the standard for fitting productivity-diversity relationships (Vallina et al. 2014; Chase and Leibold 2002).

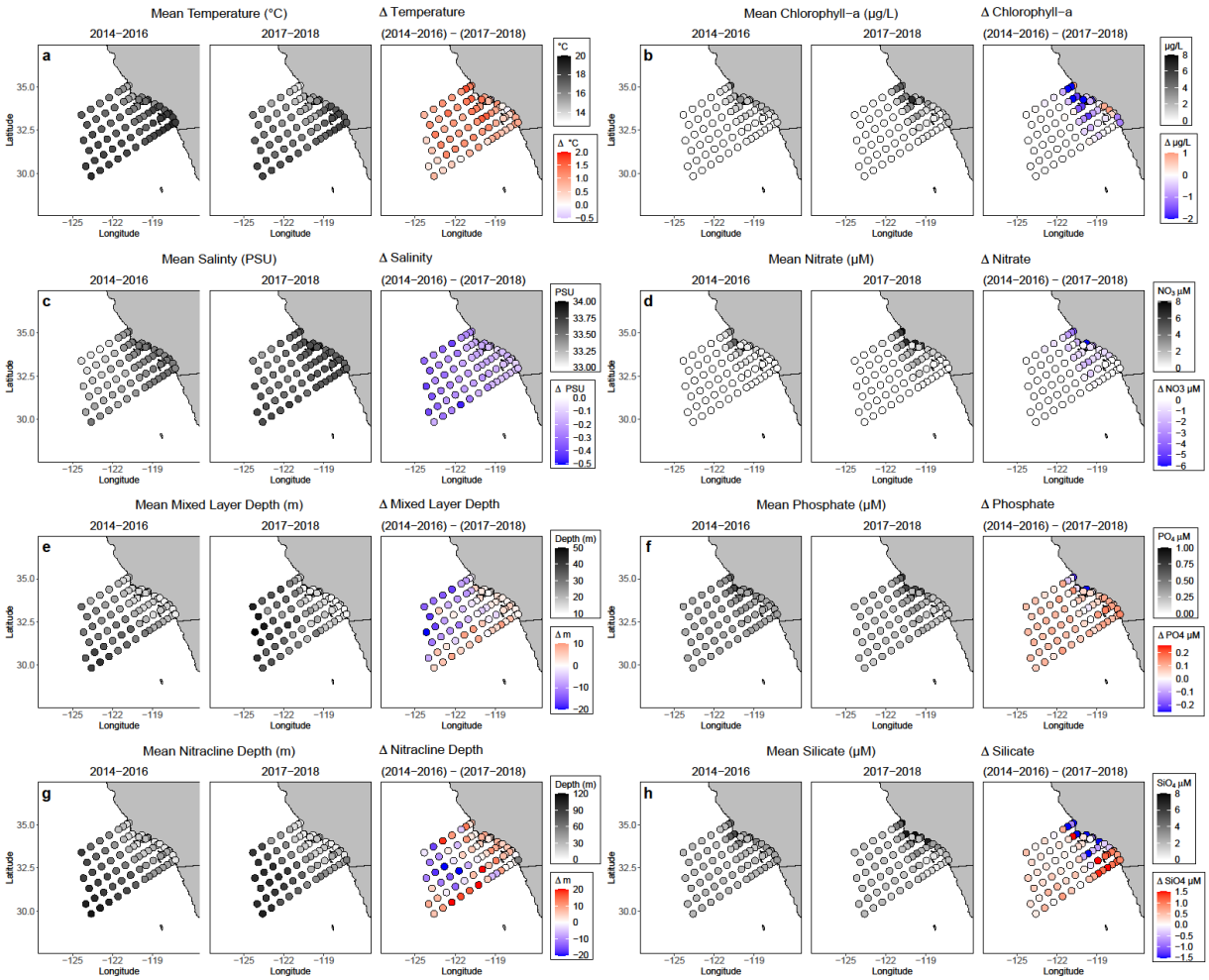


**Figure 2.11S:** Mean spatial gradients of physical and ecological variables in 2014-2016 and 2017-2018 are shown in grayscale. The difference between the two time periods is shown in color. Variables include: (a) temperature (°C), (b) salinity (PSU, practical salinity units), (c) mixed layer depth (m), (d) nitracline depth (m), (e) chlorophyll *a* (µg/L), (f) nitrate (µM), (g) phosphate (µM), and (h) silicate (µM)

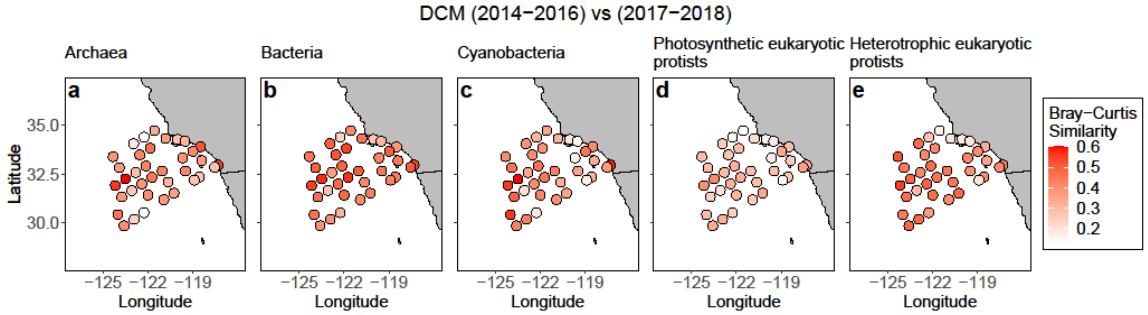


**Figure 2.12S:** Maps of the mean Bray-Curtis similarity (Legendre and Legendre 2012) between deep chlorophyll maximum (DCM) samples from the warm (2014-2016) and cool (2017-2018) phase for each station for our five main groups: (a) archaea, (b) bacteria, (c) cyanobacteria, (d) photosynthetic eukaryotic protists, and (e) heterotrophic eukaryotic protists.

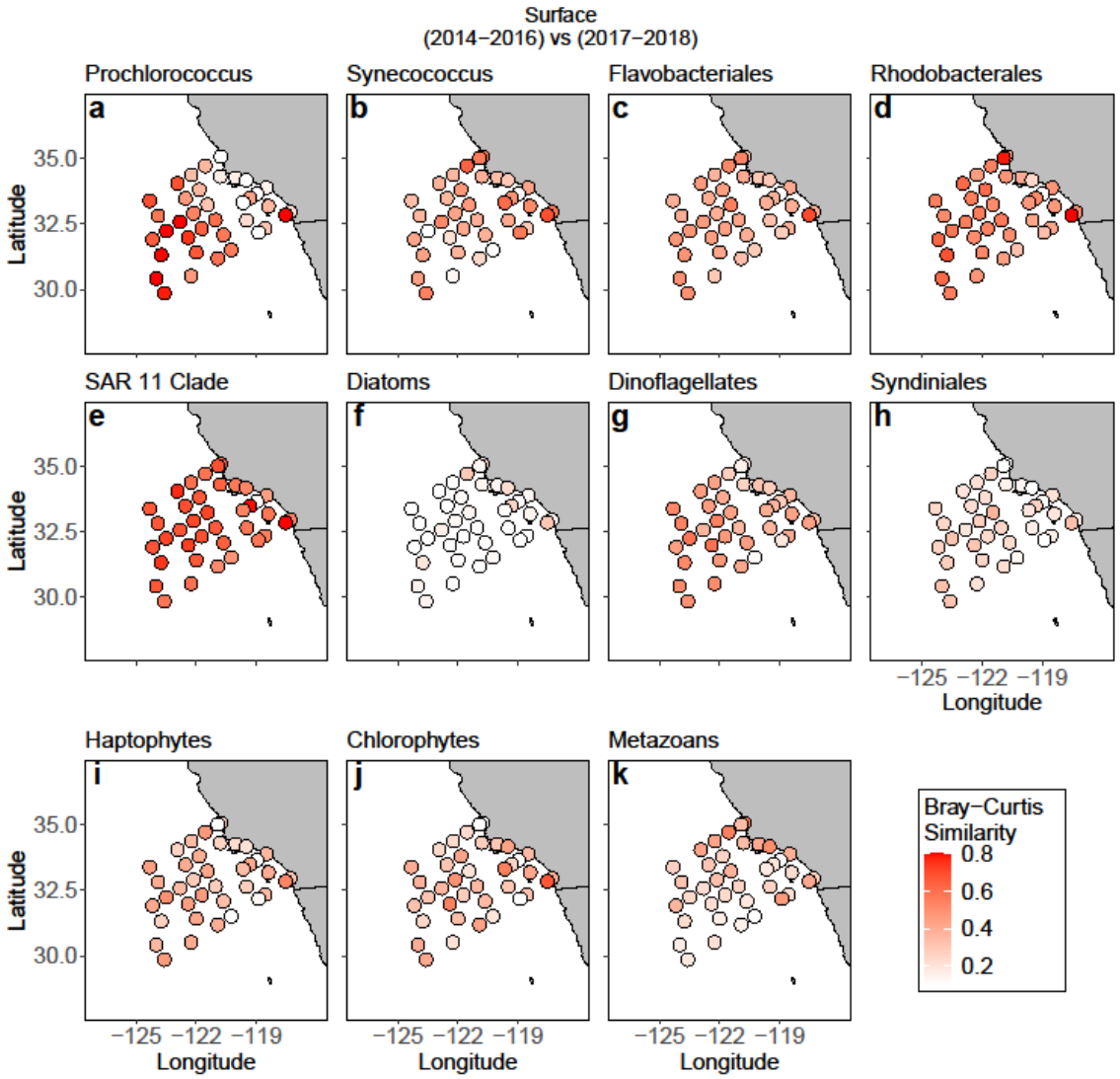


**Figure 2.13S:** Maps of the mean Bray-Curtis similarity (Legendre and Legendre 2012) between surface samples from the warm (2014-2016) and cool (2017-2018) phase for each station. Maps show surface samples for our eleven taxonomic groups: (a) Prochlorococcus, (b) Synechococcus, (c) Flavobacteriales, (d) Rhodobacteriales, (e) SAR 11 Clade, (f) Diatoms, (g) Dinoflagellates, (h) Syndiniales, (i) Haptophytes, (j) Chlorophytes, and (k) Metazoans.

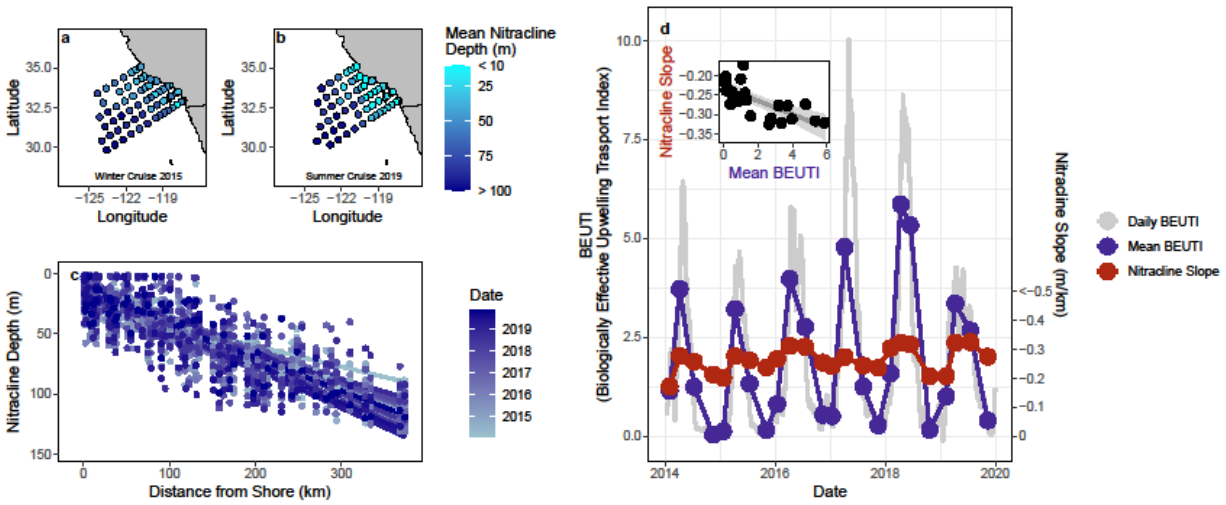


**Figure 2.14S:** Examples of cruises with variable regional nitracline slopes. (a), winter 2015 where there was a shallow regional nitracline slope (similar nitracline depth across the entire region). (b), summer 2019 where there was strong nearshore upwelling (shallow nitracline) contrasted with a deep nitracline far offshore (c), Regional slope of the nitracline for each cruise. Color of the points and lines represents the mean date for a given cruise. Slopes were fit with a generalized linear model (d), Relationship between Biological Effective Upwelling Transport Index (BEUTI) and the regional nitracline slope. Grey line shows the daily BEUTI values while the blue line shows a 3-month moving average centered around the mean cruise date (points). Red line shows the nitracline slope for each cruise. The inlaid plot shows the correlation between Mean BEUTI and Nitracline Slope where higher BEUTI values correlated with steeper slopes in the nitracline.

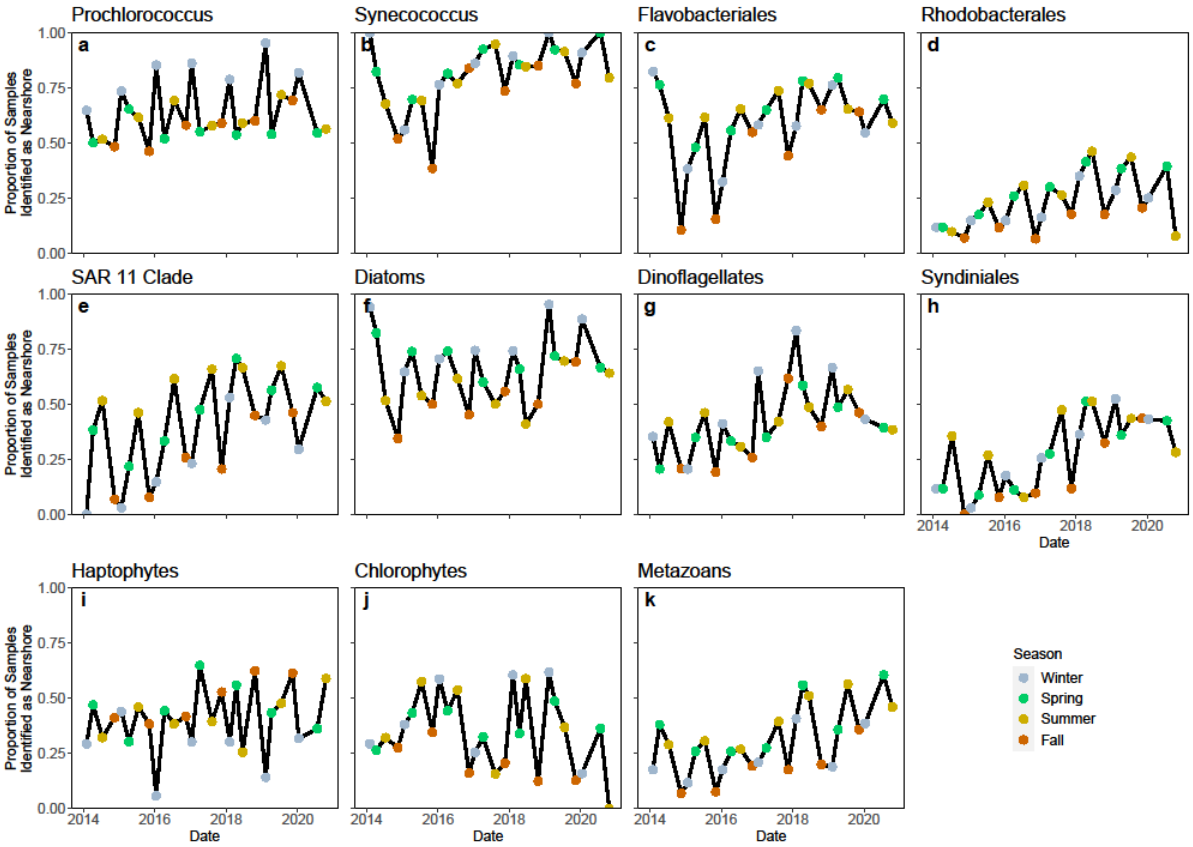


**Figure 2.15S:** Time series illustrating the proportion of samples per cruise that were identified as "nearshore" communities. Points are colored based on the season during which each cruise took place. Panel represent each of our eleven taxonomic groups: (a) Prochlorococcus, (b) Synechococcus, (c) Flavobacteriales, (d) Rhodobacteriales, (e) SAR 11 Clade, (f) Diatoms, (g) Dinoflagellates, (h) Syndiniales, (i) Haptophytes, (j) Chlorophytes, and (k) Metazoans.

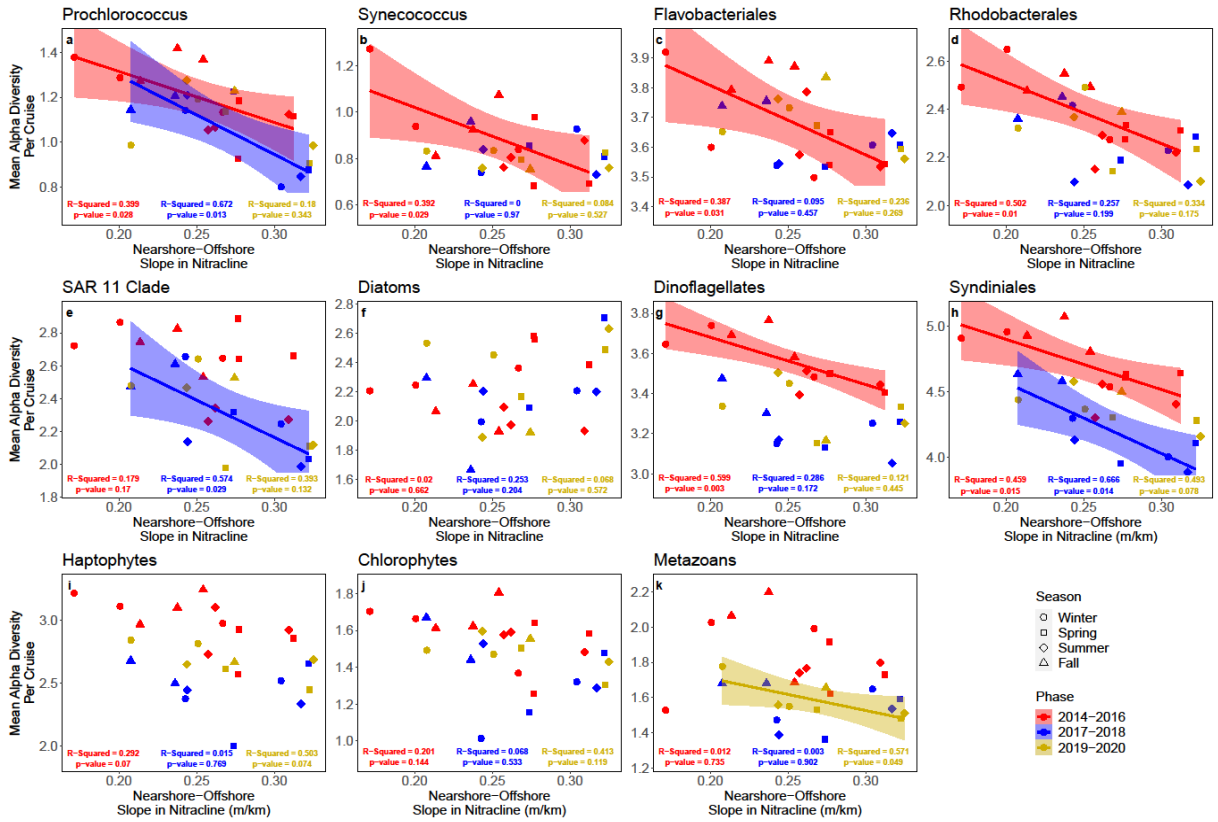


**Figure 2.16S:** Mean alpha diversity in relation to the slope in the nitracline depth across the entire region. Shapes represent the different seasons during which cruises took place (circle = winter, square = spring, diamond = summer, triangle = fall) and the colors represent samples that were collected from 2014-2016 (red), 2017-2018 (blue), or 2019-2020 (gold). Data were fitted as separate linear models per phase. Shading represents the 95% confidence interval around the model fit. Panels represent each of our eleven taxonomic groups: (a) Prochlorococcus, (b) Synechococcus, (c) Flavobacteriales, (d) Rhodobacteriales, (e) SAR 11 Clade, (f) Diatoms, (g) Dinoflagellates, (h) Syndiniales, (i) Haptophytes, (j) Chlorophytes, and (k) Metazoans.

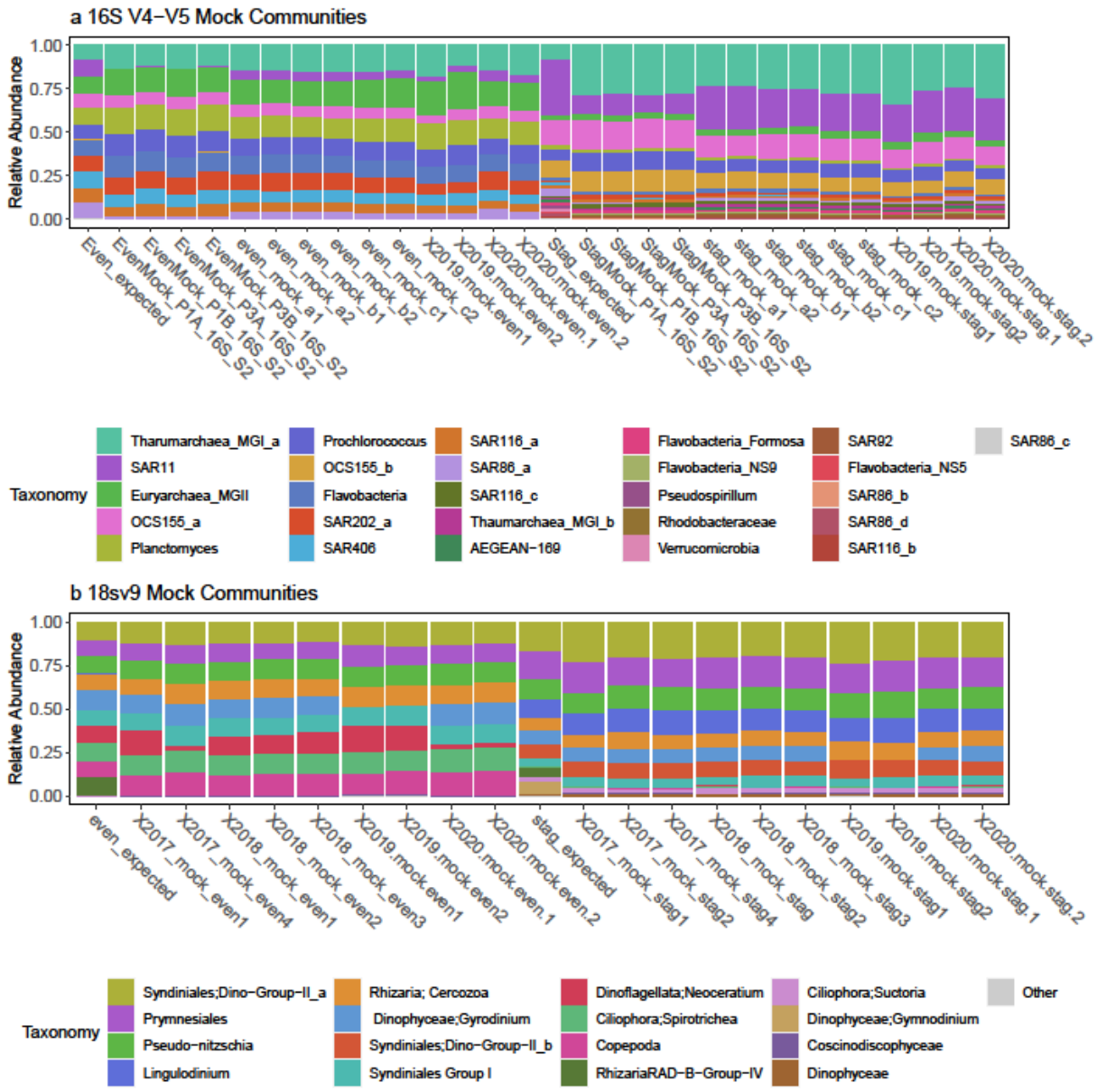


**Figure 2.17S:** Even and staggered mock communities for: (a), 16S and (b), 18S. Custom mock communities were the same as those found in Parada et al. 2016 (Parada, Needham, and Fuhrman 2016). SAR86_c is expected in our staggered community (0.002 expected relative abundance) but is not seen in any mock community samples.

## 2.7 Acknowledgements

CCJ acknowledges graduate student support by Scripps Institution of Oceanography.

This study was supported by National Science Foundation, California Current Ecosystem Long Term Ecological Research Grants, CCE-LTER Phase II and III (NSF-OCE-1026607 and NSF-

OCE-1637632), and NSF-OCE-1756884, NOAA (NOAA OAR Omics, CIMEC NA15OAR4320071, and ECOHAB NA19NOS4780181), and Gordon and Betty Moore Foundation grants GBMF3828 to AEA.

We would like to acknowledge former CalCOFI director David M. Checkley and Margot Bohan from the NOAA Office of Ocean Exploration and Research (OER) for their vision and guidance during the initial phase of the NCOG program and current CalCOFI director Brice X. Semmens for his continued support. We are also especially grateful to California Current Ecosystem, Long Term Ecological Research (CCE-LTER) and CalCOFI project and team members and crew who have assisted with the NCOG program 2014-present.

Chapter 2, in full, is a reprint of the material as it appears in James, C. C., Barton, A. D., Allen, L. Z., Lampe, R. H., Rabines, A., Schulberg, A., Zheng, H., Goericke, R., Goodwin, K. D., Allen, A. E. (2022). Influence of nutrient supply on plankton microbiome biodiversity and distribution in a coastal upwelling region. Nature Communications. The dissertation author was the primary investigator and author of this paper.

# Chapter 3 - Endemism, cosmopolitanism, and habitat specificity within a coastal marine microbiome


## Abstract

The prevalence of endemic and cosmopolitan taxa, as well as the habitat specificities of taxa within the marine microbiome are not well known. Here, using a seven-year record of prokaryotic and eukaryotic metabarcodes spanning 445 samples in the Southern California Current region, we quantify the proportion of taxa exhibiting endemic and cosmopolitan distributions in this region, as well as identify the general characteristics of habitat affinity within marine microbes. We find that the majority of taxa occupy a space between endemism and cosmopolitanism, occurring in some but not all habitats. These taxa also tend to have no habitat affinities and are relatively rare. Approximately 10% of taxa were significantly endemic while around 30% were cosmopolitan. Perhaps surprisingly most cosmopolitans while dispersed across all habitats, had an affinity to one habitat. Compared to terrestrial microbiomes, most marine microbes are not endemic to the habitat level, however they are also not cosmopolitan. Rather, most marine microbes appear to be randomly distributed and moderately rare.


## 3.1 Introduction

Marine microbes comprise the base of the marine food web and are responsible for many environmental functions with local to global effects such as primary productivity, nutrient recycling, and carbon sequestration (Field et al. 1998; Falkowski et al. 2008; Not et al. 2012). These processes and their relative local magnitude are determined by the community composition within the marine microbiome, which ultimately is shaped by the distribution of individual species. The distribution of species are shaped in relative degrees by four major ecological forces: speciation, selection, dispersal, and drift (Vellend 2010). Within the pelagic

environment, dispersal, driven by the movement of water within surface currents, is thought to play a major role in microbial distributions (Villarino et al. 2018), leading many to suggest that most marine microbes are globally distributed (Fenchel & Finlay 2004; Gibbons et al. 2013).

Species distributions fall along a wide spectrum. Within the ocean, bacterial species, such as those belonging to the SAR 11 Clade are thought to have ubiquitous distributions and are likely to occur in nearly every genomic sample of the marine environment (Morris et al. 2002). Species with global distributions, but not to the extreme of ubiquitous species, are known as cosmopolitans. Within the marine environment, cosmopolitans are thought to occur most frequently within microbes or large mobile megafauna (Costello et al. 2017). Endemic species, which fall on the opposite side of the distribution spectrum from cosmopolitans, are defined as species found in particular regions or habitats. While cosmopolitanism is thought to be more common within marine microbes, global surveys such as Tara Oceans and Malaspina have found evidence of regional microbial endemism. Within diatoms, rates of endemism were found to range from 2.3% to 53.3% depending on the region (Malviya et al. 2016). Studies of global ciliate diversity found rates of endemism ranging from 8.8% to 16% (Gimmler et al. 2016; Canals et al. 2020). This pattern is in stark contrast to terrestrial microbiomes, where endemism, rather than cosmopolitanism appears to be the norm. Within North American soil fungal communities, endemic species constitute 40% of the species richness on a sample by sample basis, and over 80% of regional species richness (Talbot et al. 2014). It is likely that the differences in habitat structure between terrestrial and marine systems drives these large differences in overall species distribution patterns.

Unlike terrestrial systems, pelagic habitats are not defined by their location, but rather by the physical and chemical makeup of internally consistent water masses, or seascapes. These

seascapes move and mix with one another as the result of currents. Within a region like the Southern California Current (SCC), seascapes expand and contract both seasonally and interannually and mix together chaotically at the submesoscale level (Lévy et al. 2018; Bograd et al. 2019; Martín et al. 2020). While endemism within a terrestrial environment may be defined as those species found in only one location, within the marine environment, locations are rarely representative of consistent habitats. Within the marine environment it is therefore difficult to define what is endemic versus what is cosmopolitan. Furthermore, the mixing of habitats at the submesoscale level can lead to organisms being found in drastically different habitats from their preferred biome.

In this study we aim to address the challenge of comparing rates of endemism and cosmopolitanism in a comparable way to terrestrial systems and ask the following questions: 1) are marine microbes found in preferred habitats and 2) what proportion of microbes are endemic, cosmopolitan, and ubiquitous within a regional context? For this study we utilize a seven-year, survey of marine microbial community composition consisting of 445 surface samples within the Southern California Current (SCC) referred to as the NOAA CalCOFI Ocean Genomics (NCOG) data. The region is characterized by a wide variety of conditions from the highly productive nearshore to the oligotrophic offshore (Fig. 3.1). Seasonal upwelling also drives large-scale changes in community composition and spatial patterning (Bograd & Lynn 2003; Barth et al. 2020; Catlett et al. 2021, James et al. 2022). To explore these questions in a way that was comparable to terrestrial systems, we first had to define consistent habitats within the SCC region. Based on a previous study by Bograd et al. 2019, we identified the three major water masses primarily responsible for the habitats present in the region's surface waters. These three habitats were: the Pacific Subarctic Water (PSUW) which is advected in from the north via the

California Current, the Eastern North Pacific Central Water (ENPCW) which enters the region in its southwestern most corner and is sourced from the subtropical gyre, and finally Pacific Equatorial Water (PEW), which is advected into the surface primarily in the nearshore and arrives to the region via the California Under Current (Bograd et al. 2019).

Following the classification of our samples into these internally consistent habitats we identify the rates of regional endemism (found in one water mass), cosmopolitanism (found in every water mass), or ubiquity (found in every sample). We also assessed what proportion of species had a significant affinity to zero, one, or two habitats within the region—testing whether selection for preferred habitats is a strong driver of regional microbial biogeographies. Finally, we examine to what extent taxa found in NCOG are found elsewhere using data collected by both Tara Oceans and Tara Polar, which provide a global context for our regional analyses. Combined this approach explores the relative effects of dispersal and selection on marine microbial distributions and identifies rates of endemism, cosmopolitanism, and ubiquity in a way that is comparable to previous terrestrial studies.

## 3.2 Results

The following results represent our analysis of a 1,000-member library ensemble (see Methods), and reported values represent the mean value across all members or are shown as a distribution of values across all members. Overall, we observed 13,012 distinct prokaryotic (16S) amplicon sequence variants (ASVs) and 24,737 distinct eukaryotic (18Sv9) ASVs across the 1,000-member ensemble in the Southern California Current region (SCC), of which many were incredibly rare. The majority of observed ASVs across both prokaryotes and eukaryotes were rare in both total abundance and occurrence, where occurrence is defined as either occurring in a sample or station (Fig. 3.2). More than half of the prokaryotic and eukaryotic ASVs had on

average fewer than 10 reads and occurred in less than 3 samples or stations. ASVs that occurred in more samples tended to also be more abundant (Fig. 3.1Sa). In both 16S and 18Sv9, we observed that 50% and 53%, respectively, of ASVs were found at most in only one sample across all 1,000 ensemble members. To better understand these ASVs that were only found in one sample, referred to as singletons, we explored the distribution of reads for singleton ASVs compared to all other ASVs (Fig. 3.1Sb). We found that the mean number of reads for singleton ASVs was 3.12 and 5.30 reads for prokaryotes and eukaryotes, respectively. In contrast, the mean number of reads for all other ASVs was 1,305 and 1,101 for prokaryotes and eukaryotes, respectively. Singletons were removed from the remaining analyses as it did not make sense to test the significance of habitat affinity for ASVs that are seen only once across the 445 samples. Thus, the following analyses were conducted across 6,771 and 12,274 non-singleton prokaryotic and eukaryotic ASVs respectively.

Habitats are typically defined in the ocean by their environmental properties such as temperature, salinity, and nutrients; habitats are constantly in motion within the ocean. Bograd et al. (2019) identified three main pelagic water masses, or habitats, in SCC region: the Pacific Subarctic Upper Water (PSUW; relatively fresh and cool), the East North Pacific Central Water (ENPCW; warm, salty, and nutrient-poor), and the Pacific Equatorial Water (PEW; cool, salty, and nutrient-rich). We used a method called Self-Organizing Maps (SOMs, Kohonen 1997) to cluster our samples into these three water masses using temperature, salinity, and NO3 + NH3 (µM) data (Fig. 3a-b). Our SOM clusters matched the hydrographic and chemical characteristics (temperature, salinity, nutrient concentrations) and spatial distributions of these three water masses described by Bograd et al. (2019).

ASVs were then categorized into four classes based on their occurrence across these three water masses: endemic, cosmopolitan, ubiquitous, and semi-dispersed. Endemic ASVs were defined as ASVs that were only present in one water mass across all 1,000 ensemble members (see Fig. 3.3c for an endemic example). Cosmopolitan ASVs were defined as ASVs that are present in at least one sample of all three water masses across all 1,000 iterations (see Fig. 3.3d for a cosmopolitan example). Ubiquitous ASVs were defined as ASVs that were seen in every sample. Semi-dispersed ASVs occupy the space between endemism and cosmopolitanism and were defined as ASVs that only occur in two water masses or are sometimes seen in one or three water masses but not across all 1,000 iterations.

Following this classification, we tested whether ASVs were significantly overabundant in each water mass, hereafter referred to as habitat affinity. The mean relative abundance per ASV per water mass was compared against a null distribution to assess whether ASVs had a significant affinity towards any given water mass (see Methods for details). In other words, is an ASV relatively more abundant than expected due to random chance in a given water mass? This test was used as both a significance test for endemism (see Fig. 3.3e for an endemic example) and to explore overall habitat affinity for water masses across all ASVs (see Fig. 3.3f for a cosmopolitan example).

Across all taxa we find a wide variety of both occurrence classes, from endemism to ubiquity. 50.0% of prokaryotes and 52% of eukaryotes fell under the semi-dispersed category, as their distributions were neither endemic nor cosmopolitan (Fig. 3.4a-b). Cosmopolitanism was the next most common category with 29.0% of prokaryotic ASVs and 31.7% of eukaryotic ASVs occurring in all three water masses. Rates of endemism were 20.5% for prokaryotes and 16.1%

for eukaryotes. Ubiquity was the rarest categorization with only 3 prokaryotic and 2 eukaryotic ASVs out of 6,771 and 12,274 total ASVs, respectively, found in all samples.

Taxa show various levels of habitat affinity, from no affinity up to an affinity for two of the three water masses (maximum). 55.1% and 60.0% of prokaryotic and eukaryotic ASVs had no significant habitat affinity (Fig. 3.4c-d). A species with no affinity for a particular water mass means that its distribution does not differ significantly from random. The bulk of ASVs with no habitat affinities were from the semi-dispersed category—65.4% for prokaryotes and 64.9% for eukaryotes. Semi-dispersed ASVs with no habitat affinities tend to have a similar rarity to endemics but appear randomly dispersed (Fig. 3.4g-h). Of the remaining ASVs, 43.6% and 39.0% for prokaryotes and eukaryotes respectively had an affinity for one water mass within the region. Only 86 prokaryotic ASVs and 123 eukaryotic ASVs had an affinity for two water masses.

Endemics and cosmopolitans (besides the handful of ubiquitous taxa) sit and the ends of regional distribution patterns, yet these taxa do not all share similar habitat affinity patterns. For both prokaryotes and eukaryotes, roughly 50% of endemics appeared to have a significant habitat affinity for their respective water masses. Thus, the rates of significant habitat-specific endemism were 10.8% and 8% for prokaryotes and eukaryotes. Even though they are found everywhere, the majority of cosmopolitan ASVs had habitat affinities for either one or two water masses. Of the five ubiquitous ASVs, only one had no habitat affinity.

One question that arises is whether there are observed relationships between overall abundance and either occurrence type or the rate of habitat affinity. For both prokaryotes and eukaryotes, there was a significant relationship (nested ANOVA, $p < 0.001$ for both) between overall (across all samples) ASV mean relative abundance and occurrence category (Endemic,

Semi-Dispersed, Cosmopolitan, Ubiquitous). Endemic ASVs were the most rare and ubiquitous ASVs were the most common (Fig. 3.4g-h). There was also a significant nested effect within each category between the number of water mass affinities and overall mean relative abundance for both prokaryotes and eukaryotes (nested ANOVA, $p < 0.001$ for both). ASVs with higher overall relative abundance had affinities for more water masses. Thus, the opposite is also true, ASVs with no water mass affinity, which are most ASVs, tend to be the most rare.

We also explored the spatial distribution of ASVs with habitat affinity across the three distinct water masses (Fig. 3.5). 43.6% and 39% of prokaryotic and eukaryotic ASVs show a habitat preference for one water mass (Fig. 3.5a-b). 1.3% and 1.0% of prokaryotic and eukaryotic ASVS show a habitat preference for two water masses (Fig. 3.5a-b). The majority of ASVs, 55% and 60% of prokaryotes and eukaryotes, show no affinity for any particular water mass (Fig. 3.5a-b). Within prokaryotes (Fig. 3.5a) affinity for PSUW, ENPCW, and PEW was 13.0%, 16.0%, and 14.6%, respectively. In contrast, the proportion of eukaryotic ASVs that had an affinity for the ENPCW (19.4%) was approximately double the proportion of ASVs with affinities for the PSUW (10.9%) and PEW (8.7%). Of the 86 prokaryotic ASVs and 123 eukaryotic ASVs with two water mass affinities the majority had affinities for the PSUW and ENPCW, 72.1% and 90.2% respectively. No ASVs had an affinity for the ENPCW + PEW.

Next, we explored to what extent ASVs that had an affinity to a given water mass dominated relative abundance within each water mass (Fig. 3.5e-h). While ASVs with an affinity for a particular water mass represent a minority of total richness we suspect that they should be more dominant in their preferred habitat. Within the PSUW, ASVs with a habitat affinity for the PSUW or PSUW + ENPCW represented 47.5% and 30.8% of the relative abundance of prokaryotes and eukaryotes, on average per sample (Fig. 3.5c-d). Within the ENPCW, habitat-

specific ASVs comprised 42.6% and 41.8% of the prokaryotic and eukaryotic relative abundance on average per sample (Fig. 3.5e-f). Within the PEW, these values were the highest, as the relative abundance of PEW-affinity prokaryotic and eukaryotic ASVs was 56.5% and 50.2% respectively (Fig. 3.5g-h). The relative abundance of ASVs with no habitat affinity varied between prokaryotes and eukaryotes. Across all water masses, prokaryotic ASVs with no habitat affinity represented on average 5% of the relative abundance per sample. In contrast, eukaryotic ASVs with no habitat affinity represented on average 22% of the relative abundance per sample. This represents a major difference between prokaryotes and eukaryotes and the relative contribution of these randomly distributed ASVs within local ecological communities.

One question that remained was whether there were taxonomic differences between those ASVs that had no affinity to a habitat versus those with an affinity to one or two particular habitats. We found that across all habitat affinity types (No Water Mass Affinities, ENPCW, PSUW, PEW, PSUW + ENCPW, PEW + PSUW) the relative dominance of taxonomic groups varied for both prokaryotes (Fig. 3.2S) and eukaryotes (Fig. 3.3S). Within prokaryotes, Alphaproteobacteria, Gammaproteobacteria, and Bacteroidia were the most taxonomically rich groups and had many species in each affinity group. For eukaryotes, Syndiniales and Dinophyceae were the most taxonomically rich and occurred across all affinity groups. Within eukaryotic communities there was a large number of ASVs that fell outside of the 30 most taxonomically rich groups (labeled as Other Eukaryotes, Fig. 3.3S). In general, the majority of ASVs per taxonomic group had no water mass affinities, however there were exceptions. Within prokaryotes, the majority of Lentisphaeria and Nitrospinia ASVs had an affinity for the PEW (Fig. 3.2Sc). Within eukaryotes, the majority of Telonemia ASVs had an affinity for the PSUW (Fig. 3.3Sc). For many groups, where the majority of ASVs have no water mass affinity, we still

find preferences for particular water masses for those ASVs with affinities. For instance, 16S ASVs within Nitrososphaeria, a class of ammonia oxidizing archaea, and 18Sv9 ASVs within the cercozoan groups Filosa-Thecofilosea both have strong affinities for the PEW even though the majority of ASVs within each group have no water mass affinity. Group wide preference could be an indication of environmental specialization occurring at a relatively high taxonomic level, whereas those groups with relatively evenly distributed occurrence across all affinity types may indicate that this level of taxonomic aggregation does not align with niche specialization.

Finally, we quantified the degree of overlap between our NCOG data and global survey data (Tara Oceans and Tara Polar). On average, 71.2% of prokaryotic ASVs and 37.3% of NCOG eukaryotic ASVs were not found in any Tara samples. Overlap between NCOG and Tara surveys was highest in the Pacific and Atlantic basins. Intermediate levels of overlap occurred between NCOG and the Indian Ocean, Red Sea, and Mediterranean Sea. Lowest levels of overlap occurred between NCOG and the Southern Ocean, Arctic Ocean, and Tara Polar North Atlantic samples (Fig. 3.6a-b). We also explored the rate of overlap between ASVs in our occurrence categories (Endemic, Semi-Dispersed, Cosmopolitan, and Ubiquitous). In general, we found that the degree of overlap between NCOG and Tara Oceans regions aligned strongly with the total diversity found in each region of Tara's globally sampling with higher regional diversity (# of ASVs) leading to a higher degree of overlap. Ubiquitous NCOG ASVs were seen in nearly every single region for both 16S and 18Sv9. Cosmopolitan NCOG ASVs had the most striking relationship between regional diversity and overlap—low-diversity regions contained few of the cosmopolitan NCOG ASVs while high-diversity regions contained up to 39.0% and 74.7% of the NCOG cosmopolitans for prokaryotes and eukaryotes, respectively (Fig. 3.6c-d).

Since Tara samples were collected from primarily open-ocean stations, we were also interested in whether ASVs with an affinity for more oligotrophic conditions would overlap more strongly with global Tara data. To explore this question, we binned ASVs with particular affinities into three groups: oligotrophic (affinities for ENPCW, PSUW, and ENPCW + PSUW), meso/eutrophic (PEW), and no water mass affinities. In low-diversity regions we observed little to no difference between these groups. However, in high-diversity regions we observed a higher rate of overlap between oligotrophic-associated ASVs in NCOG and those ASVs found from Tara when compared to both meso/eutrophic group and those ASVs with no water mass affinities (Fig. 3.6e-f). In general, ASVs with no water mass affinities had the least overlap with regional Tara datasets, a surprising finding given their numerical (richness) dominance in NCOG.

Overlap between NCOG and Tara samples varied amongst taxonomic groups and from region to region, as certain prokaryotic ASVs (Fig. 3.4S) and eukaryotic ASVs (Fig. 3.5S) found in NCOG were more or less likely to be found globally. Within prokaryotes, Thermoplasmata ASVs were overrepresented in the overlap of NCOG and Tara for all regions but the Southern Ocean. Within the Southern Ocean, the limited overlap was driven almost completely by the three most diverse prokaryotic groups (Alphaproteobacteria, Gammaproteobacteria, and Bacteroidia, Fig. 3.4Sb). Within eukaryotes overrepresented taxonomic groups were more variable from region to region than prokaryotes. Eukaryotic groups that were overrepresented in multiple regional overlaps included: Choanoflagellatea, Prymnesiophyceae, and Filosa-Imbricatea. The overlap between eukaryotic NCOG ASVs and eukaryotic Southern Ocean ASVs included far more unique taxonomic groups compared to prokaryotic groups, even though the overlap between NCOG and the Southern Ocean was small in both cases (Fig. 3.5Sb).

## 3.3 Discussion

Overall, we found that the majority of ASVs identified within the Southern California Current (SCC) region were rare in both abundance and occurrence (Fig. 3.2). Roughly half of the prokaryotic and eukaryotic diversity was represented by singletons ASVs that only occurred in one sample. Within the non-singleton ASVs, the majority had no habitat specificity (Fig. 3.4). These species were relatively rare (occurrence and abundance, Fig. 3.4g-h) and on average had less overlap with global samples than those ASVs with habitat preferences (Fig. 3.6c-f). These results align with previous findings which have highlighted the overabundance of rare species within the marine microbiome (Bachy & Worden 2014; Logares et al. 2014; Ser-Giacomi et al. 2018). A possible explanation for the overabundance of rare taxa in our observations and their overabundance generally within the marine microbiome could be the dynamic mixing of habitats within the ocean. In a study by Martín et al. 2020 (Martín et al. 2020), researchers suggested that the chaotic mixing of water masses, characterized by well-mixed areas interspersed with steep gradients, effectively isolates species from larger patches of suitable habitat, reducing the likelihood of highly abundant species, while also reducing competition across the entire range of a given habitat—preserving populations of ecological equivalent or even maladapted species within a particular habitat. That these rare ASVs had relatively low overlap across Tara regions may indicate that at a global scale, the relative effects of selection and drift outweigh the regional effects of chaotic mixing.

Cosmopolitan species were the second-most numerically rich non-singleton category representing 29.0% and 31.7% of the non-singleton prokaryotic and eukaryotic ASVs respectively. The majority of cosmopolitans had an affinity to one water mass and tended to be more abundant when compared to endemics or other rare taxa (Fig. 3.4). Large abundances combined with habitat preferences may indicate that mass effects (the diffusion of populations

from areas of high to low density) may lead to their cosmopolitan distributions at the regional scale (Leibold et al. 2004). Cosmopolitans also showed the strongest positive relationship between regional diversity and the percent overlap between Tara and NCOG (Fig. 3.6c-d). Regions of low total diversity within Tara correspond to either dispersal bottlenecks (Mediterranean and Red Sea) or strong environmental gradients (Arctic and Southern Ocean) and as such, may present strong barriers to the immigration of NCOG cosmopolitans. Combined, these results indicate that for abundant marine microbes, selection and dispersal may be the strongest factors for determining both local and global biogeographies.

Endemic species were the third-most common non-singleton category and were evenly split between ASVs with no significant habitat affinity, and those with a significant habitat affinity. In total, 10.8% of non-singleton prokaryotes and 8% of non-singleton eukaryotes were significantly overabundant and endemic to a single habitat. Compared to terrestrial microbiomes these rates are relatively low, 40% of per sample richness and 80% of regional richness was endemic to a given location (Talbot et al. 2014). Like other rare taxa, marine microbial endemics may be able to survive local extinction via the chaotic mixing of water masses at the submesoscale level (Martín et al. 2020).

Finally, only five taxa across both prokaryotic and eukaryotic ASVs occurred in every sample within NCOG. Of these five, only one eukaryotic taxon showed no affinity towards a particular habitat or habitats. Ubiquitous ASVs had a nearly perfect overlap across all Tara global regions. Yet, some ubiquitous ASVs were missing from the Tara regions with the lowest prokaryotic and eukaryotic richness respectively (Fig. 3.6c-d). Like cosmopolitans, the biogeographies of these ubiquitous taxa appear to be largely structured by their ecological

preferences, however, unlike cosmopolitans it appears that even the steepest dispersal barriers or selective gradients do not hinder the global occurrence of these taxa.

If selection were the most important factor for structuring marine microbial communities, we would expect increased dominance of ASVs with habitat affinities in their respective habitats. Within prokaryotic assemblages, this appeared to be the case. ASVs that had an affinity with a particular habitat had higher relative abundances within samples in that habitat. In contrast with prokaryotes, a relatively large proportion of eukaryotic communities were comprised of rare taxa with no water mass affinities. This difference between prokaryotic and eukaryotic communities aligns with results from a study by Logares et al. 2020 (Logares et al. 2020), which suggested that within the marine microbiome, prokaryotic communities may be more structured by selection whereas eukaryotic communities may be more structured by dispersal. While selection still can influence eukaryotic community composition, non-specific eukaryotic taxa may be more abundant, particularly in a well-mixed region like the SCC.

Through the combined sampling across both NCOG and Tara we have explored rates of marine microbial endemism, cosmopolitanism, ubiquity, and habitat specificity across regional to global scales. We find continued evidence that the marine microbiome is dominated by numerically rare taxa which may persist locally due to the dynamic mixing of ocean habitats. Abundant taxa tend to have increased habitat specificity but higher dispersal potential leading to regional cosmopolitanism. The distributions of more abundant taxa appears to be driven by both selection and dispersal at both local and global scales, barring a few exceptionally common taxa. Finally, within prokaryotes, habitat-specific community structure is largely dominated by taxa specific to that habitat. Eukaryotic communities are also dominated by taxa specific to their respective habitats, however, non-specific taxa are far more common in these communities

relative to prokaryotic communities. Combined, these results suggest that within the SCC, selection may be a stronger community-structuring force in prokaryotic communities relative to eukaryotic communities. Overall this study confirms recent suggestions that the majority of marine microbes are unlikely to be globally ubiquitous (van der Gast 2015; Ward et al. 2021), but rather show distinct biogeographies that, while perhaps at different scales relative to macroorganisms and terrestrial systems, are driven by the same core ecological processes of dispersal, selection, and drift (Vellend 2010).

## 3.4 Methods

### 3.4.1 Study location and sample collection

NOAA CalCOFI Genomics Project (NCOG) data were collected within the Southern California Current (SCC) region, a productive eastern boundary current ecosystem. The data analyzed here consist of 445 surface (nominally at 10m depth) samples collected quarterly from 2014-2020. Cardinal stations on CalCOFI lines 80 (stations 55.0, 70.0, 80.0, 100.0), 81.8 (station 46.9), and 90 (stations 37.0, 53.0, 70.0, 90.0, 120.0) were sampled every cruise. Primary productivity stations, which measure $^{14}C$ primary production at approximately noon were also sampled. Productivity stations vary from cruise to cruise depending on where the ship is located each day at approximately midday.

Both molecular (described in more detail below) and environmental data (temperature, salinity, nutrients) were collected via a CTD rosette. Temperature and salinity were measured with the Seabird 911 CTD. Salinity measurements were compared to bottle samples that were measured with a Guildline Portasal Salinometer model 8410A. Nutrients ($NO_3$ and $NH_3$) were measured with a QuAAtro continuous flow autoanalyzer (SEAL analytical). For a

comprehensive description of collection and processing methods related to the NCOG database, see James et al. (2022).

### 3.4.2 DNA collection and extraction

Approximately 0.5 – 2 L of seawater was filtered through a 0.22 µm Sterivex-GP filter unit (MilliporeSigma, Burlignton, MA, USA) for all DNA samples. Samples were immediately sealed with a sterile luer-lock plug and hematocrit sealant, wrapped in aluminum foil, and flash frozen in liquid nitrogen. DNA was extracted with the NucleoMag Plant Kit for DNA purification (Macherey-Nagel, Düren, Germany) on an epMotion 5057TMX (Eppendorf, Hamburg, Germany) as described here: https://dx.doi.org/10.17504/protocols.io.bc2hiyb6. DNA was assessed on a 1.8% agarose gel after extraction.

### 3.4.3 Amplicon sequencing and analysis

Amplicon sequence variant (ASV) libraries used in this analysis targeted the V4-V5 region of the 16S rRNA gene for prokaryotes and V9 region of the 18S rRNA gene for eukaryotes. See James et al. 2022 for a more detailed description of both the primer sets and methodology. For full protocols visit: https://www.protocols.io/view/amplicon-library-preparation-bmuck6sw.

### 3.4.4 Rarefaction of amplicon data (1,000-member library ensemble)

NCOG and Tara samples had a wide variety of library sizes. We therefore rarefied our data to a consistent level of 20,024 reads for prokaryotes and 30,347 reads for eukaryotes, representing the $99^{th}$ percentile of library sizes across all samples for prokaryotes and eukaryotes respectively, to remove the effect of sequencing noise (library size) on our results. Repeated rarefaction is preferred to a singular rarefaction step as rarefaction can lead to the loss of novel but rare taxa, which is particularly important for this study as we aimed to identify the rates of

endemism, cosmopolitanism, and habitat specificity within the region (Cameron *et al.* 2021). Thus, we repeated the rarefaction step 1,000 times and then ran our analyses on each of these rarefied microbiome tables. Results for all figures represented the mean values across the entire 1,000-member ensemble, with standard deviations highlighted where appropriate. In doing so, we were able to statistically assess the occurrence and habitat specificity of rare taxa that might not be observed through a single rarefaction step.

### 3.4.5 Self-organizing maps (SOMs)

We used SOMs to categorize and differentiate marine habitats based upon environmental conditions. SOMs are a machine learning approach capable of reducing highly dimensional data into a two-dimensional map (Kohonen 1997). SOMs have previously been used to identify 'seascapes' along the Western Antarctic Peninsula (Bowman *et al.* 2018). We followed a similar procedure using three physical parameters: temperature, salinity, and $NO_3$ + $NH_3$ to construct our SOM. Once the SOM was generated, we used hierarchical clustering to cluster the SOM into three seascapes. These seascapes, called the Eastern North Pacific Central Water (ENPCW), Pacific Subarctic Upper Water (PSUW), and Pacific Equatorial Water (PEW), align with the temperature, salinity, and nutrient concentrations of the three major water masses that comprise surface waters in the SCC (Bograd *et al.* 2019). The spatial distribution of three core ocean habitats varied through time, reflecting the dynamic nature of the marine environment.

### 3.4.6 Null model for water mass affinity

We used a null model test to assess whether ASVs had a significant affinity for any given seascape. Null models were generated for each ASV and each rarified library. Below we have outlined the step-by-step processes for identifying the water mass affinity for an individual taxon

within each water mass for one ensemble member. This process was then applied across all taxa and across all 1,000 ensemble members.

First, we calculated the mean relative abundance of a taxon in each water mass ($\bar{\mu}_{ENPCW}$, $\bar{\mu}_{PSUW}$, $\bar{\mu}_{PEW}$). Then, the relative abundances of a taxon were reshuffled across all samples (with replacement) 1,000 times. From these 1,000 surrogate abundance distributions we calculated 1,000 null mean relative abundances per water mass. P-values were calculated for actual mean relative abundances compared to their respective null distributions with the following equation:

$$p = \frac{number\ of\ surrogate\ means > actual\ mean}{total\ number\ of\ surrogate\ means} \quad (1)$$

To assess the overall significance of water mass affinity within a taxon across all ensemble members we calculate the mean p-value per water mass per taxa across all 1,000 libraries.

### 3.4.7 Tara Oceans and Tara Polar samples

In this study, we explored the overlap between NCOG ASVs and Tara Oceans and Tara Polar ASVs. For both 16S and 18Sv9, we only used samples that were collected by filtering seawater via a peristaltic pump (excluding net samples for 18Sv9). For 16S, the size fractioned filter was 0.22 µm to 1.6 µm. For 18Sv9, this included multiple size fractions ranging from 0.22 µm to 200 µm.

### 3.4.8 Data availability

The 16S rDNA raw reads are available for the 2014-2016, 2017-2019, and 2020 periods at NCBI under Bioproject IDs PRJNA555783, PRJNA665326 and PRJNA804265 and Biosample accessions SAMN25705811-SAMN25706151, SAMN16250568-SAMN16251083, and SAMN25756929-SAMN25757078. The 18S rDNA raw reads for the 2014-2016, 2017-2019, and 2020 periods have been deposited at NCBI under Bioproject IDs PRJNA555783,

PRJNA665326, and PRJNA804265 and Biosample accessions SAMN25710021-SAMN25710361, SAMN16251281-SAMN16251796, and SAMN25757352-SAMN25757501.

Tara Oceans and Tara Polar 18Sv9 sequences can be found at the European Nucleotide Archive under the project accession IDs PRJEB6610 and PRJEB9737. Tara Oceans and Tara Polar 16S and 18Sv9 sequences can be found at the European Nucleotide Archive under the project ID: PRJEB402.

## 3.5 Figures

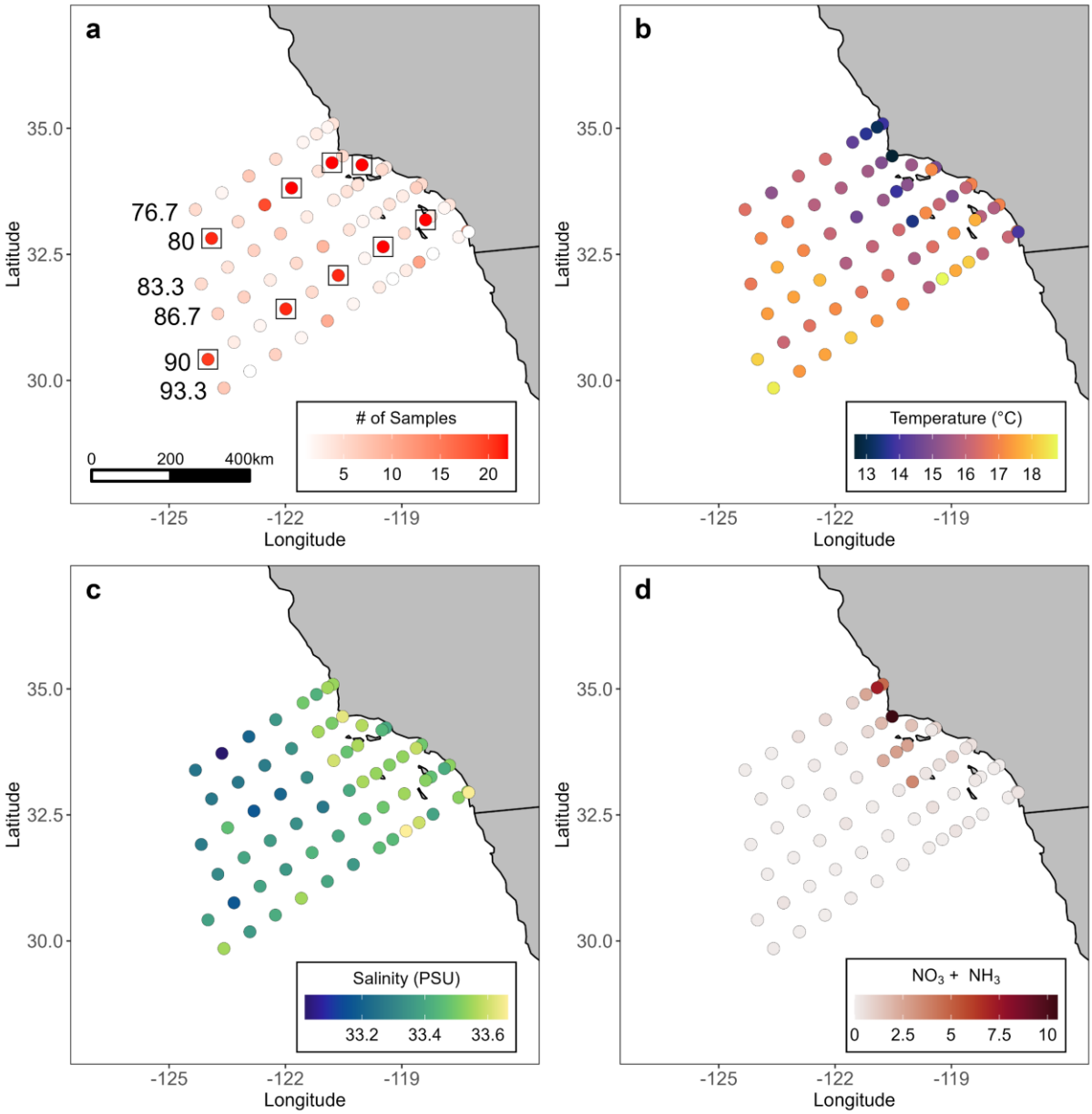


**Figure 3.1**: Description of the sampling regime and physical environment over 2014-2020. a, Number of samples per station. Squares highlight Cardinal stations which are sampled every cruise. b, Mean temperature (°C) per station. c, mean salinity (PSU) per station. d, mean NO3 + NH3, µM per station.

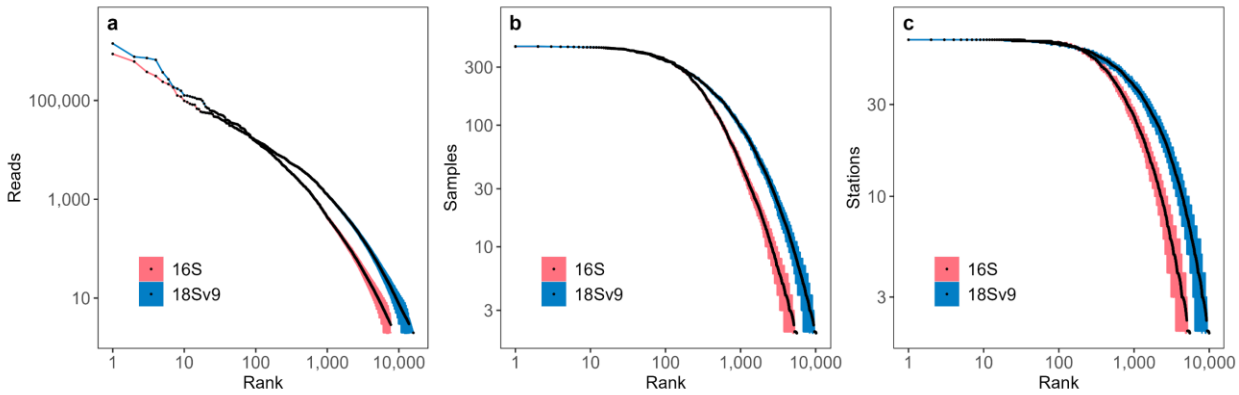


**Figure 3.2**: Rank curves for 16S and 18Sv9. a, log10-log10 relationship between mean abundance (reads) and abundance rank. b, log10-log10 relationship between mean occurrence (samples) and occurrence rank. c, log10-log10 relationship between mean occurrence (stations) and occurrence rank. Color indicates either 16S (pink) or 18Sv9 (blue) ASVs. Shading around the means (points) show the upper (95%) and lower (5%) percentiles of either abundance or occurrence, calculated from the 1,000-member library ensemble.

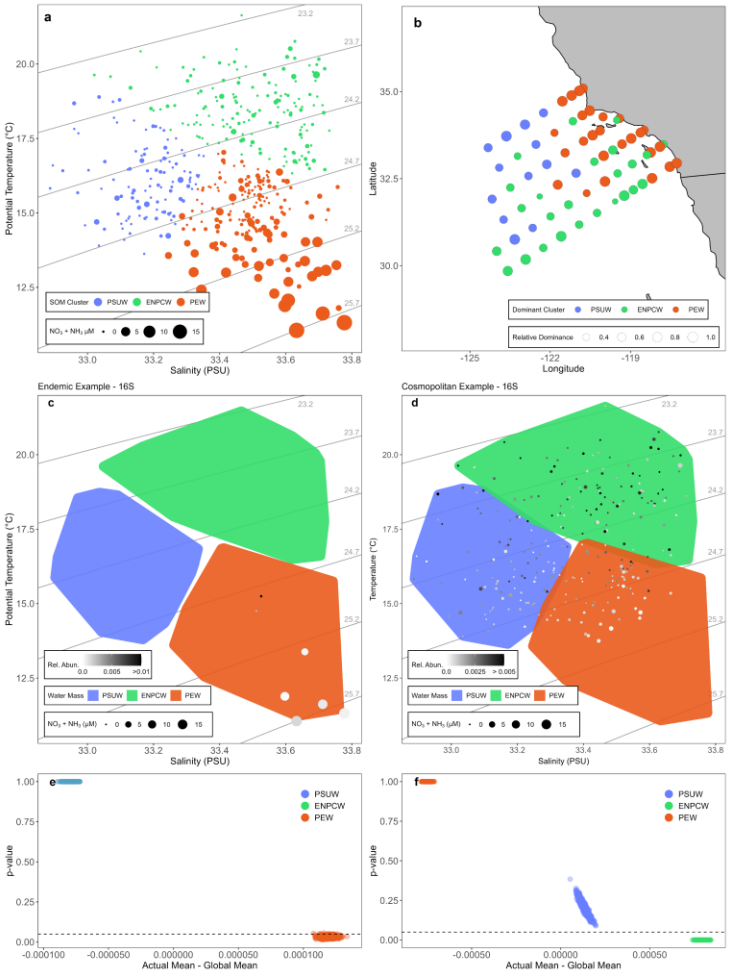


**Figure 3.3**: a, Three-dimensional environmental space for the 445 samples with temperature, salinity, and NO3 + NH3 data. Axes show potential temperature (°C) and salinity (PSU). The size of the points represents NO3 + NH3 (µM) of each sample. Color of the points represents the identified SOM clusters which align with known water masses: Pacific Subarctic Upper Water (PSUW, blue), East North Pacific Central Water (ENPCW, green), and Pacific Equatorial Water (PEW, orange). Solid black lines indicate isopycnals of constant seawater density (also in c and d). b, Map showing the most dominant water mass per station, where the size of the circles represents the frequency with which that water mass is observed at a given station. c-d, Example temperature and salinity diagram showing the occurrence and relative abundance of an endemic and cosmopolitan ASV, respectively, across all 445 samples. The color of the points represents the relative abundance of the ASV per sample. Blue, green, and orange shaded regions show the boundary of each water mass. The size of the points represents NO3 + NH3 (µM) of each sample. e-f, Example significance vs abundance diagrams for the endemic and cosmopolitan ASVs in c-d, highlighting which water mass(es) the ASVs had a significant affinity for (p-value < 0.05, dashed line). The x-axis shows the mean relative abundance within a water mass – the overall (across all samples) mean relative abundance for that ASV. The y-axis shows the p-value associated with each mean relative abundance per water mass (see methods for p-value calculation). A high value along the x-axis means that the abundance within a water mass is higher than the mean overall abundance for that taxon across all samples. Values below the dashed line on the y-axis represent ensemble members where the abundance was significantly greater in a water mass then the null (p-value < 0.05). Thus, in this example, the endemic ASV is significantly overabundant in the water mass it is observed in, while the cosmopolitan species, while found everywhere, is only significantly overabundant in the ENPCW.

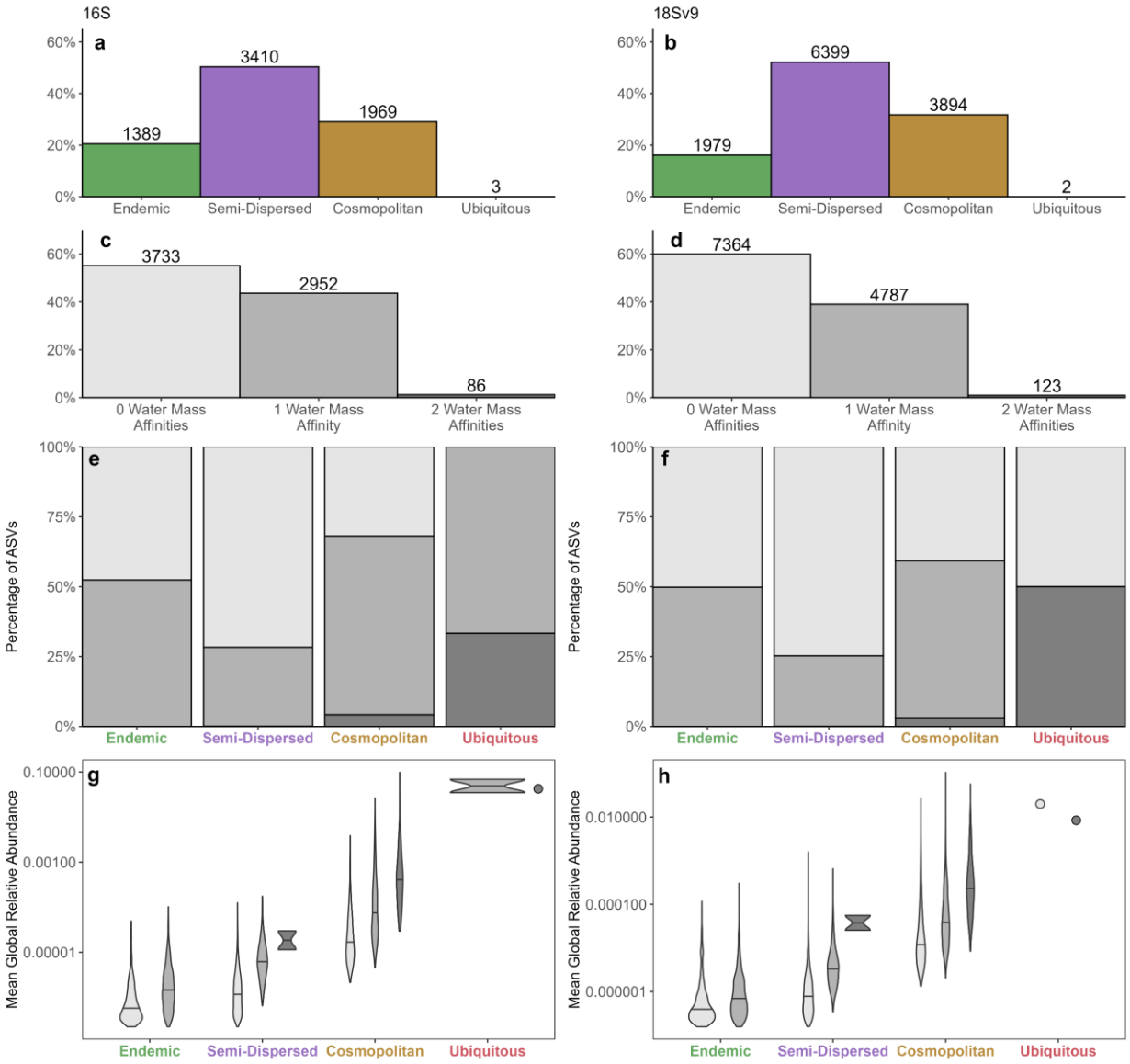


**Figure 3.4**: a-b, Percentage of 16S and 18Sv9 ASVs respectively, in each descriptive category (Endemic, Semi-Dispersed, Cosmopolitan, and Ubiquitous). Values above each bar show the number of ASVs in each category. c-d, Percentage of 16S and 18Sv9 ASVs respectively, in each affinity (0 water mass affinities, 1 water mass affinity, or 2 water mass affinities). Values above each bar show the number of ASVs in each category. e-f, Percentage of 16S and 18Sv9 ASVs respectively, in each affinity level per descriptive category. g-h, Distributions of mean overall (across all samples) relative abundance across all 16S and 18Sv9 ASVs respectively. Distributions are separated per descriptive category and affinity level.

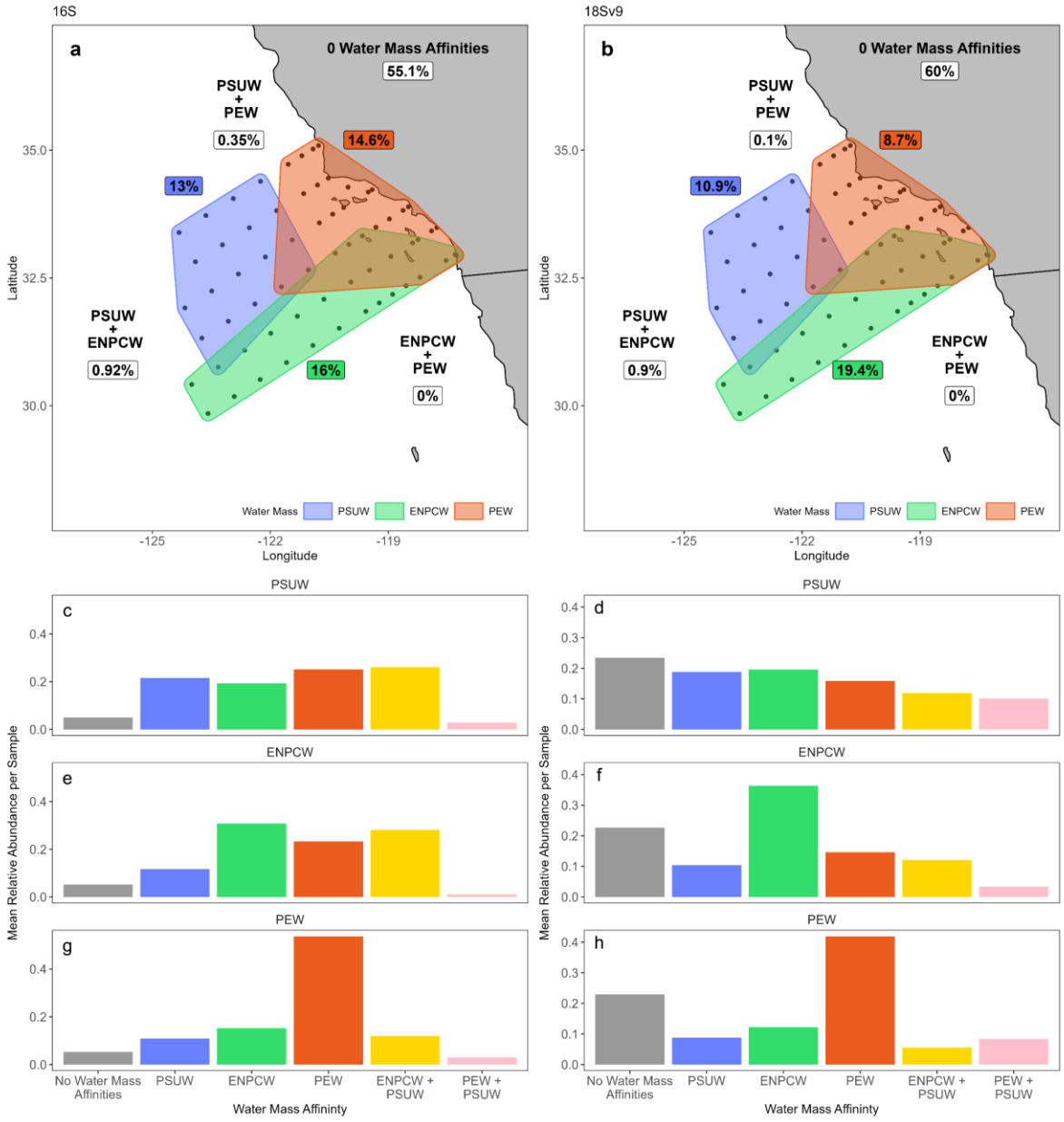


**Figure 3.5**: a-b, Map indicating the percentage (numbers in boxes) of 16S and 18Sv9 ASVs respectively that have an affinity for: no water masses, one water mass (PSUW, ENPCW, or PEW), or two water masses (PSUW + ENPCW, PSUW + PEW, or ENPCW + PEW). c-d, Mean per sample relative abundance of 16S and 18Sv9 ASVs respectively within PSUW samples for each affinity group. e-f, Mean per sample relative abundance of 16S and 18Sv9 ASVs respectively within ENPCW samples for each affinity group. g-h, Mean per sample relative abundance of 16S and 18Sv9 ASVs respectively within PEW samples for each affinity group. Mean per sample relative abundances are calculated by summing the relative abundances of ASVs in each affinity group per sample, then an average is calculated per affinity group based on samples within each water mass.

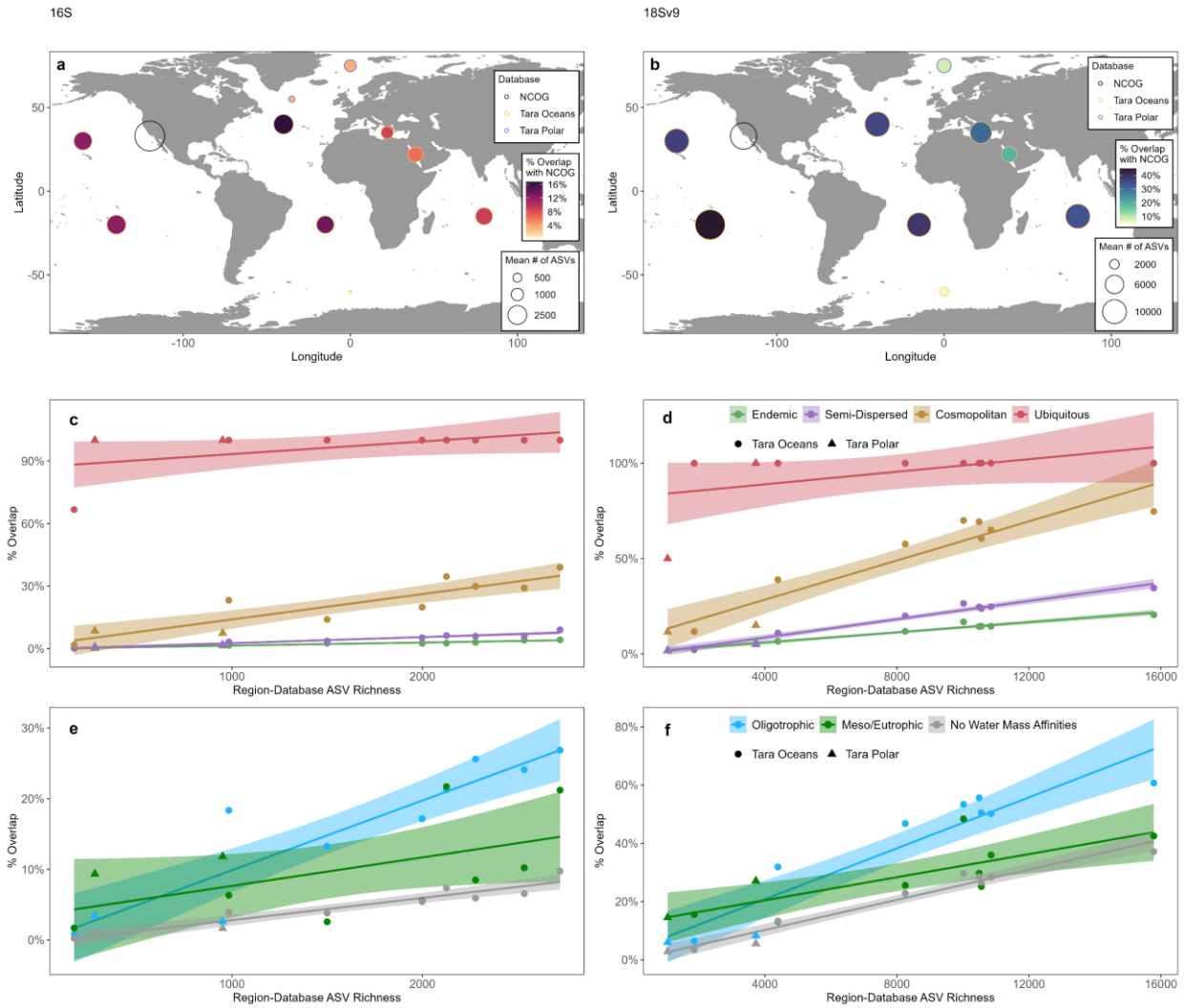


**Figure 3.6**: a-b, Maps highlighting the overlap between NCOG and Tara Oceans and Tara Polar. Size of circles indicates the mean number of ASVs identified per region per database. Stroke color represents the database for the respective data (NCOG, Tara Ocean, or Tara Polar). a, fill color represents the percentage of NCOG ASVs found in each respective region/database for 16S. b, fill color represents the percentage of endemic NCOG ASVs (observed at only one station in NCOG) found in each respective region/database for 16S. c, Relationship between regional richness (per database) and the % overlap between NCOG 16S ASVs and regional 16S ASVs. Colors represent the four occurrence categories (Endemic, Other, Cosmopolitan, and Ubiquitous). d, Relationship between regional richness (per database) and the % overlap between NCOG 18Sv9 ASVs and regional 18Sv9 ASVs. Colors represent the four occurrence categories (Endemic, Other, Cosmopolitan, and Ubiquitous). e, Relationship between regional richness (per database) and the % overlap between NCOG 16S ASVs and regional 16S ASVs. Colors represent combined affinity categories (Oligotrophic, Eutrophic, and No Water Mass Affinities). f, Relationship between regional richness (per database) and the % overlap between NCOG 18Sv9 ASVs and regional 18Sv9 ASVs. Colors represent combined affinity categories (Oligotrophic, Eutrophic, and No Water Mass Affinities).

## 3.6 Supplementary Information

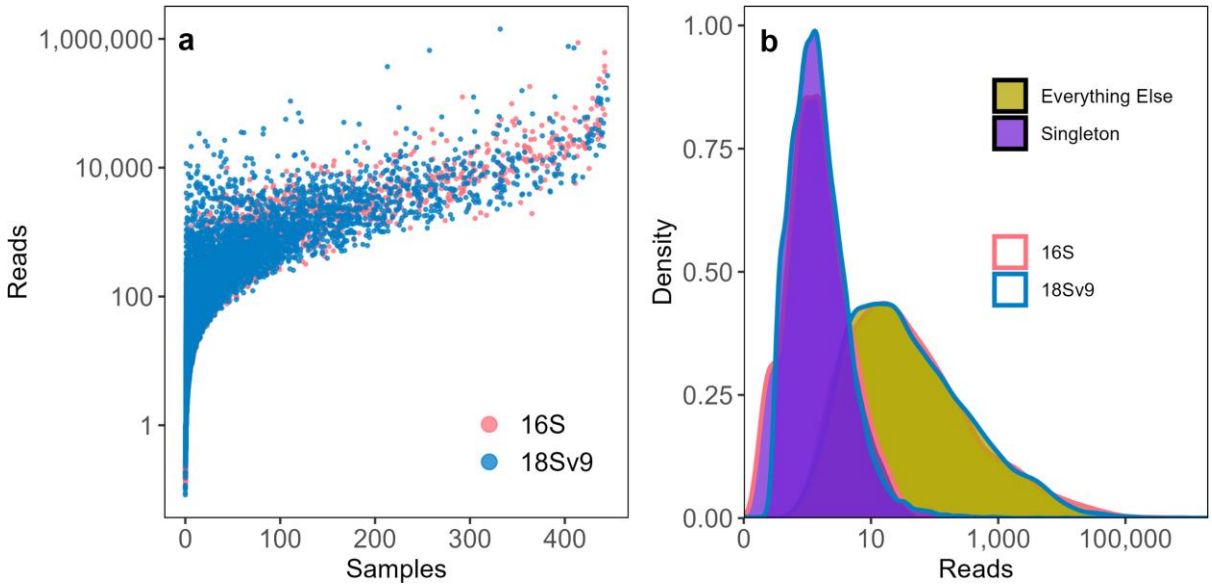


**Figure 3.1S**: a, Occurrence versus log10 number of reads, or abundance, for 16S and 18Sv9. Color indicates either 16S (pink) or 18Sv9 (blue) ASVs. b, Histograms of mean total reads (log10) per ASV split between Singletons (ASVs only seen in one sample, purple) and everything else (yellow) for 16S and 18Sv9. Outline color indicates either 16S (pink) or 18Sv9 (blue) ASVs.

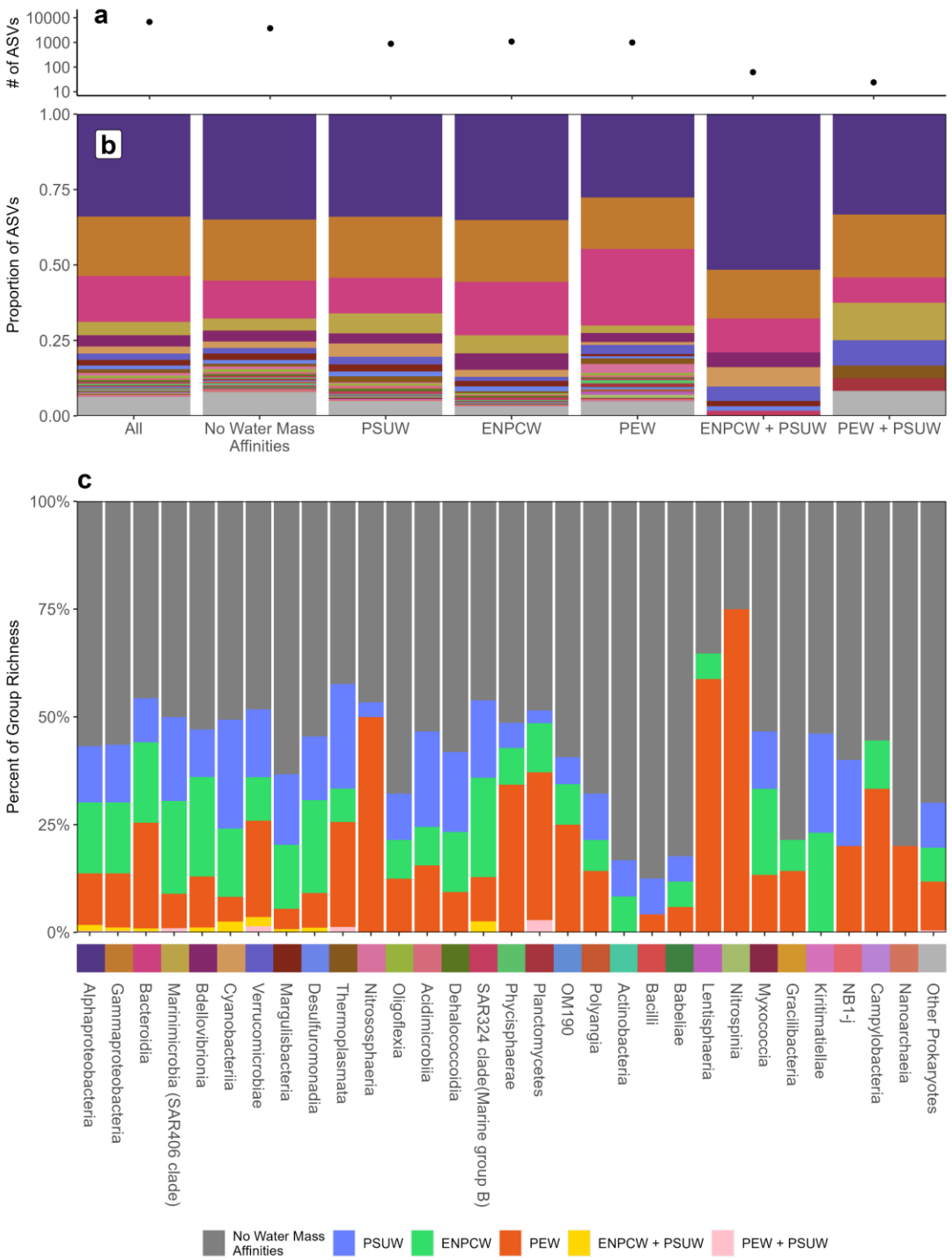


**Figure 3.2S**: Taxonomic composition of 16S ASVs with significant affinities per water mass or water mass combination. a, Total number of ASVs per affinity type (All ASVs, No Water Mass Affinities, PSUW, ENPCW, PEW, ENPCW + PSUW, and PEW + PSUW). b, Relative proportion of ASVs in broad taxonomic groups per affinity type (All, No Water Mass Affinities, PSUW, ENPCW, PEW, ENPCW + PSUW, and PEW + PSUW). c, Percentage of ASVs per taxonomic group in each affinity group (bar colors). Colors next to taxonomic groups indicate the color for that taxonomic group in subpanel b.

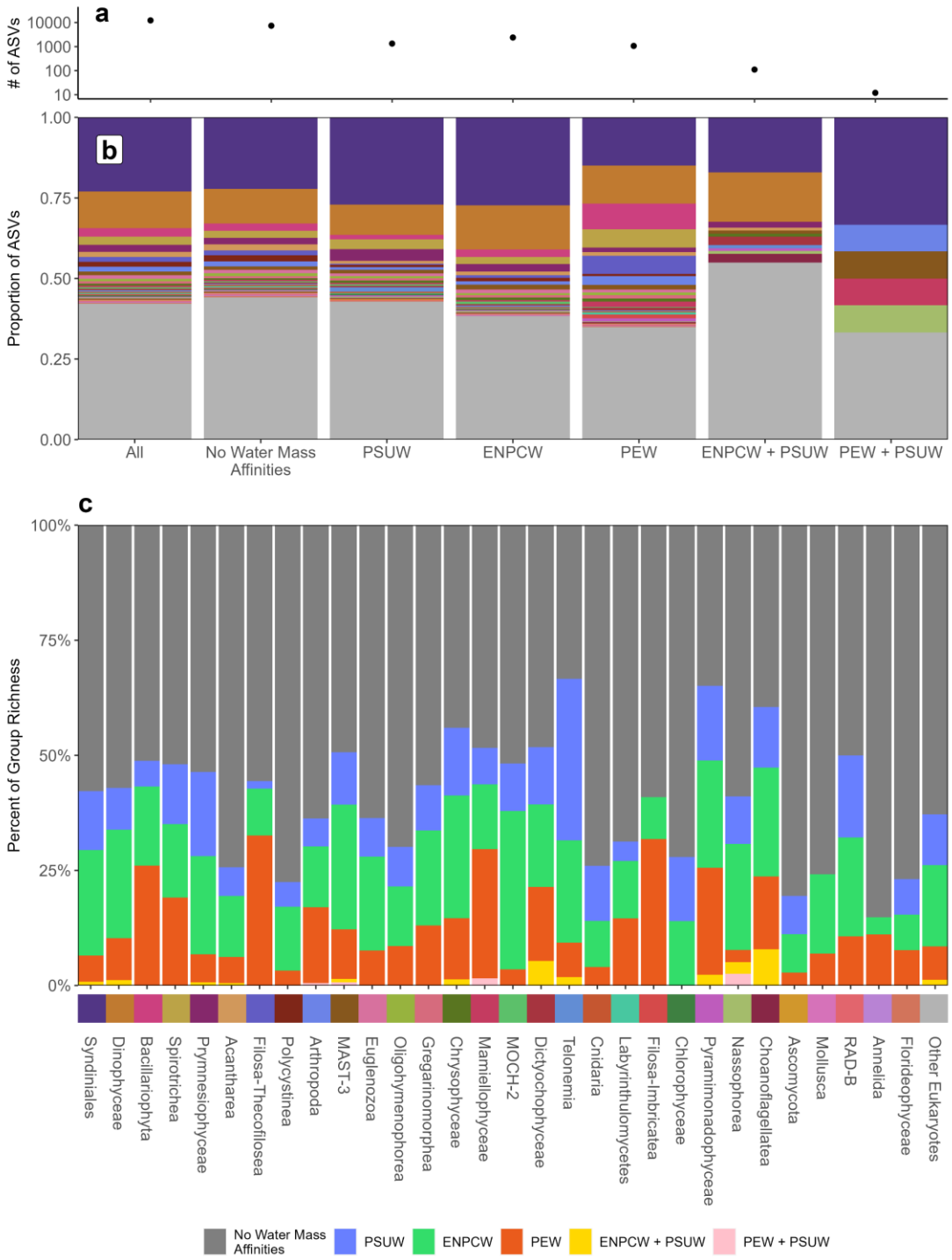


**Figure 3.3S**: Taxonomic composition of 18Sv9 ASVs with significant affinities per water mass or water mass combination. a, Total number of ASVs per affinity type (All ASVs, No Water Mass Affinities, PSUW, ENPCW, PEW, ENPCW + PSUW, and PEW + PSUW). b, Relative proportion of ASVs in broad taxonomic groups per affinity type (All, No Water Mass Affinities, PSUW, ENPCW, PEW, ENPCW + PSUW, and PEW + PSUW). c, Percentage of ASVs per taxonomic group in each affinity group (bar colors). Colors next to taxonomic groups indicate the color for that taxonomic group in subpanel b.

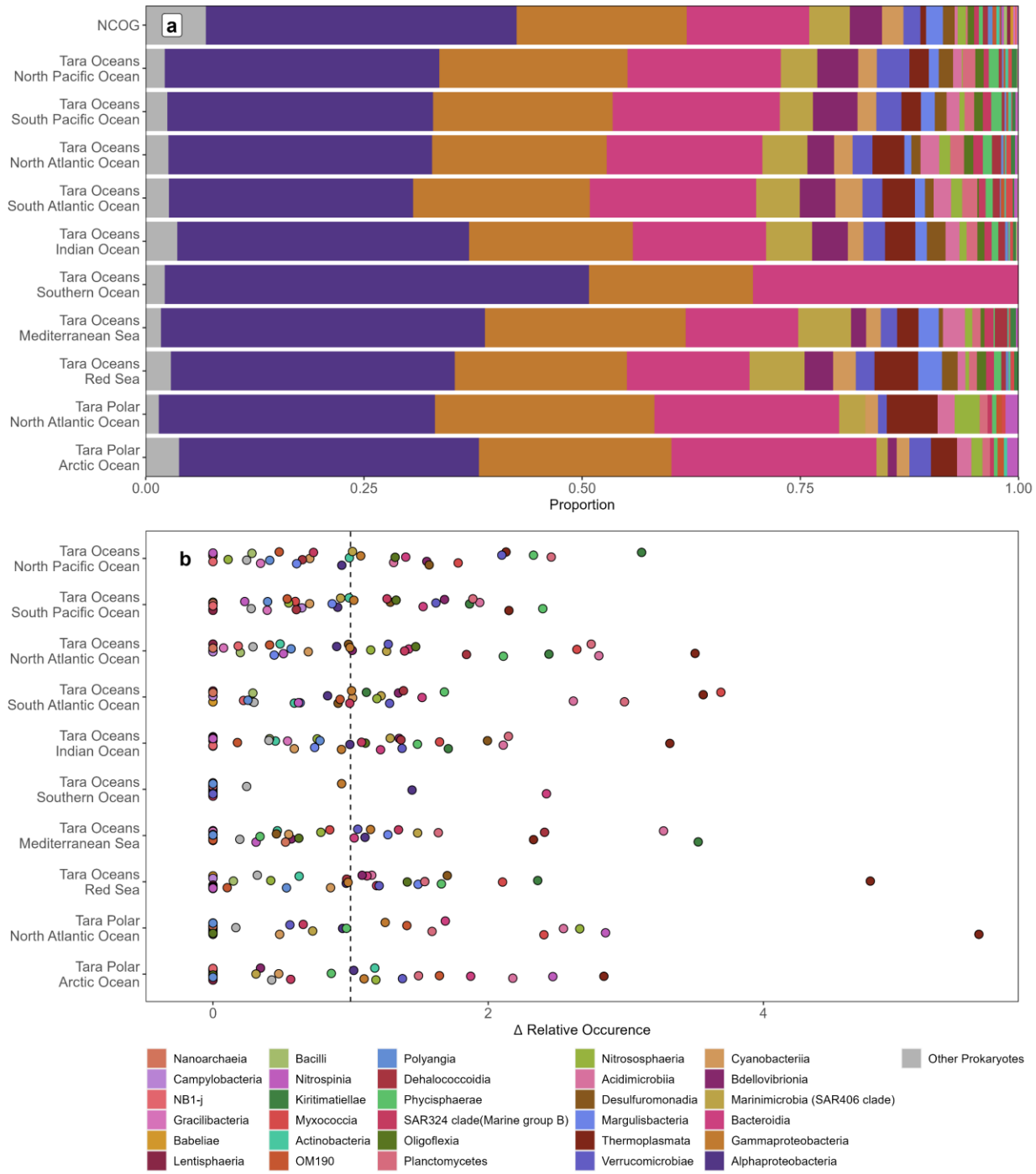


**Figure 3.4S**: Taxonomic composition of 16S ASVs per region. a, Relative proportion of ASVs in broad taxonomic groups per regional dataset. b, Representation of broad taxonomic groups (# ASVs) in each regional dataset relative to their representation across all NCOG 16S ASVs. Δ Relative Occurrence = Proportional Richness per region / Proportional Richness within NCOG per taxonomic group. Larger numbers indicate that a given taxonomic group represents a larger proportion of a particular regional richness compared to its proportional richness in NCOG. Zero values indicate that that taxonomic group is not found in a particular region.

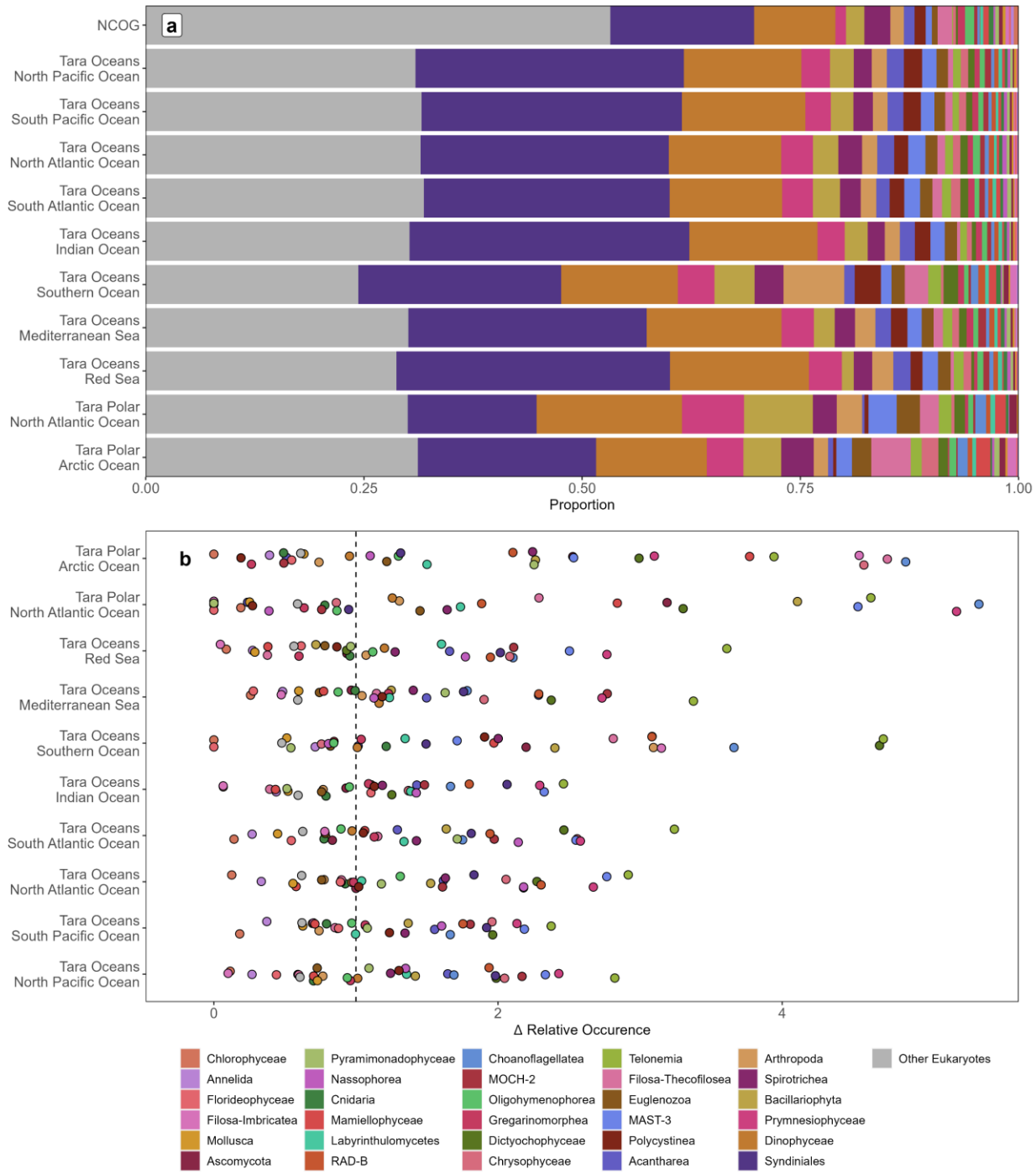


**Figure 3.5S**: Taxonomic composition of 18Sv9 ASVs per region. a, Relative proportion of ASVs in broad taxonomic groups per regional dataset. b, Representation of broad taxonomic groups (# ASVs) in each regional dataset relative to their representation across all NCOG 18Sv9 ASVs. Δ Relative Occurrence = Proportional Richness per region / Proportional Richness within NCOG per taxonomic group. Larger numbers indicate that a given taxonomic group represents a larger proportion of a particular regional richness compared to its proportional richness in NCOG. Zero values indicate that that taxonomic group is not found in a particular region.

## 3.7 Acknowledgements

CCJ acknowledges graduate student support by Scripps Institution of Oceanography. This study was supported by National Science Foundation, California Current Ecosystem Long Term Ecological Research Grants, CCE-LTER Phase II and III (NSF-OCE-1026607 and NSF-OCE-1637632), and NSF-OCE-1756884, NOAA (NOAA OAR Omics, CIMEC NA15OAR4320071, and ECOHAB NA19NOS4780181), and Gordon and Betty Moore Foundation grants GBMF3828 to AEA.

We would like to acknowledge former CalCOFI director David M. Checkley and Margot Bohan from the NOAA Office of Ocean Exploration and Research (OER) for their vision and guidance during the initial phase of the NCOG program and current CalCOFI director Brice X. Semmens for his continued support. We are also especially grateful to California Current Ecosystem, Long Term Ecological Research (CCE-LTER) and CalCOFI project and team members and crew who have assisted with the NCOG program 2014-present.

Chapter 3, in part is currently being prepared for submission for publication of the material. James, C.C., Allen, A. E., Lampe, R. H., Barton, A. D. The dissertation author was the primary investigator and author of this paper.

# Chapter 4 - Metatranscriptomics reveal marine microbial realized niche and ecological function


## Abstract

The taxonomic and functional composition of the marine microbiome dictates the magnitude of many key ecosystem services within the pelagic system including the relative rates of primary productivity, nutrient recycling, and carbon sequestration. Using metatranscriptomic data from a coastal upwelling region sampled over seven years (2014-2020), we explore the relationships between environmental gradients and their effect on taxonomic niche partitioning and functional biogeography. Across all prokaryotes and eukaryotes, we find that niche optimums within temperature and nitracline depth are variable, niche optimums vary among taxa, and are selective, populations decrease quickly away from optimums, indicative of strong niche partitioning among taxa across these environmental gradients. Across all taxa, functional gene composition varies most strongly across gradients in salinity, however, temperature and nitracline depth tend to be more important in shaping the relative abundance of different functional gene classes within broad taxonomic groups such as eukaryotic phytoplankton. Finally, we find that most groups show community-wide shifts in functional responses to environmental conditions, except eukaryotic phytoplankton, in which functional shifts occur with large shifts in community dominance, highlighting the importance of ecologically uneven but high-productivity eukaryotic phytoplankton communities in shaping the structure and function of the coastal oceanic region.


## 4.1 Introduction

Spatial and temporal differences in community structure and functional composition within the marine microbiome dictate the local magnitude of the ecosystem functions at the base

of the pelagic food web (Falkowski, Fenchel, and Delong 2008; Fuhrman 2009; Fuhrman, Cram, and Needham 2015). Some ecosystem functions, such as primary productivity are important for regional economic success of coastal fisheries (Stock et al. 2017), while others like carbon sequestration, a natural process for shunting atmospheric carbon to the deep ocean, have global climate implications and can vary largely as a result of changes to the active microbial community (Abrantes et al. 2016).

Within coastal upwelling regions like the Southern California Current (SCC), strong cross shore gradients, shaped by a nutrient rich nearshore and an oligotrophic offshore, lead to drastically different environmental, and subsequent ecological communities across the region (Checkley and Barth 2009; Venrick 2009; Taylor and Landry 2018). Across the region, communities vary in terms of species richness, community evenness, and ecological function (Hayward and Venrick 1998; Venrick 2009; Taylor et al. 2015, James et al. 2022). Changes in function are often assumed in relation to changes in taxonomy, or estimated via bulk measurements (Taylor et al. 2015; Stukel et al. 2017). However, these methods remain incapable of resolving the direct connections between species and their function. Metatranscriptomics provide a dynamic view of ecological communities, highlighting those species that are actively expressing genes in response to environmental conditions, and in doing so provides a tool for how changes in community composition can affect spatio-temporal patterns in community function (Dupont et al. 2015; Harke and Gobler 2015; Landa et al. 2017; Kolody et al. 2019; 2022).

Marine metatranscriptomes have been utilized to explore a variety of ecological questions, from changes in community function in day-night cycles (Kolody et al. 2019), global patterns of community function (Vorobev et al. 2020), and niche partitioning within ecologically

important taxa, such as diatoms (Alexander et al. 2015). However, the scope of these studies has often been restricted taxonomically, functionally, or as a result of constraints to the sampling scheme and thus many general ecological inquires remain unexplored.

Here we use a seven-year regional dataset of marine microbial metatranscriptomes sampled as part of the quarterly CalCOFI (California Cooperative Fisheries Investigation) surveys. These 323 metatranscriptomic samples are part of the NOAA CalCOFI Ocean Genomics project (NCOG, James et al. 2022). Data is collected at each station from both the surface and deep chlorophyll maximum layer with stations spanning from the nutrient-rich nearshore to the oligotrophic offshore (Fig. 4.1, Bograd, Schroeder, and Jacox 2019). Given the contrasting nearshore and offshore environments, these data capture relatively large gradients in temperature and nutrient supply within the surface ocean (Fig. 4.1S). As previously highlighted, major ecological functions in the marine environment are set by the microbiome and are a combination of which taxa are there and the functions those organisms carry out. The goals of this study will therefore be twofold, to identify the patterns and processes that shape the active microbial community and to understand the biogeography of functional gene classes and their relationship to both community structure and environmental conditions. We attempt to address the first question by identifying the niche optimums of taxa across a multitude of environmental gradients including: temperature, salinity, $NO_3$, $PO_4$, $SiO_3$, nitracline depth, mixed layer depth, and sample depth. Next, we explore the biogeographies of functional gene classes using the EuKaroyotic Orthologous Groups (KOG) database. We explore whether environmental gradients align with overall shifts in the relative abundance of KOG classes and examine the individual relationships between KOG classes and environmental covariates.

Across all taxa, we find that niche optimums are significantly different than mean environmental conditions, with taxa more active and abundant in warmer, nutrient poor waters. This aligns with previous findings from metabarcoding work which identified higher diversity in the oligotrophic stations of the region (James et al. 2022). Niche optimums across both temperature and nitracline depth were unique as they were both variable (optimums were more distinct from one another) and selective (per taxa abundances declined significantly away from their respective optimums). This pattern was observed not only across all taxa but within select groups including heterotrophic bacteria, cyanobacteria, eukaryotic phytoplankton, and heterotrophic eukaryotic protists, indicating that temperature and nitracline depth lead to strong niche partitioning across and within many taxonomic groups of marine microbes. Across all taxa, functional biogeography was driven by large changes in the relative expression of genes related to energy production and conversion, which were most correlated with changes in salinity. Most other KOG classes were negatively correlated with increases in temperature and salinity, indicating increased functional diversity in the nutrient-rich nearshore. The majority of these shifts occurred across the entire community with the response shared across taxa. However certain groups, like eukaryotic phytoplankton shifted dramatically in terms of function in association with changes in community evenness. Combined, these analyses capture the direct link between environmental conditions and community structure and function and in doing so uncover the dynamics that determine key ecological services at the regional scale.

## 4.2 Results

The NCOG metatranscriptomic data analyzed in this study consists of 323 samples from 2014-2020 collected within the Southern California Current (SCC) region. Samples were collected at the surface and deep chlorophyll maximum (DCM) at each station (Fig. 4.1). For this

study we examined open reading frames (orfs) whose best taxonomic hit aligned with known prokaryotic and eukaryotic taxa. We identified 1,998 distinct taxa categorized into the following five groups: archaea (60), heterotrophic bacteria (1,571), cyanobacteria (76), eukaryotic phytoplankton (159), and heterotrophic eukaryotic protists (132).

We used these 1,998 taxa to assess the taxonomic niche partitioning of marine microbes within the SCC, with niches defined by eight environmental variables (temperature, salinity, NO3, PO4, SiO3, mixed layer depth, nitracline depth, and sample depth Fig. 4.2, Fig. 4.1S). For each taxon, we calculated the total number of transcripts per liter per sample. We used these values as our "weighting" to identify niche optimums, for each taxon across all environmental covariates (example shown in Fig. 4.2a, see Eq. 3 in methods). For each taxon, we also calculated the slope between total transcripts $L^{-1}$ and environmental distance to the optimum, where negative slopes indicate decreases in population size away from the optimum. Slopes that are not significantly negative ($p > 0.05$) or positive were set to zero. To make the value more intuitive, we take the absolute value of the slope following the previous step and refer to this value as habitat specificity ($\gamma$), with large positive values indicating the most habitat-specific relationships and zero values indicating no habitat-specific relationships. Following the identification of individual niche optimums, we assessed the distribution of optimums across all eight environmental variables (Fig. 4.2b, Fig. 4.2S). Distributions can vary in terms of their means (different from the mean environmental conditions), variability (narrower or wider), and internally (whether taxa show habitat specificity to their optimums). For example, a narrow distribution (low variability) could be the result of many taxa with similar niche optimums and high habitat specificity (Fig. 4.2c) or low habitat specificity (Fig. 4.2d). Similarly, a wide

distribution, where optimums range widely across an environmental gradient may occur alongside high (Fig. 4.2e) or low (Fig. 4.2f) habitat specificity on the taxon level.

We first assessed whether the means of the distributions were greater or less than the mean of the environmental variables (two-sided t-test, p-value < 0.01). We found that every distribution of optimums was significantly greater or less than the mean of the associated environmental gradient (Fig 3a). Temperature and nitracline depth optimums were higher than the mean environmental conditions. All other variables had lower optimum distributions compared to the means of the respective environmental variables. The variability of optimum distributions was different across all environmental parameters. Salinity had the highest variability while nutrients (NO3, PO4, and SiO3) and depth had the lowest variability across optimums (Fig. 4.3a). Within our broad groups, archaea had the fewest significant differences between optimums and the mean environment (Fig. 4.3Sa). Salinity optimums were most variable in heterotrophic bacteria (Fig. 4.3Sc) and less so for all other groups, suggesting that the overall variability in salinity optimums was largely attributable to variability amongst heterotrophic bacteria. Eukaryotic phytoplankton showed strong preferences for shallow nitracline, mixed layer, and sample depth, all much shallower than their respective means (Fig. 4.3Sg). Overall, some patterns were consistent across groups. For instance, temperature optimums were always warmer than the mean temperature (though this wasn't significant in archaea).

Finally, we compared the variability of optimum distributions to the mean habitat specificity across all environmental parameters (Fig. 4.3b, see Fig. 4.4S for distributions of habitat specificity). While salinity had the highest variability across all environmental parameters, it also had the lowest mean habitat specificity, suggesting that optimums do not

necessarily align with changes in abundance, and that populations are not found in particular salinity ranges. In contrast, both temperature and nitracline depth had intermediate levels of variability coupled with the highest mean values of habitat specificity. This could be indicative of niche partitioning, where taxa fall into relatively narrow individual niches distributed across a wide variety of both temperatures and nitracline depths. Nutrients optimums tended to have relatively low variability but had intermediate levels of habitat specificity. Since nutrients are taken up rapidly within the surface ocean, this weaker signal in both variability and habitat specificity when compared to nitracline depth was likely due to a mismatch between current population size and current nutrient levels, the latter of which isn't necessarily indicative of the conditions responsible for the observed population sizes. Across the broad groups both temperature and nitracline depth optimums tended to show intermediate levels of variance and some of the highest levels of habitat specificity (Fig. 4.3S). Within eukaryotic phytoplankton, both depth and nitracline depth had high levels of habitat specificity, highlighting that many eukaryotic phytoplankton are more active and abundant at shallower depths (surface vs DCM) and in regions with shallower nitracline depths (nearshore).

Following the analysis of taxonomic niche distributions, we identified relationships between the environment and the functional response of the marine microbial community. Using the EuKaroyotic Orthologous Groups (KOG) database, orfs were binned into 23 functional classes of genes. We first assessed the relationship between the relative abundance of the 23 KOG classes and our environmental parameters. Since our response variable was multinomial (23 classes) we used Dirichlet regressions to assess the relationships between environmental variables and the relative abundances of KOG classes (Fig. 4.4). Across all taxa, we found that salinity was most highly correlated with the relative abundance of all 23 KOG classes (Fig. 4.4a).

Within three of the five broad taxonomic groups, temperature was the most highly correlated with changes in the relative abundance of KOG classes, followed by nitracline depth and salinity (Fig. 4.4b). Some groups showed alternative relationships, such as Archaea, whose function was most highly correlated with gradients in PO4 and cyanobacteria, where functional response aligned with changes in depth. Across all taxa, the largest changes in function were primarily driven by the relative abundances of genes associated with energy production and conversion, which decreased with increasing salinity. Since many KOG classes are relatively low abundance compared to the major groups, we also explored the correlations between individual KOG classes and environmental parameters (Fig. 4.5a). Most KOG classes showed significant negative relationships with both temperature and nitracline depth indicating an increased importance of most functional classes relative to energy production and conversion in cooler temperatures and shallower nitracline depths.

To test whether changes in functional composition were the result of community wide responses or shifts in taxonomy we explored two community metrics: 1) the correlation between the per-sample total relative abundance of a KOG class across all taxa and the per-sample mean relative abundance of a KOG class across taxa and 2) the community evenness, measured by total transcripts $L^{-1}$ per taxon (Fig. 4.5b). Across all taxa the trend in energy production and conversion appeared to be strongly associated with a community-wide response. In contrast, few patterns emerged because of large shifts in community evenness. Across all KOG classes we observed a handful of conserved seasonal patterns. The relative abundance of energy production and conversion was lowest in the spring and relatively consistent through the other seasons. In contrast, KOG groups like amino acid transport and metabolism, RNA processing and modification, and coenzyme transport and metabolism showed increased relative abundance in

the spring. Another pattern common to many KOG groups was an increased relative abundance in the fall, which occurred for many KOG groups such as cell motility, defense mechanisms, and nuclear structure (Fig. 4.5c). Finally, across all seven years the relative abundances of KOG classes were quite stable. From 2018-2019 we observed a dip in the relative abundance of energy production and conversion which coincided with an increase in most other categories during this time (Fig. 4.5d).

Beyond the patterns observed across all taxa, we also explored the relationships between KOG classes and environmental variables within each of the five broad taxonomic groups (Fig. 4.5-9S). Within archaea, many relationships between individual KOG classes and environmental variables were not significant. Most archaeal KOG classes showed community-wide functional shifts with little to no change in community evenness (Fig. 4.5S). Within heterotrophic bacteria the correlation structure between KOG classes and environmental parameters was similar to the structure observed across all taxa. Overall, most changes in functional relative abundance appeared to occur community wide. However, some functions such as RNA processing and modification and inorganic ion transport and metabolism appeared to be driven by changes in the community—increases in the relative abundance of these KOG classes coincided with decreases in community evenness (Fig. 4.6S). Cyanobacteria again showed a similar correlation structure to heterotrophic bacteria, however, certain KOG classes had more prominent environmental relationships, such as carbohydrate transport and metabolism and posttranslational modification, protein turnover, chaperones (Fig. 4.7S). Across all broad groups, eukaryotic phytoplankton show the most drastic departure from the structure observed in other groups. Most KOG classes within eukaryotic phytoplankton showed significant correlations with environmental parameters. Strong correlations with shallower nitracline depths and higher nutrient concentrations occurred

across most Metabolism KOG classes. Notably, almost all functional KOG classes had strong negative correlations with eukaryotic phytoplankton community evenness (Fig. 4.8S). In other words, the most dominant taxa drive large-scale changes in the function of the community. Finally, within heterotrophic eukaryotic protist communities we again found strong relationships between KOG classes and environmental variables. However, unlike eukaryotic phytoplankton, the function of heterotrophic eukaryotes appeared to align with community-wide shifts in function rather than shifts in community evenness (Fig. 4.9S).

While KOG classes provide an informative mid-level description of functional changes to the community, there is immense gene diversity below the level of KOG class. To better understand which genes were responsible for observed shifts in community function we identified which orfs aligned most closely with the overall changes in relative abundance of each KOG class (Tables 4.1-3S). Aconitase, a component within the first steps of the TCA cycle appeared as one of the primary genes responsible for changes within the energy production and conversion KOG class. Likewise, Acetyl-CoA transporters which falls under the KOG class inorganic ion transport and metabolism, were presumably acting to transport Acetyl-CoA to the TCA cycle (Table 4.1S). Many of the genes within the broader KOG category of cellular processing and signaling have been shown to be upregulated in the evening in day-night cycle studies and within this study were positively correlated with depth across all taxa (Fig. 4.5a), providing further evidence that these genes may be less active under UV-stress (Kolody et al. 2019, Table 4.2S). Ribosomal proteins were common within the broad category of information processing and storage and are known to be strongly associated with cellular growth rates (Ottesen et al. 2013, Table 4.3S). As such, KOG classes within the broad category of information

processing and storage correlated with increased nutrients and shallow nitracline depths, indicating growth was maximal for many taxa in these nutrient replete conditions (Fig. 4.5a).

## 4.3 Discussion

The results presented here represent a fundamental step forward in spatial and temporal sampling of the active pelagic microbiome. The strong gradients in many environmental variables across the region from the meso/eutrophic nearshore to the oligotrophic offshore provided the ideal testing ground to observe spatial and temporal patterns in both taxonomic and functional biogeographies across prokaryotic and eukaryotic microbial assemblages. Previous work found that microbial community structure and diversity within the SCC was largely structured by the availability of nutrients to the surface ocean (where nitracline depth represents a proxy for nutrient availability) across space and time (James et al. 2022). Here we found that both temperature and nutrient availability most strongly structured the niches of active prokaryotic and eukaryotic taxa within the region (Fig. 4.3). While variables like salinity showed the highest variability in niche optimums, it appeared populations of most taxa do not decline significantly as they moved away from salinity optimums (Fig. 4.3b, Fig. 4.4Sb). In contrast, both temperature and nitracline depth showed intermediate levels of variability amongst niche optimums and relatively high habitat specificity (Fig. 4.3b). We found that this was not only the case across all taxa, but that these environmental gradients appeared to partition niches within a majority of the broad taxonomic groups as well (Fig. 4.3S).

One explanation for the increased importance of temperature could be that we are identifying niche optimums by a weighted centroid approach wherein the weightings are total transcripts $L^{-1}$ per taxon. Total transcripts $L^{-1}$ are the result of both abundance (more cells lead to more transcripts $L^{-1}$) and activity (per cell levels of activity may change according to

environmental conditions). As such, the niche optimum represents a combination of both abundance and activity and is likely to change as a result of environmental conditions more than abundance alone. While this blending of two factors could be seen as a disadvantage of the methodology, we believe that the identification of an "active" niche presents a unique lens with which to view the community structure—highlighting those active members that contribute to the *in-situ* magnitude of ecological functions (such as primary productivity and carbon sequestration) across the region. We note that dormant taxa referred to as the "microbial seed bank" can be important drivers of changes in community function and may not be detected as readily through metatranscriptomics alone, becoming active episodically in response to changing conditions (Lennon and Jones 2011; Gibbons et al. 2013). While this method may not capture their full biogeographic extent, we believe that in highlighting these taxa at their most abundant and active we may better identify the conditions during which these taxa contribute most to the overall functional landscape of the community.

Overall, we found that changes to the functional biogeography of the regional microbiome were most correlated with gradients in salinity (Fig. 4.4). Across disparate biomes, salinity has been shown to be a major determinant of bacterial diversity (Lozupone and Knight 2007). Within the marine environment however, global studies have found temperature and light are more highly correlated with diversity and function than salinity (Sunagawa et al. 2015). Within the SCC, salinity is an indicator of the major water masses that contribute to the region's waters (Bograd, Schroeder, and Jacox 2019). While we do not dismiss salinity's role in determining the community composition across microbes globally and between biomes, we believe that within this regional context, differences between water masses, which align with gradients in salinity, are most likely to structure the observed changes in functional composition.

Across all ASVs, changes in functional composition were largely driven by changes in the relative abundance of genes related to energy production and conversion (Fig. 4.4a). The saltiest waters within the region are part of the Pacific Equatorial Water (PEW) and tend to be subsurface and nutrient rich, upwelling to the surface in the nearshore (Bograd, Schroeder, and Jacox 2019). In these samples, the relative abundance of energy production and conversion reached its lowest point, while other KOG classes such as translation, ribosomal structure and biogenesis increased. In contrast, within the oligotrophic environment the bulk of transcription occurred within the energy production and conversion KOG class, likely as the result of limited resources making other functions too metabolically expensive.

In general, shifts in function appeared to occur across the entire community, with taxa responding to gradients in environmental conditions in similar ways (Fig. 4.5b). However, eukaryotic phytoplankton exhibited a different pattern: shifts in the relative abundance of most functional classes were negatively correlated with community evenness (Fig. 4.8S). Within the nearshore environment, a select few eukaryotic phytoplankton tend to be dominant at an given time or location, forming blooms (Not et al. 2012; Needham and Fuhrman 2016). As such these blooms can significantly alter the functional landscape within the nearshore environment. Offshore, eukaryotic phytoplankton communities tended to be more even, and energy production and conversion represented a larger proportion of the relative abundance of transcripts, aligning with observations across all taxa. Eukaryotic phytoplankton, such as diatoms, are important ecologically as they shape the magnitude of many regional and global ecological functions generated within the marine microbiome (Taylor et al. 2015; Malviya et al. 2016; Tréguer et al. 2017). That functional shifts within eukaryotic phytoplankton are driven by only a few taxa

should provide further impetus to better understand the resilience of these ecologically crucial taxa.

Regionally comprehensive transcriptional datasets such as this one represent a major leap forward in our understanding of how marine microbiomes respond to environmental conditions. Here we demonstrate a potential path forward for exploring the entire community assemblage (prokaryotes and eukaryotes) through the active lens provided by metatranscriptomes. We find that temperature and nitracline depth drive separation of the niche optimums of taxa, while simultaneously leading to high intra-taxon habitat specificity. We identify salinity, a marker of various regional water masses, as the strongest correlate of community-wide shifts in the relative abundance of functional gene classes—a pattern driven by large shifts in the relative abundance of genes related to energy processing and conversion. Finally, we note that changes in the relative abundance of functional gene classes within eukaryotic phytoplankton assemblages often aligns with changes in community evenness, indicating that functional shifts in this ecologically important community are driven by few, dominant taxa. Combined, these analyses aim to address fundamental ecological questions about taxonomic niche partitioning and functional biogeography but at a far greater resolution than has previously been possible, setting the framework for how transcriptional data can be used to unravel longstanding inquiries in the field.

## 4.4 Methods

### 4.4.1 Study location and sample collection

The NOAA CalCOFI Genomics Project (NCOG) metatranscriptome data was collected from 2014-2020 within the Southern California Current (SCC) region. This region is part of a highly productive eastern boundary current that is both ecologically diverse and economically

important. For this study we examined 323 samples that were collected from the surface (0m) to 150m. NCOG data is collected quarterly (winter, spring, summer, and fall).

Metatranscriptome and environmental data were collected with a CTD rosette. Temperature and salinity were measured with the Seabird 911 CTD. Salinity measurements were compared to bottle samples that were measured with a Guildline Portasal Salinometer model 8410A. Nutrients were measured with a QuAAtro continuous flow autoanalyzer (SEAL analytical).

### 4.4.2 RNA collection, extraction, and sequencing

0.2-2.2 L of seawater was filtered through a 0.22 µm Sterivex-GP filter unit (MilliporeSigma, Burlignton, MA, USA) for RNA samples. Samples were immediately sealed with a sterile luer-lock plug and hematocrit sealant, wrapped in aluminum foil, and flash frozen in liquid nitrogen. For a full step-by-step sampling procedure see: https://www.protocols.io/view/noaa-calcofi-ocean-genomics-ncog-sample-collection-eq2lypdorlx9/v1

RNA extraction is done using a Macherey-Nagel NucleoMag RNA kit. Automated liquid handling was performed on an eppendorf EpMotion 5075t with multi-hcannel pipettes. For full extraction procedure see: https://www.protocols.io/view/sterivex-rna-extraction-n92ldy277l5b/v1

From 2014-2019, 100ng of total RNA as input, ribosomal RNA was removed using Ribo-Zero Magnetic kits (Illumina). We modified the composition of Removal Solutions with the mixture of plant, bacterial, and human/mouse/rat Removal Solution in a ratio of 2:1:1. Agilent TapeStation 2200 checked the quality of rRNA removal RNA. The rRNA-deplete total RNA was used for cDNA synthesis by Ovation RNA-Seq System V2 (TECAN, Redwood City, USA).

Double stranded cDNA was fragmented using Covaries E210 system with the target size of 300bp. 100ng of fragmented cDNA as input into the Ovation Ultralow System V2 (TECAN, Redwood City, USA) following the manufactures protocol. Ampure XP beads (Beckman Coulter) were used for final library purification. Library quality was analyzed on a 2200 TapeStation System with Agilent High Sensitivity DNA 1000 ScreenTape System (Agilent Technologies, Santa Clara, CA, USA). Resulting libraries were subjected to paired-end Illumina sequencing.

Due to the discontinuing of Ribo-Zero magnetic kits (Illumina), we used riboPOOL for rRNA deletion for 2020 samples. Thus, for 2020 samples 80ng of total RNA as input, ribosomal RNA was removed using riboPOOL Seawater kit (Galen Laboratory Supplies, North Haven, Connecticut, USA). The riboPOOL Seawater kit is customized for us with the composition of Removal Solutions with the mixture of Pan-Prokaryote riboPOOL, Pan-Plant riboPOOL and Pan-Mammal in a ratio of 6:1:1. Agilent TapeStation 2200 checked the quality of rRNA removal RNA. The rRNA-deplete total RNA was used for cDNA synthesis by Ovation RNA-Seq System V2 (TECAN, Redwood City, USA).

For 2020 samples, Double stranded cDNA was fragmented using Covaries E210 system with the target size of 300bp. 50ng of fragmented cDNA as input into the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA, USA) following the manufactures protocol. Ampure XP beads (Beckman Coulter) were used for final library purification. Library quality was analyzed on a 2200 TapeStation System with Agilent High Sensitivity DNA 1000 ScreenTape System (Agilent Technologies, Santa Clara, CA, USA). Resulting libraries were subjected to paired-end Illumina sequencing.

### 4.4.3 RNASeq assembly and annotation

Input paired-end fastq sequences were trimmed of any adapters, primers and low quality bases by using either blastall program (NCBI, v2.2.25) (Altschul et al. 1990) or single-step fastp program with trimmomatic option (fastp, v0.22.0 (Chen et al. 2018); trimmomatic, v0.36) (Bolger, Lohse, and Usadel 2014)). The trimmed paired and unpaired sequences were then depleted of rRNA sequences with riboPicker, v0.4.3 (Schmieder, Lim, and Edwards 2012) and separated from RNA Spike-in standards 1 and 8 (ThermoFisher).

The command-line program clc assembler, v5.2.1 (Qiagen) was used to assemble processed sequences into contigs in the following sequence: assembled individual libraries; assembled contigs (minimum contig size: 200 bases) from individual assemblies in groups (group = samples in the same cluster by their 18S rRNA content) and merged the resulting contigs with cd-hit (Li and Godzik 2006) requiring a minimum overlap of 300 bases and sequence identity of 0.95; finally, assembled merged group assemblies into a global set of contigs using the same method described for merging individual assemblies into group-specific contigs. The preceding steps are performed in two separate analyses for NCOG libraries generated in years 2014 through 2018 and later in 2019-2020, following which transabyss-merge program (Robertson et al. 2010) was used to merge the two sets of assemblies into one with the settings of '--mink 23 --maxk 31'.

Open reading frames (orfs) were generated from the contigs in the global assembly by using ORF-caller FragGeneScan, v1.31 (Rho, Tang, and Ye 2010). Trimmed sequences were mapped to the predicted ORFs using the command-line program clc mapper, v5.2.1 (Qiagen) to generate mapped read counts for each ORF. The identified ORFs in the assembled global contigs were annotated by Timelogic© tera-blastp (Active Motif Inc., Carlsbad, CA), HMMER, v3.3.2 (Eddy 2011), kofamscan, v1.3.0 (Aramaki et al. 2020) analysis programs using PhyloDB (JCVI,

internal), PFAM [https://pfam.xfam.org/], TMHMM (Krogh et al. 2001), KOFAM/KEGG (https://github.com/takaram/kofam_scan), transporters (JCVI, internal), organelle (JCVI, internal) and KOG (https://mycocosm.jgi.doe.gov/help/kogbrowser.jsf) (Tatusov et al. 2003) databases. The ORFs were assigned to the best taxonomic species/group as determined by LPI (Lineage Probability Index) (Podell and Gaasterland 2007) generated from the BLAST search of the taxonomy subset of PhyloDB database.

### 4.4.4 Transcripts per liter calculation

Transcripts per liter were calculated using spike-ins quantities of known standards. Following the calculation of transcripts per million (TPM) within each sample we calculated the ratio of expected versus observed TPM values of each spike per sample. We calculated the average ratio between the two spikes per sample, then divided this ratio by the known volume of seawater filtered per sample to get a scalar between TPM and the actual number of transcripts per liter. To get transcripts per liter (TPL) per sample, we multiplied the sample specific scalar value by the TPM per orf for each sample, to get a final TPL value for each orf per sample.

### 4.4.5 Data normalization

Each environmental variable, E, was normalized to a $\mu = 0$ and $\sigma = 1$ with the following equation:

$$E_{i,j} = \frac{\left(E_{i,j} - \bar{E}_i\right)}{\sigma_i} \quad (1)$$

across *i* environmental variables (E) and *j* samples.

To calculate habitat specificity, Transcripts $L^{-1}$ (TPL) were also normalized to a $\mu = 0$ and $\sigma = 1$ with the following equation

$$TPL_{i,j,k} = \frac{\left(TPL_{i,j,k} - \overline{TPL}_{i,k}\right)}{\sigma_{i,k}} \quad (2)$$

Across *i* environmental variables, *j* samples, and *k* taxa.

### 4.4.6 Calculating niche optimums (weighed centroid)

Niche optimums were calculated for each taxon across each environmental variable. Optimums were calculated using a weighted centroid approach where the "weighting" was the total transcription of a taxon. The equation for niche optimums is as follows:

$$Niche\ Optimum_{i,k} = \frac{\sum_{j=1}^{n} E_{i,j} \times k_j}{\sum_{j=1}^{n} k_j} \quad (3)$$

Where $E_i$ represents a given environmental variable scaled to ($\mu = 0$, $\sigma = 1$, Eq. 1), *k*, represents the total transcripts $L^{-1}$ for a taxon, and *j*, represents a sample from 1 to *n*, where *n* is the total number of samples (323).

### 4.4.7 Dirichlet regression for relative expression of KOG classes

To explore which environmental parameters best predicted the relative abundance of our 23 KOG classes we ran a Dirichlet regression using the DirichletReg package (Maier 2014) in R. Unlike a binomial logistic regression which is used to fit a dichotomous response variable (such as Yes/No), a Dirichlet regression allows us to fit a multinomial response variable—which in this case is the relative abundance of our 23 KOG classes. For this study we compared the fits of singular explanatory variables using Akaike Information Criterion (AIC) values, where the lowest value indicates the variable with the best fit.

## 4.5 Figures

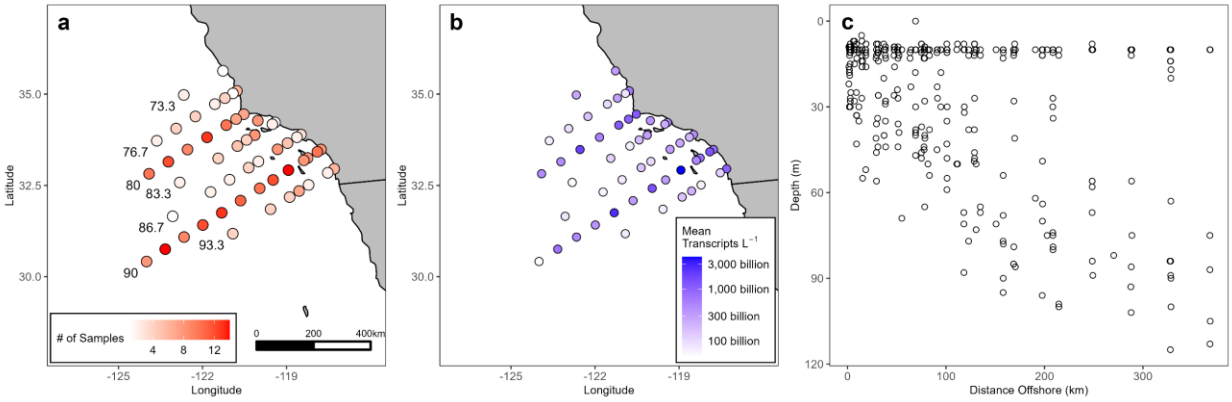


**Figure 4.1**: a, number of samples per station from 2014-2020 b, mean total transcripts $L^{-1}$ per station c, vertical and horizontal location of samples. Samples were collected at the surface and deep chlorophyll maximum (DCM). For this study we explored only samples less than 150m deep.

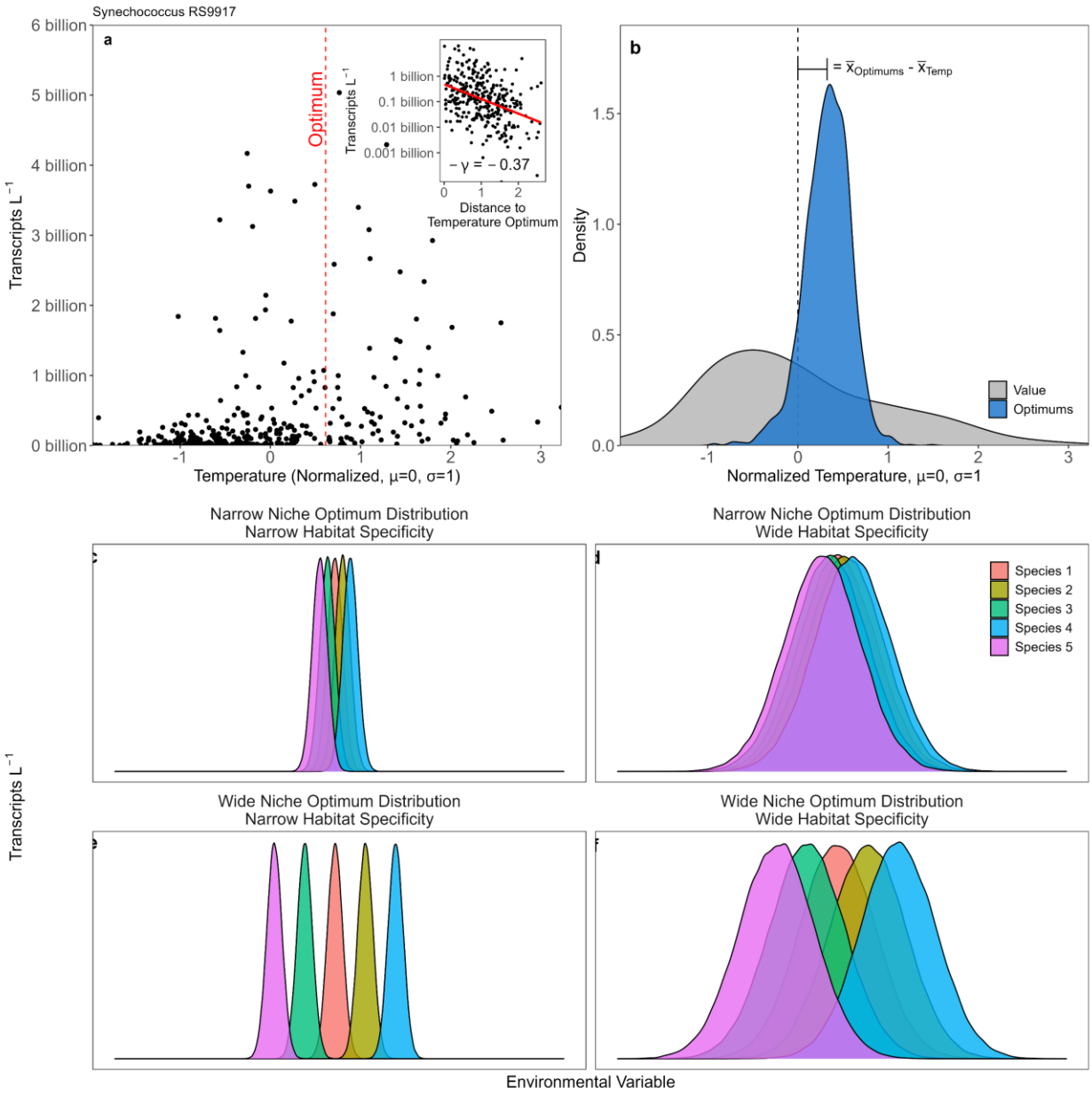


**Figure 4.2**: a, example relationship between Temperature (normalized to μ = 0, σ = 1) and total transcripts $L^{-1}$ for Synechococcus RS9917 across all 323 samples. Dashed red line indicates the weighted centroid of temperature where the weighting is the total transcripts $L^{-1}$ for this taxon. Inset graph shows the relationship between Total transcripts $L^{-1}$ and the distance from the centroid. Pearson's correlation is used to assess the significance of this relationship and is used to determine the habitat specificity of each taxa across each environmental variable. b, Distribution of normalized temperature values (grey) across all 323 samples and the distribution of weighted centroids for temperature (blue) across all 1,998 taxa c-f, example distributions of Transcripts $L^{-1}$ across environmental parameters. c, Narrow distribution of niche optimums between species, each with steep declines in abundance away from the optimum. d, Narrow distribution of niche optimums between species, but with less of a decrease in abundance away from the optimum. e, Wide distribution of niche optimums between species, each with steep declines in population away from the optimum f, Wide distribution of niche optimums between species, but with less of a decrease in abundance away from the optimum.

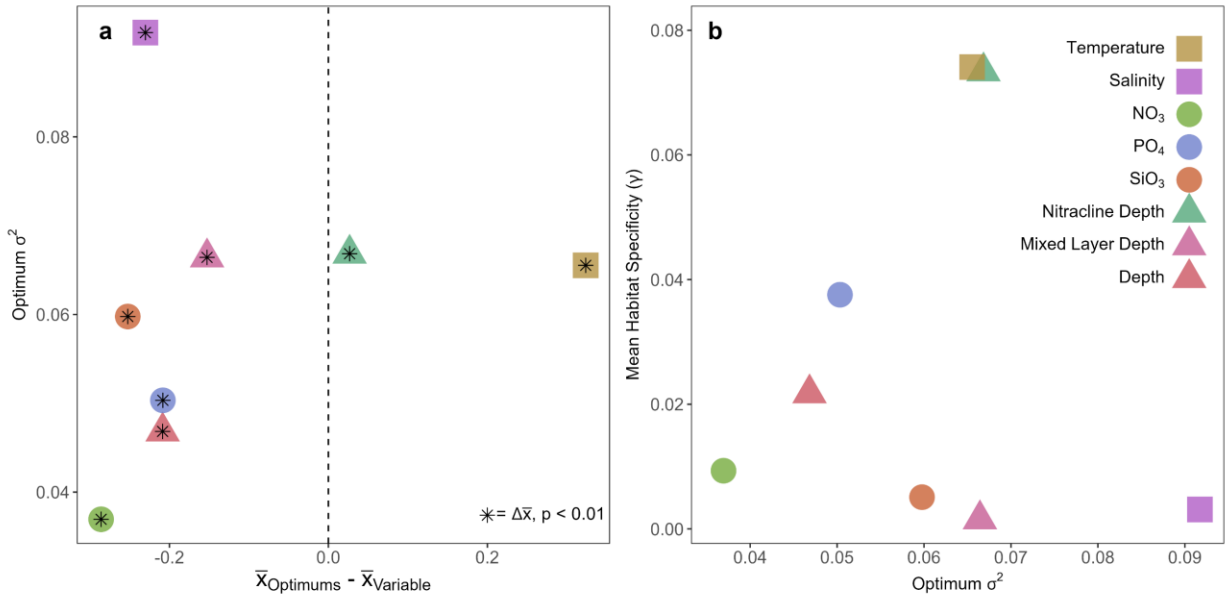


**Figure 4.3**: a, summary of weighted centroid distributions across all variables. X-axis represents the difference between the mean of the weighted centroids versus the mean of the environmental variable (asterisks represent p-values < 0.01, two-tailed t-test). Y-axis shows the variances across all weighted centroid distributions. b, relationship between the variance of weighted centroid distribution and the mean habitat specificity for each environmental variable.

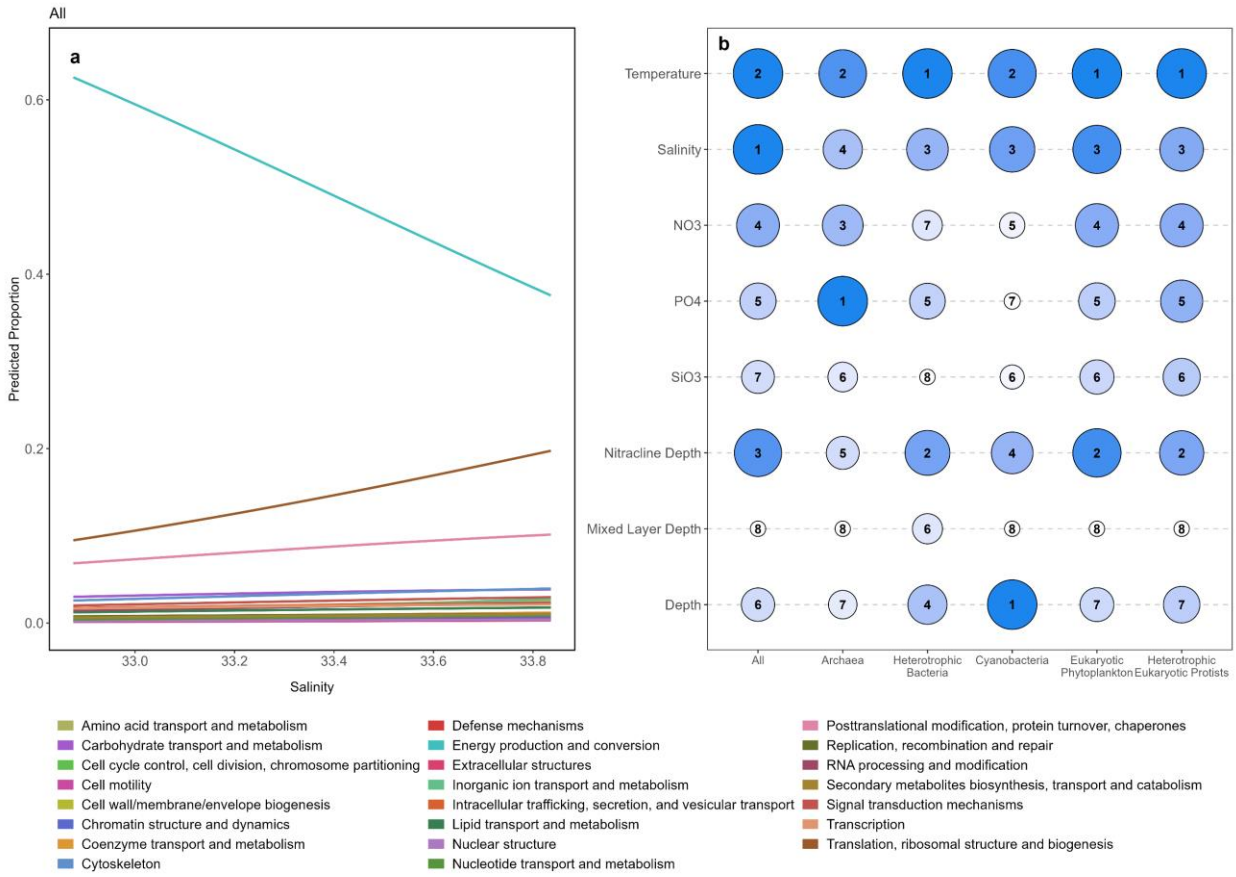


**Figure 4.4**: a, example relationship between salinity and the 23 KOG classes for all 1,998 taxa. Relationship was fit with a Dirichlet regression. b, table of AIC values showing which variables are most predictive of the relative abundance of all 23 KOG classes (Dirichlet regression). Largest, darkest circles represent the lowest AIC scores. Scores are also ranked from 1-8 where 1 represents the most significant relationship and 8 represents the least significant relationship.

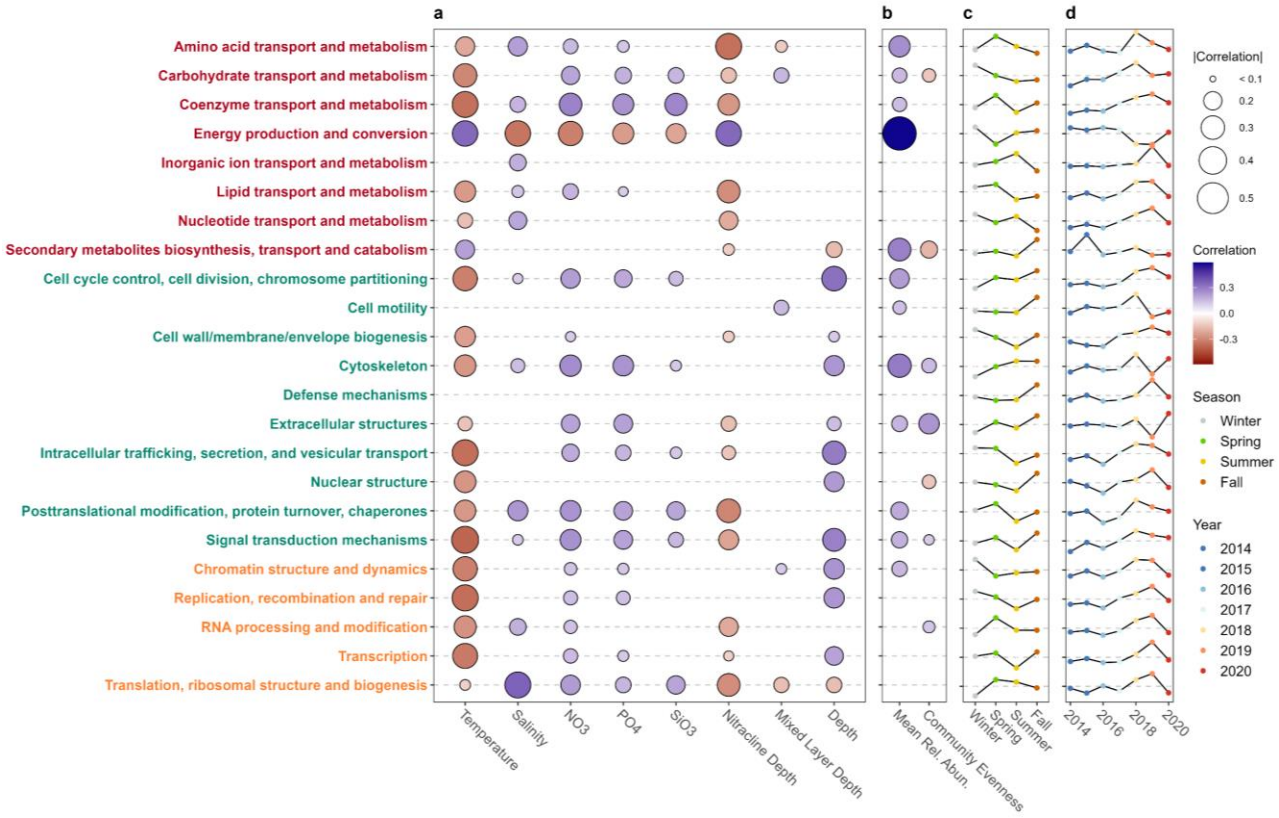


**Figure 4.5**: Individual KOG Class summary a, Relationship between individual KOG classes and physical and chemical variables. Colors represent the direction and magnitude of correlation while the size of the circles represents the absolute magnitude of the correlation. No circle indicates the relationship is not significant. b, Relationship between the relative abundance of a KOG class ecological community parameters (mean relative abundance and group evenness). Where mean relative abundance is calculated as the mean relative abundance of a given KOG class across all prokaryotic and eukaryotic genera and group evenness is calculated as the Shannon Evenness Index for total transcription of all eukaryotic phytoplankton genera. c, Mean seasonal trend in the relative abundance of KOG classes d, yearly trend in the relative abundance of all KOG classes.

## 4.6 Supplementary Information

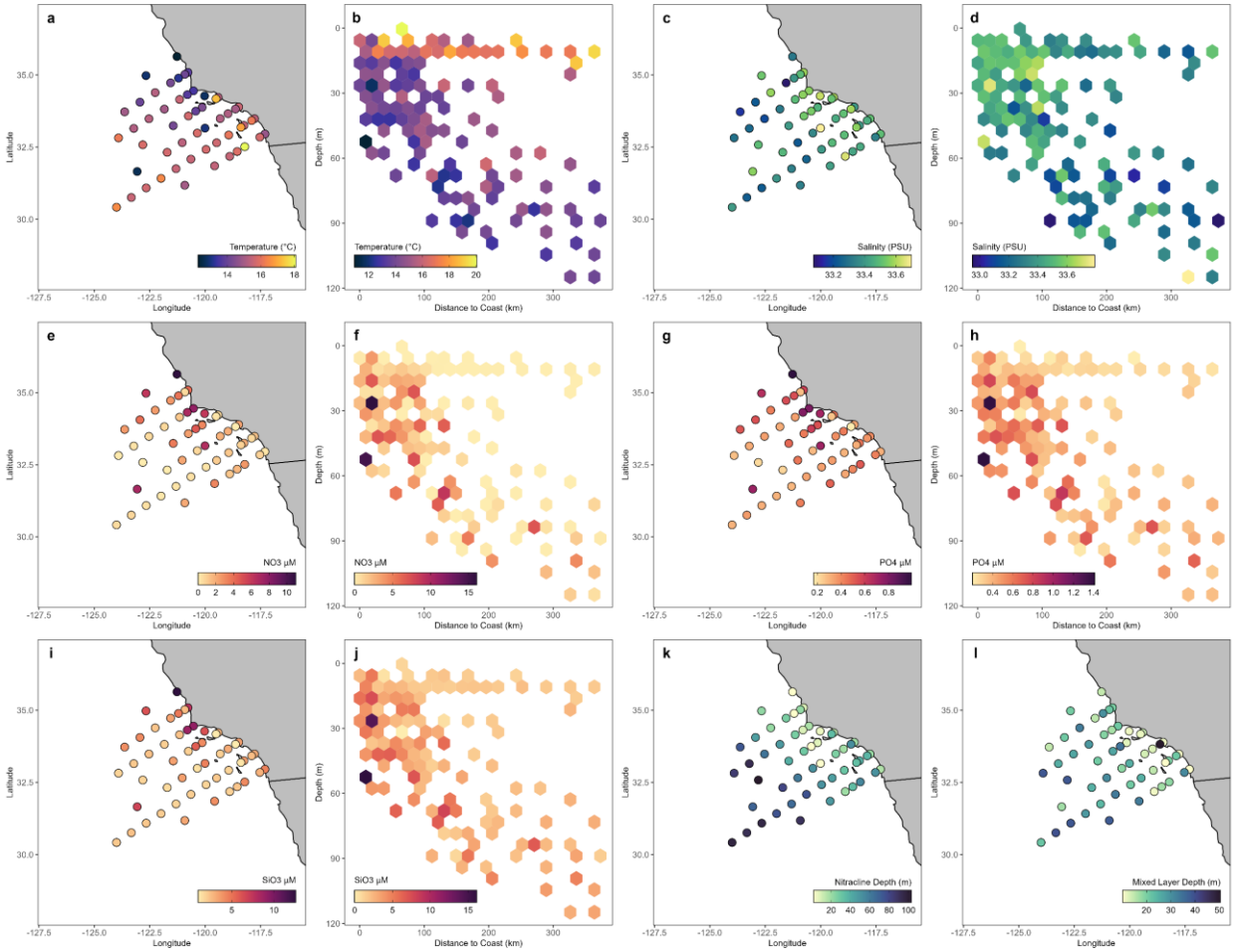


**Figure 4.1S**: Maps highlighting spatial gradients in each environmental parameter (a Temperature, c Salinity, e NO3uM, g PO4uM, i SiO3uM, k Nitracline Depth, and l Mixed Layer Depth). Cross sectional view highlighting gradients in environmental variables across both depth and distance offshore (b Temperature, d Salinity, f NO3uM, h PO4uM, j SiO3uM).

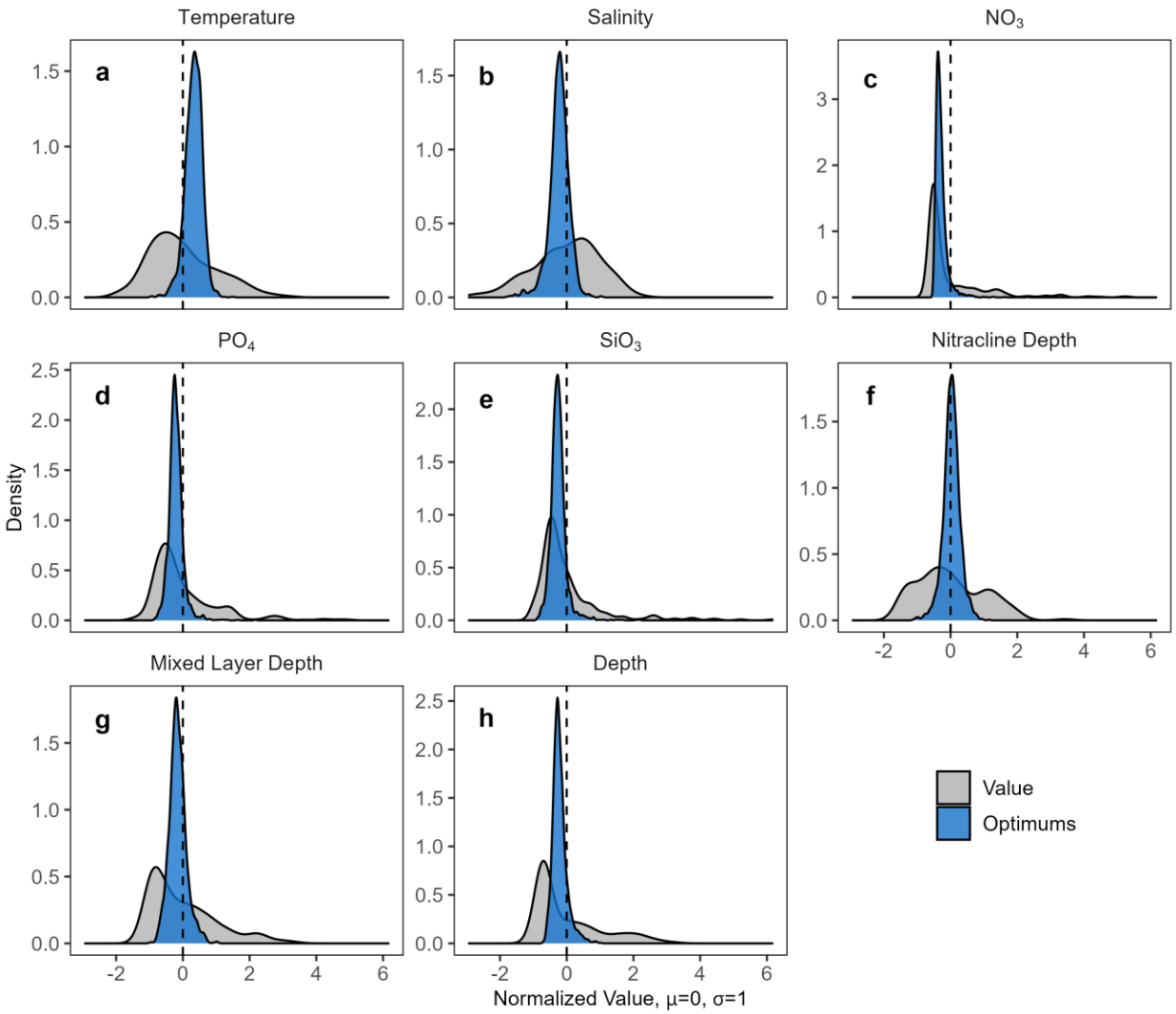


**Figure 4.2S**: Distributions of normalized environmental values (grey) across all 323 samples and the distribution of weighted centroids for each environmental value (blue) across all 1,998 taxa. Environmental variables include: a temperature, b Salinity, c NO3, d PO4, e SiO3, f nitracline depth, g mixed layer depth, and h depth.

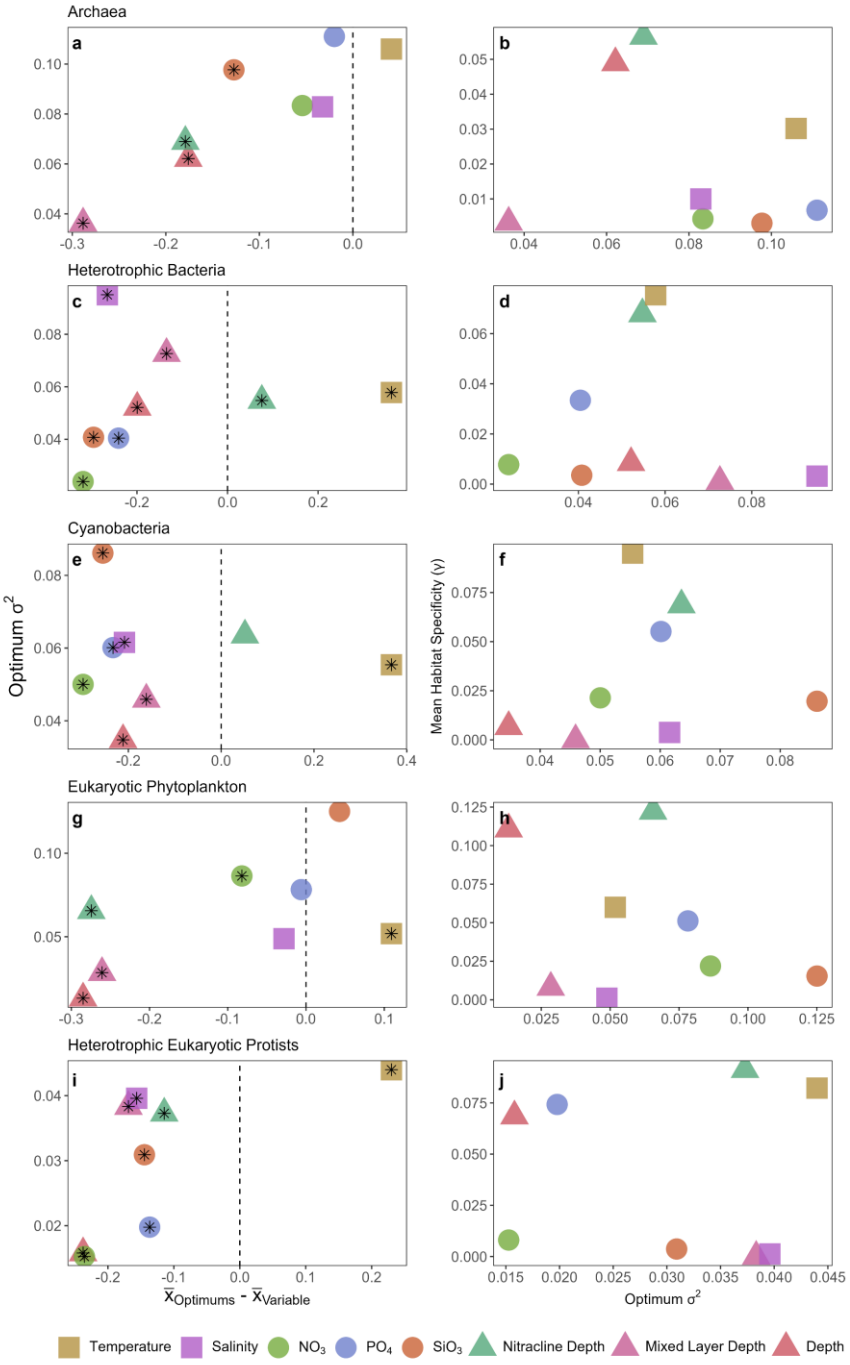


**Figure 4.3S**: Summary of weighted centroid distributions across all variables for the five major groups (a Archaea, c Heterotrophic Bacteria, e Cyanobacteria, g Eukaryotic Phytoplankton, and i Heterotrophic Eukaryotic Protists). X-axis represents the difference between the mean of the weighted centroids versus the mean of the environmental variable (asterisks represent p-values < 0.01, two-tailed t-test). Y-axis shows the variances across all weighted centroid distributions. Relationships between the variance of weighted centroid distribution and the mean habitat specificity for each environmental variable for each of the five major groups (b Archaea, d Heterotrophic Bacteria, f Cyanobacteria, h Eukaryotic Phytoplankton, and j Heterotrophic Eukaryotic Protists).

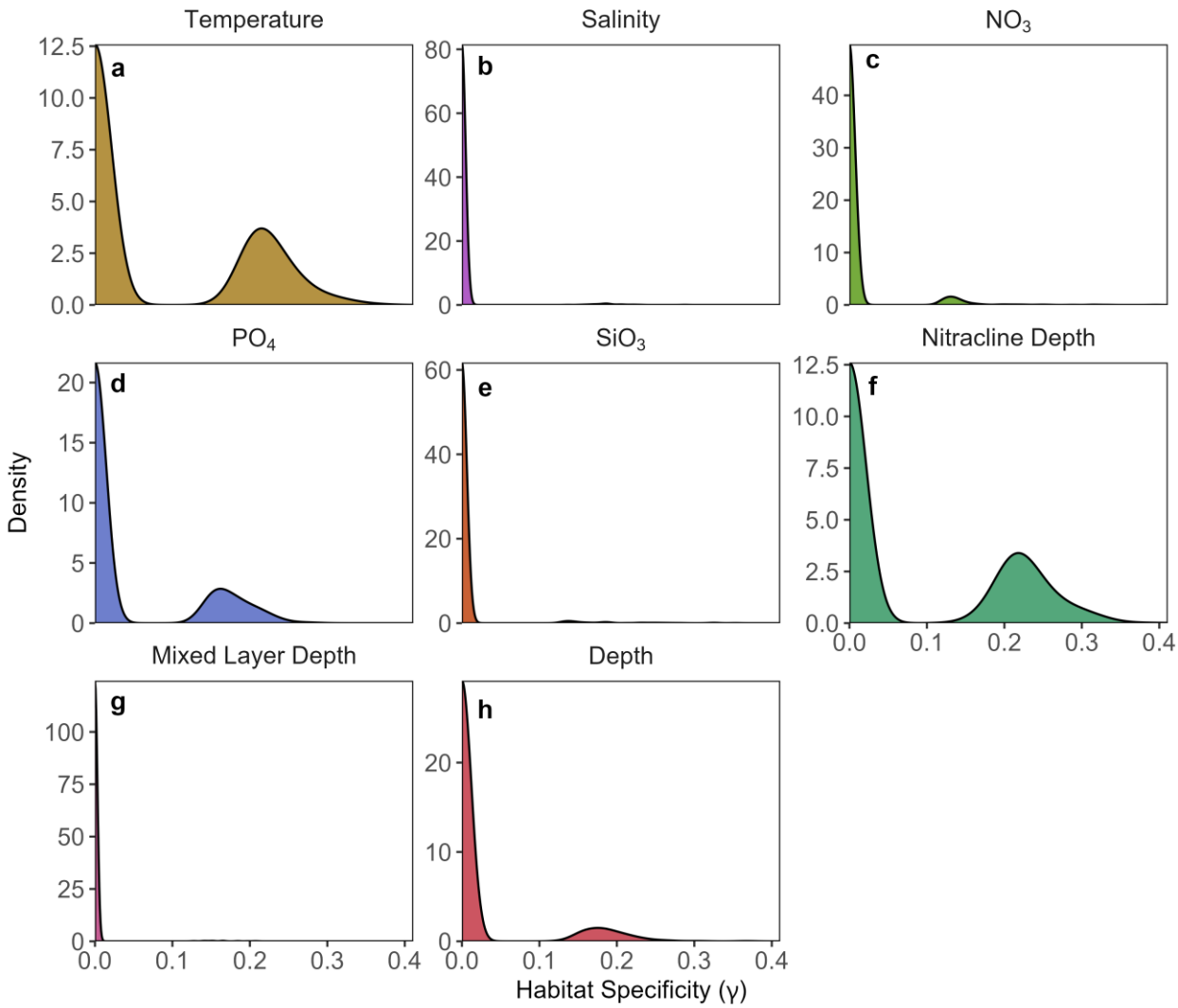


**Figure 4.4S**: Distributions of habitat specificity across all 1,998 taxa within each environmental variable. Environmental variables include: a temperature, b Salinity, c NO3, d PO4, e SiO3, f nitracline depth, g mixed layer depth, and h depth.

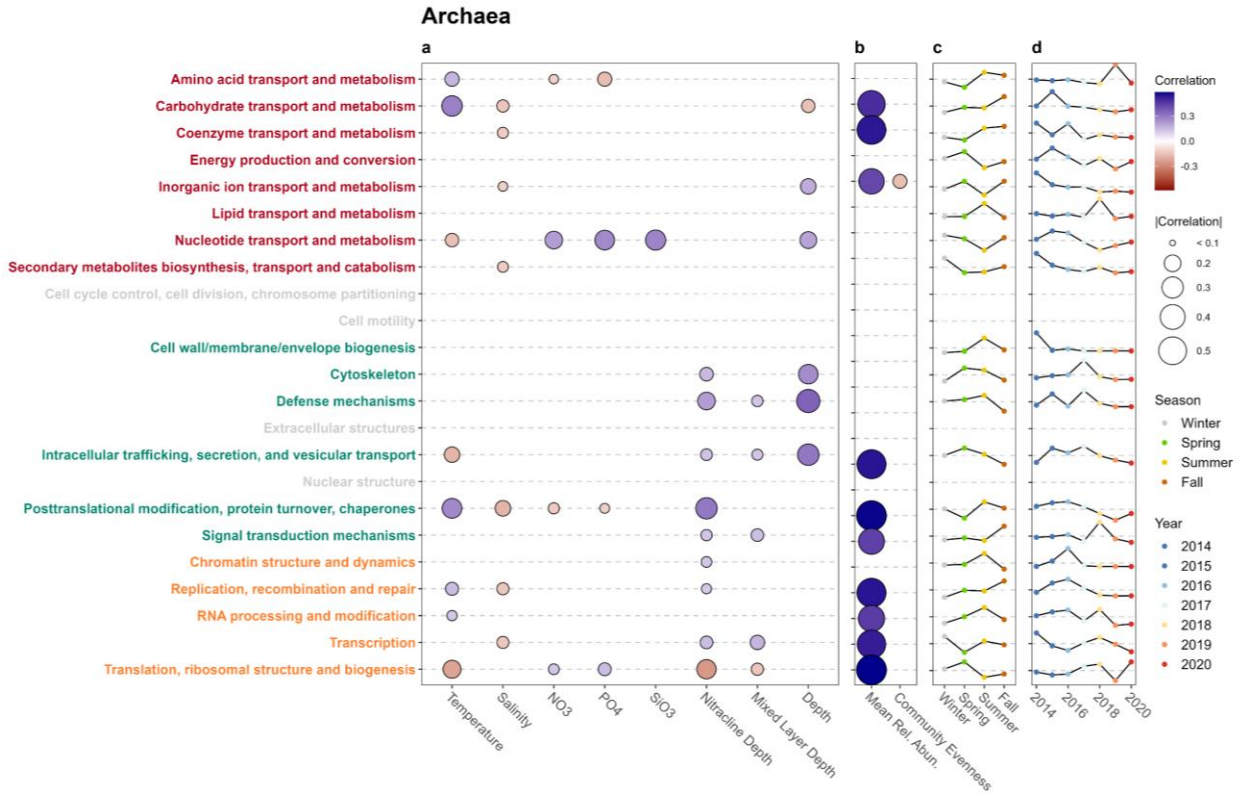


**Figure 4.5S**: Individual KOG Class summary for Archaea. a, Relationship between individual KOG classes and physical and chemical variables. Colors represent the direction and magnitude of correlation while the size of the circles represents the absolute magnitude of the correlation. No circle indicates the relationship is not significant. b, Relationship between the relative abundance of a KOG class ecological community parameters (mean relative abundance and group evenness). Where mean relative abundance is calculated as the mean relative abundance of a given KOG class across all prokaryotic and eukaryotic genera and group evenness is calculated as the Shannon Evenness Index for total transcription of all eukaryotic phytoplankton genera. c, Mean seasonal trend in the relative abundance of KOG classes d, yearly trend in the relative abundance of all KOG classes.

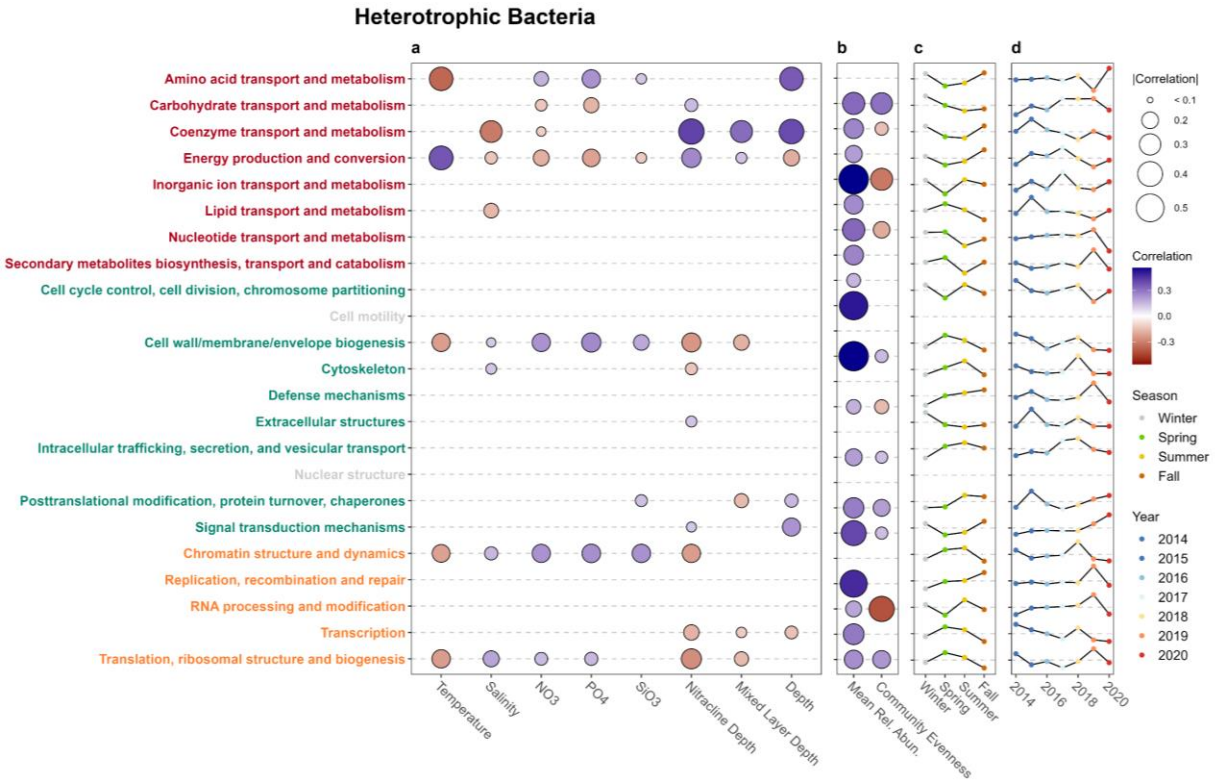


**Figure 4.6S**: Individual KOG Class summary for Heterotrophic Bacteria. a, Relationship between individual KOG classes and physical and chemical variables. Colors represent the direction and magnitude of correlation while the size of the circles represents the absolute magnitude of the correlation. No circle indicates the relationship is not significant. b, Relationship between the relative abundance of a KOG class ecological community parameters (mean relative abundance and group evenness). Where mean relative abundance is calculated as the mean relative abundance of a given KOG class across all prokaryotic and eukaryotic genera and group evenness is calculated as the Shannon Evenness Index for total transcription of all eukaryotic phytoplankton genera. c, Mean seasonal trend in the relative abundance of KOG classes d, yearly trend in the relative abundance of all KOG classes.

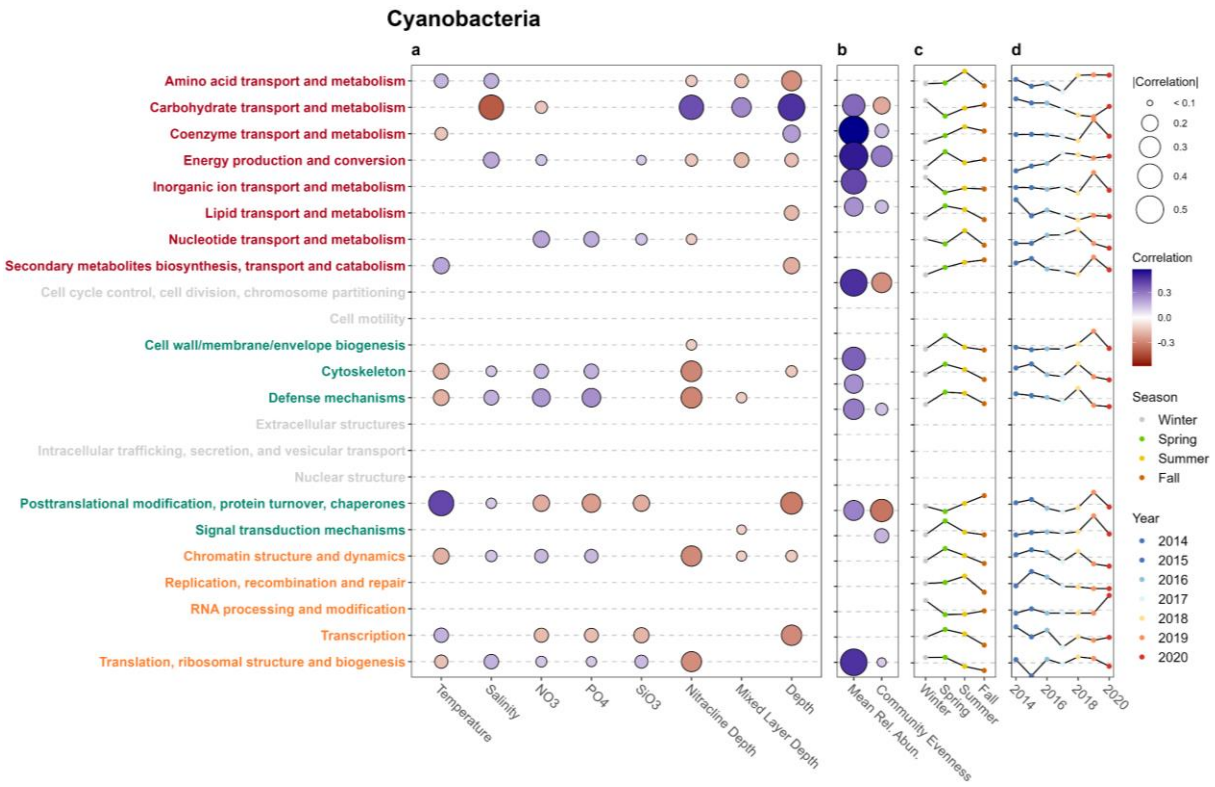


**Figure 4.7S**: Individual KOG Class summary for Cyanobacteria. a, Relationship between individual KOG classes and physical and chemical variables. Colors represent the direction and magnitude of correlation while the size of the circles represents the absolute magnitude of the correlation. No circle indicates the relationship is not significant. b, Relationship between the relative abundance of a KOG class ecological community parameters (mean relative abundance and group evenness). Where mean relative abundance is calculated as the mean relative abundance of a given KOG class across all prokaryotic and eukaryotic genera and group evenness is calculated as the Shannon Evenness Index for total transcription of all eukaryotic phytoplankton genera. c, Mean seasonal trend in the relative abundance of KOG classes d, yearly trend in the relative abundance of all KOG classes.

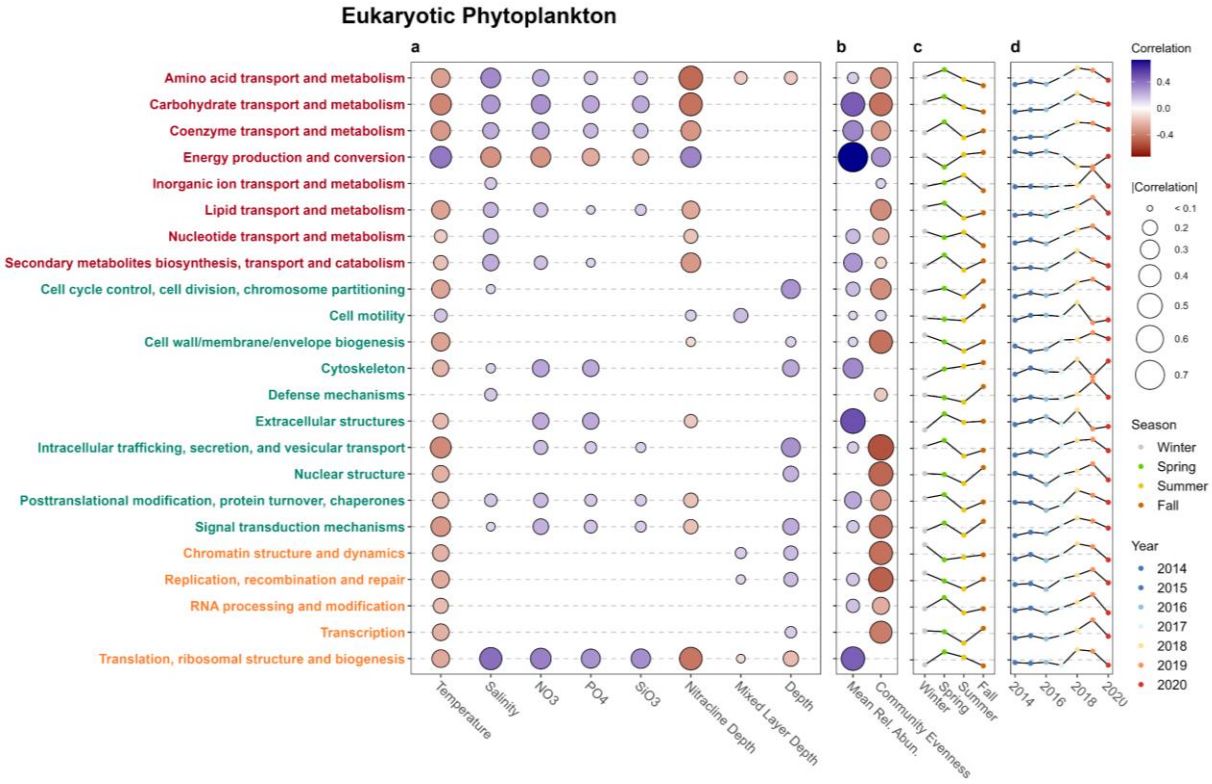


**Figure 4.8S**: Individual KOG Class summary for Eukaryotic Phytoplankton. a, Relationship between individual KOG classes and physical and chemical variables. Colors represent the direction and magnitude of correlation while the size of the circles represents the absolute magnitude of the correlation. No circle indicates the relationship is not significant. b, Relationship between the relative abundance of a KOG class ecological community parameters (mean relative abundance and group evenness). Where mean relative abundance is calculated as the mean relative abundance of a given KOG class across all prokaryotic and eukaryotic genera and group evenness is calculated as the Shannon Evenness Index for total transcription of all eukaryotic phytoplankton genera. c, Mean seasonal trend in the relative abundance of KOG classes d, yearly trend in the relative abundance of all KOG classes.

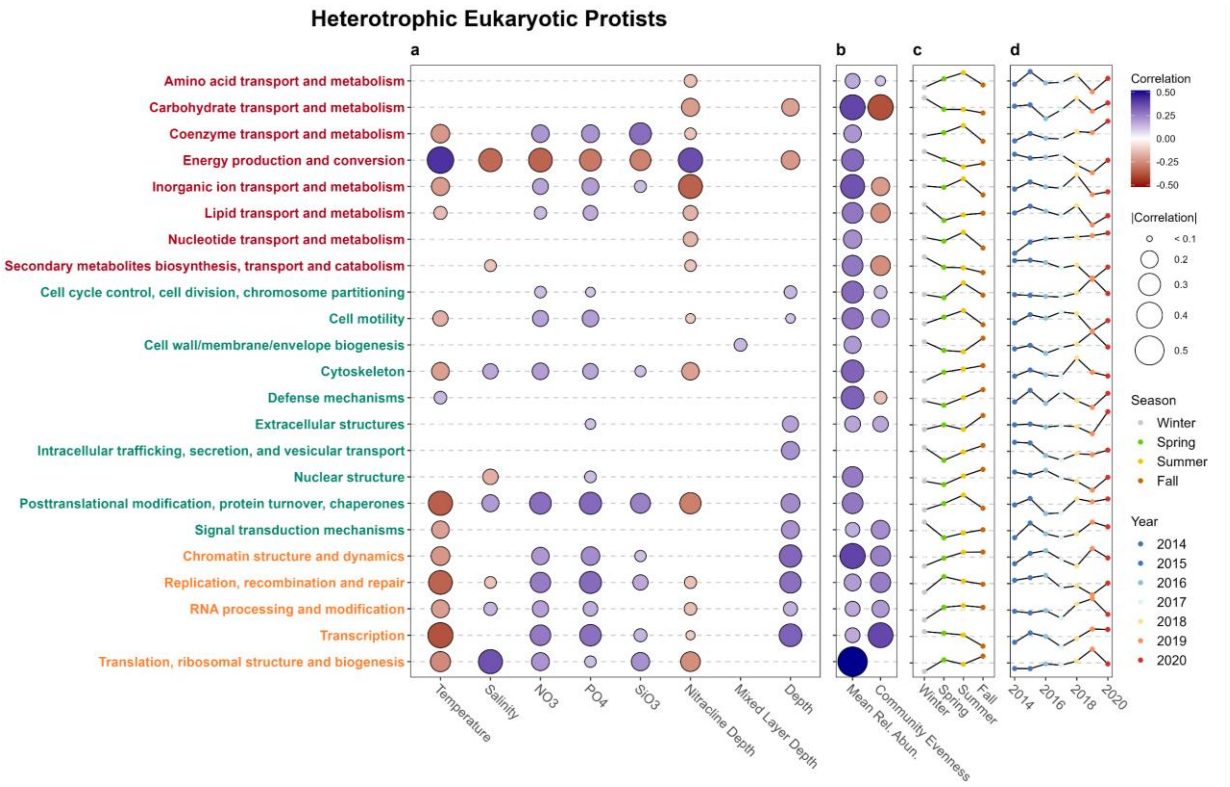


**Figure 4.9S**: Individual KOG Class summary for Heterotrophic Eukaryotic Protists. a, Relationship between individual KOG classes and physical and chemical variables. Colors represent the direction and magnitude of correlation while the size of the circles represents the absolute magnitude of the correlation. No circle indicates the relationship is not significant. b, Relationship between the relative abundance of a KOG class ecological community parameters (mean relative abundance and group evenness). Where mean relative abundance is calculated as the mean relative abundance of a given KOG class across all prokaryotic and eukaryotic genera and group evenness is calculated as the Shannon Evenness Index for total transcription of all eukaryotic phytoplankton genera. c, Mean seasonal trend in the relative abundance of KOG classes d, yearly trend in the relative abundance of all KOG classes.

**Table 4.1S**: KOG classes under the broader 'Metabolism' KOG group. Individual orfs that are the most correlated with their overall class and represent the largest proportion of transcripts.

Metabolism

| Broad Group | KOG Class | KOG Description | KEGG Description | PFAM Description | Major Taxonomic Group | Correlation | Maximum Proportion |
|---|---|---|---|---|---|---|---|
| All | Energy production and conversion | Aconitase/homoaconitase (aconitase superfamily) | Aconitase Aconitase_C AcnX | Aconitase family (aconitate hydratase)\|\|_GAP_\|\|Aconitase C-terminal domain | SAR406_X | 0.51 | 0.01 |
| All | Coenzyme transport and metabolism | 5-aminolevulinate synthase | Aminotran_1_2 Cys_Met_Meta_PP Aminotran_5 Preseq_ALAS Beta_elim_lyase DegT_DnrJ_EryC1 | Aminotransferase class I and II\|\|_GAP_ | Other Eukaryota | 0.21 | 0.11 |
| All | Carbohydrate transport and metabolism | 1, 2-alpha-mannosidase | Glyco_hydro_47 | Glycosyl hydrolase family 47 | Dinophyceae | 0.34 | 0.01 |
| All | Amino acid transport and metabolism | 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes | PAPS_reduct | Phosphoadenosine phosphosulfate reductase family\|\|_GAP_ | Pelagophyceae | 0.73 | 0.02 |
| All | Inorganic ion transport and metabolism | Acetyl-CoA transporter | Acatn | | Mamiellophyceae | 0.69 | 0.04 |
| All | Secondary metabolites biosynthesis, transport and catabolism | Alcohol dehydrogenase, class III | ADH_zinc_N ADH_N ADH_zinc_N_2 TrkA_N 2-Hacid_dh_C AlaDh_PNT_C UDPG_MGDP_dh_N | Zinc-binding dehydrogenase\|\|_GAP_ | Mamiellophyceae | 0.83 | 0.23 |
| All | Nucleotide transport and metabolism | 5' nucleotidase | 5_nucleotid_C Metallophos | _GAP_\|\|5'-nucleotidase, C-terminal domain\|\|_GAP_ | Pelagophyceae | 0.08 | 0.17 |
| Heterotrophic Bacteria | Energy production and conversion | Aconitase/homoaconitase (aconitase superfamily) | Aconitase Aconitase_C AcnX | Aconitase family (aconitate hydratase)\|\|_GAP_\|\|Aconitase C-terminal domain | SAR406_X | 0.13 | 0.01 |
| Heterotrophic Bacteria | Lipid transport and metabolism | 3-Methylcrotonyl-CoA carboxylase, biotin-containing subunit/Propionyl-CoA carboxylase, alpha chain/Acetyl-CoA carboxylase, biotin carboxylase subunit | CPSase_L_D2 Biotin_carb_N Biotin_carb_C Biotin_lipoyl ATP-grasp Biotin_lipoyl_2 Dala_Dala_lig_C HlyD_3 HlyD_D23 ATP-grasp_3 RnfC_N GCV_H | Carbamoyl-phosphate synthase L chain, ATP binding domain\|\|Biotin carboxylase C-terminal domain\|\|_GAP_ | SAR86 | 0.96 | 0.54 |
| Heterotrophic Bacteria | Carbohydrate transport and metabolism | 2-oxoglutarate dehydrogenase, E1 subunit | Transket_pyr E1_dh OxoGdeHyase_C 2-oxogl_dehyd_N | Dehydrogenase E1 component | SAR11 clade | 0.90 | 0.56 |
| Heterotrophic Bacteria | Amino acid transport and metabolism | 3-isopropylmalate dehydratase (aconitase superfamily) | Aconitase_C | Aconitase C-terminal domain\|\|_GAP_ | Other Bacteria | 0.08 | 0.06 |
| Heterotrophic Bacteria | Coenzyme transport and metabolism | Methylenetetrahydrofolate dehydrogenase/methylenetetrahydrofolate cyclohydrolase | THF_DHG_CYH_C THF_DHG_CYH AlaDh_PNT_C | Tetrahydrofolate dehydrogenase/cyclohydrolase, catalytic domain | SAR11 clade | 0.97 | 1.00 |
| Heterotrophic Bacteria | Secondary metabolites biosynthesis, transport and catabolism | Dehydrogenases with different specificities (related to short-chain alcohol dehydrogenases) | adh_short adh_short_C2 KR ApbA Shikimate_DH | short chain dehydrogenase\|\|_GAP_ | SAR116 | 0.59 | 1.00 |
| Heterotrophic Bacteria | Nucleotide transport and metabolism | Dihydroorotate dehydrogenase | DHO_dh Glu_synthase | Dihydroorotate dehydrogenase | SAR11 clade | 0.99 | 1.00 |
| Heterotrophic Bacteria | Inorganic ion transport and metabolism | Alkaline phosphatase | Alk_phosphatase Metalloenzyme | _GAP_\|\|Alkaline phosphatase | Alteromonadales | 0.65 | 1.00 |
| Eukaryotic Phytoplankton | Energy production and conversion | Aconitase/homoaconitase (aconitase superfamily) | Aconitase Aconitase_C AcnX | Aconitase family (aconitate hydratase) | Dinophyceae | 0.15 | 0.01 |
| Eukaryotic Phytoplankton | Coenzyme transport and metabolism | Bifunctional GTP cyclohydrolase II/3,4-dihydroxy-2butanone-4-phosphate synthase | DHBP_synthase GTP_cyclohydro2 ORC6 | GTP cyclohydrolase II\|\|_GAP_ | Mamiellophyceae | 0.81 | 0.18 |
| Eukaryotic Phytoplankton | Amino acid transport and metabolism | 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase (PAPS reductase)/FAD synthetase and related enzymes | PAPS_reduct | Phosphoadenosine phosphosulfate reductase family\|\|_GAP_ | Pelagophyceae | 0.63 | 0.02 |
| Eukaryotic Phytoplankton | Carbohydrate transport and metabolism | 1, 2-alpha-mannosidase | Glyco_hydro_47 | Glycosyl hydrolase family 47 | Dinophyceae | 0.36 | 0.02 |
| Eukaryotic Phytoplankton | Secondary metabolites biosynthesis, transport and catabolism | Alcohol dehydrogenase, class III | ADH_zinc_N ADH_N ADH_zinc_N_2 TrkA_N 2-Hacid_dh_C AlaDh_PNT_C UDPG_MGDP_dh_N | _GAP_\|\|Zinc-binding dehydrogenase\|\|_GAP_ | Mamiellophyceae | 0.80 | 0.33 |
| Eukaryotic Phytoplankton | Inorganic ion transport and metabolism | Acetyl-CoA transporter | Acatn | | Mamiellophyceae | 0.64 | 0.11 |
| Eukaryotic Phytoplankton | Nucleotide transport and metabolism | 5' nucleotidase | 5_nucleotid_C Metallophos | _GAP_\|\|5'-nucleotidase, C-terminal domain\|\|_GAP_ | Pelagophyceae | 0.05 | 0.39 |
| Eukaryotic Phytoplankton | Lipid transport and metabolism | 17 beta-hydroxysteroid dehydrogenase type 3, HSD17B3 | adh_short adh_short_C2 Polysacc_synt_2 KR Epimerase | short chain dehydrogenase\|\|_GAP_ | Mamiellophyceae | 0.80 | 0.07 |
| Heterotrophic Eukaryotic Protists | Energy production and conversion | Acyl carrier protein/NADH-ubiquinone oxidoreductase, NDUFAB1/SDAP subunit | PP-binding | _GAP_\|\|Phosphopantetheine attachment site | Other Eukaryota | 0.50 | 0.03 |
| Heterotrophic Eukaryotic Protists | Carbohydrate transport and metabolism | 1,4-alpha-glucan branching enzyme/starch branching enzyme II | CBM_48 Alpha-amylase DDAH_eukar | | Other Eukaryota | 0.18 | 0.05 |
| Heterotrophic Eukaryotic Protists | Inorganic ion transport and metabolism | Acetyl-CoA transporter | Acatn MFS_1 YwcE DUF2614 | | Other Eukaryota | 0.39 | 0.08 |
| Heterotrophic Eukaryotic Protists | Secondary metabolites biosynthesis, transport and catabolism | Alcohol dehydrogenase, class III | Esterase ADH_zinc_N ADH_N Peptidase_S9 Hydrolase_4 Lipase_3 | | Ciliophora | 0.51 | 0.42 |
| Heterotrophic Eukaryotic Protists | Amino acid transport and metabolism | 3-Methylcrotonyl-CoA carboxylase, non-biotin containing subunit/Acetyl-CoA carboxylase carboxyl transferase, subunit beta | Carboxyl_trans ECH_1 DUF5623 | Carboxyl transferase domain | Syndiniales | 0.24 | 0.53 |
| Heterotrophic Eukaryotic Protists | Nucleotide transport and metabolism | Adenine deaminase/adenosine deaminase | A_deaminase A_deaminase_N RIG-I_C | Adenosine/AMP deaminase | Ciliophora | 0.46 | 0.96 |
| Heterotrophic Eukaryotic Protists | Coenzyme transport and metabolism | 5-aminolevulinate synthase | Aminotran_1_2 Cys_Met_Meta_PP Aminotran_5 Preseq_ALAS Beta_elim_lyase DegT_DnrJ_EryC1 | Aminotransferase class I and II\|\|_GAP_ | Other Eukaryota | 0.17 | 1.00 |
| Archaea | Amino acid transport and metabolism | Puromycin-sensitive aminopeptidase and related aminopeptidases | Peptidase_M1 ERAP1_C Peptidase_M1_N Aminopep DUF2808 UB2H | | Thaumarchaeota_X | 1.00 | 1.00 |
| Archaea | Energy production and conversion | Cytochrome b | Cytochrome_B Cytochrom_B_N_2 DUF4405 DUF373 | | Halobacteria | 0.47 | 1.00 |
| Cyanobacteria | Energy production and conversion | ATP synthase F0 subunit 6 and related proteins | ATP-synt_A | ATP synthase A chain | Cyanobacteria_X | 0.58 | 0.83 |
| Cyanobacteria | Lipid transport and metabolism | SAM-dependent methyltransferases | Methyltransf_11 MetW Methyltransf_12 Methyltransf_23 Methyltransf_25 | | Cyanobacteria_X | 0.78 | 1.00 |
| Cyanobacteria | Carbohydrate transport and metabolism | D-ribulose-5-phosphate 3-epimerase | RibuL_P_3_epim His_biosynth DHO_dh | Ribulose-phosphate 3 epimerase family | Cyanobacteria_X | 0.27 | 1.00 |
| Cyanobacteria | Amino acid transport and metabolism | Tryptophan synthase beta chain | PALP | Pyridoxal-phosphate dependent enzyme | Cyanobacteria_X | 0.39 | 1.00 |
| Cyanobacteria | Inorganic ion transport and metabolism | Fe-S cluster biosynthesis protein ISA1 (contains a HesB-like domain) | Fe-S_biosyn | Iron-sulphur cluster biosynthesis | Cyanobacteria_X | 0.93 | 1.00 |
| Cyanobacteria | Coenzyme transport and metabolism | Delta-aminolevulinic acid dehydratase | ALAD | Delta-aminolevulinic acid dehydratase | Cyanobacteria_X | 1.00 | 1.00 |
| Cyanobacteria | Secondary metabolites biosynthesis, transport and catabolism | Amine oxidase | Amino_oxidase NAD_binding_8 MCRA Pyr_redox_2 DAO Pyr_redox FAD_binding_2 FAD_oxidored HI0933_like NAD_binding_9 Thi4 FAD_binding_3 GIDA | Flavin containing amine oxidoreductase | Cyanobacteria_X | 1.00 | 1.00 |
| Cyanobacteria | Nucleotide transport and metabolism | Nucleoside diphosphate kinase | NDK PSI_PsaJ | Nucleoside diphosphate kinase | Cyanobacteria_X | 1.00 | 1.00 |

**Table 4.2S**: KOG classes under the broader 'Cellular Processing and Signaling' KOG group. Individual orfs that are the most correlated with their overall class and represent the largest proportion of transcripts.

Cellular Processing and Signaling

| Broad Group | KOG Class | KOG Description | KEGG Description | PFAM Description | Major Taxonomic Group | Correlation | Maximum Proportion |
|---|---|---|---|---|---|---|---|
| All | Cytoskeleton | Actin-binding protein Coronin, contains WD40 repeats | DUF1899 WD40_4 WD40 ANAPC4_WD40 | Domain of unknown function (DUF1900) | Bacillariophyta | 0.49 | 0.09 |
| All | Posttranslational modification, protein turnover, chaperones | 20S proteasome, regulatory subunit alpha type PSMA1/PRE5 | Proteasome Proteasome_A_N Phytase-like | Proteasome subunit | Mamiellophyceae | 0.79 | 0.10 |
| All | Intracellular trafficking, secretion, and vesicular transport | 40 kDa farnesylated protein associated with peroxisomes | Pex19 TM231 SRP54_N | Pex19 protein family | Other Eukaryota | 0.54 | 0.02 |
| All | Signal transduction mechanisms | 1-aminocyclopropane-1-carboxylate synthase, and related proteins | Aminotran_1_2 | Aminotransferase class I and II | Cryptophyta | 0.56 | 0.03 |
| All | Cell cycle control, cell division, chromosome partitioning | Anaphase-promoting complex (APC), Cdc23 subunit | TPR_1 TPR_2 ANAPC8 TPR_14 TPR_8 TPR_19 TPR_11 TPR_16 TPR_12 TPR_7 TPR_17 ANAPC3 TPR_10 TPR_15 Coatomer_E ChAPs PPTA Cohesin_load Fis1_TPR_C | PF13414\|\|_GAP_ | Mamiellophyceae | 0.59 | 0.19 |
| All | Cell wall/membrane/envelope biogenesis | 1,3-beta-glucan synthase/callose synthase catalytic subunit | Glucan_synthase FKS1_dom1 | 1,3-beta-glucan synthase component | Mamiellophyceae | 0.45 | 0.56 |
| All | Cell motility | Myosin VII, myosin IXB and related myosins | Myosin_head MyTH4 FERM_M SH3_2 SH3_1 SH3_9 FERM_N RA Alkyl_sulf_C eIF3_subunit PID | Myosin head (motor domain) | Dinophyceae | 0.93 | 1.00 |
| All | Defense mechanisms | Bax-mediated apoptosis inhibitor TEGT/BI-1 | Bax1-I | Uncharacterised protein family UPF0005 | Ciliophora | 0.59 | 0.86 |
| All | Extracellular structures | Cell adhesion complex protein bystin | Bystin | Bystin | Bacillariophyta | 0.34 | 1.00 |
| All | Nuclear structure | Karyopherin (importin) beta 1 | HEAT HEAT_EZ HEAT_2 IBN_N Vac14_Fab1_bd MMS19_C DUF577 YukD | | Mamiellophyceae | 0.83 | 1.00 |
| Heterotrophic Bacteria | Posttranslational modification, protein turnover, chaperones | 26S proteasome regulatory complex, ATPase RPT6 | Peptidase_M41 AAA AAA_lid_3 FtsH_ext AAA_5 TIP49 AAA_16 RuvB_N AAA_22 Zeta_toxin AAA_33 AAA_17 Sigma54_activat AAA_2 NACHT | FtsH Extracellular\|\|_GAP_\|\|ATPase family associated with various cellular activities (AAA) | SAR11 clade | 0.45 | 0.53 |
| Heterotrophic Bacteria | Defense mechanisms | Flavonol reductase/cinnamoyl-CoA reductase | Epimerase 3Beta_HSD NAD_binding_10 GDP_Man_Dehyd NAD_binding_4 adh_short NmrA | | SAR11 clade | 0.97 | 1.00 |
| Heterotrophic Bacteria | Intracellular trafficking, secretion, and vesicular transport | Mitochondrial Fe/S cluster exporter, ABC superfamily | ABC_membrane ABC_tran SMC_N AAA AAA_5 AAA_16 AAA_22 RsgA_GTPase MMR_HSR1 AAA_29 AAA_18 DEAD AAA_24 AAA_21 Sigma54_activat AAA_7 Zeta_toxin AAA_23 | ABC transporter transmembrane region | SAR11 clade | 0.99 | 1.00 |
| Heterotrophic Bacteria | Cell wall/membrane/envelope biogenesis | UDP-glucose 4-epimerase/UDP-sulfoquinovose synthase | Epimerase GDP_Man_Dehyd 3Beta_HSD RmlD_sub_bind Polysacc_synt_2 NAD_binding_10 NAD_binding_4 NmrA KR adh_short TrkA_N DAO | NAD dependent epimerase/dehydratase family | Other Bacteria | 0.99 | 1.00 |
| Eukaryotic Phytoplankton | Cytoskeleton | Actin-binding protein Coronin, contains WD40 repeats | DUF1899 WD40_4 WD40 ANAPC4_WD40 | Domain of unknown function (DUF1900) | Bacillariophyta | 0.30 | 0.13 |
| Eukaryotic Phytoplankton | Posttranslational modification, protein turnover, chaperones | 20S proteasome, regulatory subunit alpha type PSMA1/PRE5 | Proteasome Proteasome_A_N Phytase-like | Proteasome subunit | Mamiellophyceae | 0.79 | 0.03 |
| Eukaryotic Phytoplankton | Intracellular trafficking, secretion, and vesicular transport | 40 kDa farnesylated protein associated with peroxisomes | Pex19 | Pex19 protein family | Cryptophyta | 0.75 | 0.02 |
| Eukaryotic Phytoplankton | Signal transduction mechanisms | 1-aminocyclopropane-1-carboxylate synthase, and related proteins | Aminotran_1_2 | Aminotransferase class I and II | Cryptophyta | 0.53 | 0.09 |
| Eukaryotic Phytoplankton | Cell cycle control, cell division, chromosome partitioning | Anaphase-promoting complex (APC), Cdc23 subunit | TPR_1 TPR_2 ANAPC8 TPR_14 TPR_8 TPR_19 TPR_11 TPR_16 TPR_12 TPR_7 TPR_17 ANAPC3 TPR_10 TPR_15 Coatomer_E ChAPs PPTA Cohesin_load Fis1_TPR_C | PF13414\|\|_GAP_ | Mamiellophyceae | 0.57 | 0.51 |
| Eukaryotic Phytoplankton | Cell wall/membrane/envelope biogenesis | 1,3-beta-glucan synthase/callose synthase catalytic subunit | Glucan_synthase FKS1_dom1 | 1,3-beta-glucan synthase component | Mamiellophyceae | 0.43 | 0.80 |
| Eukaryotic Phytoplankton | Cell motility | Myosin VII, myosin IXB and related myosins | Myosin_head MyTH4 FERM_M SH3_2 SH3_1 SH3_9 FERM_N RA Alkyl_sulf_C eIF3_subunit PID | _GAP_\|\|FERM central domain\|\|_GAP_ | Dinophyceae | 0.86 | 1.00 |
| Eukaryotic Phytoplankton | Defense mechanisms | Bax-mediated apoptosis inhibitor TEGT/BI-1 | Bax1-I | Uncharacterised protein family UPF0005 | Dinophyceae | 0.24 | 0.90 |
| Eukaryotic Phytoplankton | Nuclear structure | Karyopherin (importin) beta 1 | HEAT HEAT_EZ HEAT_2 IBN_N Vac14_Fab1_bd MMS19_C DUF577 YukD | | Mamiellophyceae | 0.88 | 1.00 |
| Eukaryotic Phytoplankton | Extracellular structures | Cell adhesion complex protein bystin | Bystin | Bystin | Bacillariophyta | 0.59 | 1.00 |
| Heterotrophic Eukaryotic Protists | Posttranslational modification, protein turnover, chaperones | 20S proteasome, regulatory subunit alpha type PSMA1/PRE5 | Proteasome Proteasome_A_N | Proteasome subunit | Ciliophora | 0.38 | 0.28 |
| Heterotrophic Eukaryotic Protists | Intracellular trafficking, secretion, and vesicular transport | Adaptor complexes medium subunit family | Adap_comp_sub Clat_adaptor_s | Adaptor complexes medium subunit family | Ciliophora | 0.64 | 0.54 |
| Heterotrophic Eukaryotic Protists | Signal transduction mechanisms | Acetylcholine receptor | Neur_chan_LBD Neur_chan_memb | Neurotransmitter-gated ion-channel ligand binding domain | Other Eukaryota | 0.25 | 0.07 |
| Heterotrophic Eukaryotic Protists | Cytoskeleton | Actin-binding protein Coronin, contains WD40 repeats | DUF1899 WD40_4 WD40 ANAPC4_WD40 Coatomer_WDAD WDCP | Domain of unknown function (DUF1900)\|\|_GAP_ | Lobosa | 0.55 | 0.17 |
| Heterotrophic Eukaryotic Protists | Cell wall/membrane/envelope biogenesis | Ankyrin | Ank_2 Ank_4 Ank_3 Ank Ank_5 TPR_1 | PF13637\|\|PF12796 | Other Eukaryota | 0.09 | 0.44 |
| Heterotrophic Eukaryotic Protists | Cell cycle control, cell division, chromosome partitioning | Anaphase promoting complex, Cdc20, Cdh1, and Ama1 subunits | WD40 ANAPC4_WD40 eIF2A Ge1_WD40 ELYS-bb | WD domain, G-beta repeat\|\|_GAP_ | Other Eukaryota | 0.19 | 1.00 |
| Heterotrophic Eukaryotic Protists | Defense mechanisms | Bax-mediated apoptosis inhibitor TEGT/BI-1 | Bax1-I | Uncharacterised protein family UPF0005 | Ciliophora | 0.99 | 1.00 |
| Heterotrophic Eukaryotic Protists | Extracellular structures | Extracellular matrix glycoprotein Laminin subunit beta | Laminin_EGF Laminin_N F5_F8_type_C | Laminin EGF-like (Domains III and V) | Other Eukaryota | 0.18 | 1.00 |
| Heterotrophic Eukaryotic Protists | Nuclear structure | Karyopherin (importin) beta 1 | HEAT HEAT_EZ IBN_N HEAT_2 Vac14_Fab1_bd DUF577 | _GAP_\|\|PF13513\|\|_GAP_ | Other Eukaryota | 0.37 | 1.00 |
| Heterotrophic Eukaryotic Protists | Cell motility | Myosin VII, myosin IXB and related myosins | Myosin_head MyTH4 FERM_M IQ SH3_9 SH3_2 FERM_N SH3_1 AAA_22 ABC_tran AAA_16 PID | _GAP_\|\|MyTH4 domain\|\|_GAP_ | Other Eukaryota | 1.00 | 1.00 |
| Archaea | Intracellular trafficking, secretion, and vesicular transport | Transport protein Sec61, alpha subunit | SecY Plug_translocon Phage_holin_5_1 | _GAP_\|\|Plug domain of Sec61p\|\|eubacterial secY protein | Thermoplasmata | 0.88 | 1.00 |
| Archaea | Posttranslational modification, protein turnover, chaperones | 20S proteasome, regulatory subunit beta type PSMB5/PSMB8/PRE2 | Proteasome | Proteasome subunit | Thaumarchaeota_X | 0.05 | 1.00 |
| Cyanobacteria | Posttranslational modification, protein turnover, chaperones | 26S proteasome regulatory complex, ATPase RPT2 | AAA Peptidase_M41 AAA_lid_3 AAA_5 AAA_2 TIP49 AAA_16 RuvB_N AAA_14 AAA_22 FtsH_ext Mg_chelatase AAA_33 IstB_IS21 AAA_28 Sigma54_activ_2 AAA_17 Sigma54_activat AAA_7 AAA_18 AAA_24 ABC_tran TsaE | _GAP_\|\|ATPase family associated with various cellular activities (AAA)\|\|_GAP_ | Cyanobacteria_X | 0.66 | 1.00 |
| Cyanobacteria | Cell wall/membrane/envelope biogenesis | GDP-mannose pyrophosphorylase/mannose-1-phosphate guanylyltransferase | NTP_transferase NTP_transf_3 Fucokinase | | Cyanobacteria_X | 0.94 | 1.00 |
| Cyanobacteria | Defense mechanisms | Serpin | Serpin | _GAP_\|\|Serpin (serine protease inhibitor) | Cyanobacteria_X | 1.00 | 1.00 |

**Table 4.3S**: KOG classes under the broader 'Information Storage and Processing' KOG group. Individual orfs that are the most correlated with their overall class and represent the largest proportion of transcripts.

Information Storage and Processing

| Broad Group | KOG Class | KOG Description | KEGG Description | PFAM Description | Major Taxonomic Group | Correlation | Maximum Proportion |
|---|---|---|---|---|---|---|---|
| All | Transcription | Activating signal cointegrator 1 | | | Mamiellophyceae | 0.65 | 0.04 |
| All | Translation, ribosomal structure and biogenesis | 40s ribosomal protein s10 | S10_plectin | Plectin/S10 domain | Ciliophora | 0.08 | 0.88 |
| All | RNA processing and modification | 60S ribosomal protein 15.5kD/SNU13, NHP2/L7A family (includes ribonuclease P subunit p38), involved in splicing | Ribosomal_L7Ae PELOTA_1 FlaC_arch | Ribosomal protein L7Ae/L30e/S12e/Gadd45 family | Mamiellophyceae | 0.50 | 0.17 |
| All | Chromatin structure and dynamics | Beta-transducin family (WD-40 repeat) protein | WD40 ANAPC4_WD40 eIF2A LisH IKI3 Coatomer_WDAD Nbas_N | WD domain, G-beta repeat\|\|_GAP_ | Pelagophyceae | 0.74 | 0.94 |
| All | Replication, recombination and repair | 3'-5' exonuclease | RNase_T | _GAP_\|\|Exonuclease\|\|_GAP_ | Bacillariophyta | 0.56 | 0.37 |
| Heterotrophic Bacteria | Transcription | RNA polymerase II, second largest subunit | RNA_pol_Rpb2_6 RNA_pol_Rpb2_3 RNA_pol_Rpb2_7 RNA_pol_Rpb2_2 RNA_pol_Rpb2_1 RNA_pol_Rpb2_45 | RNA polymerase Rpb2, domain 6 | SAR11 clade | 1.00 | 1.00 |
| Heterotrophic Bacteria | Translation, ribosomal structure and biogenesis | 40S ribosomal protein S15/S22 | Ribosomal_S8 | Ribosomal protein S8 | Flavobacteria | 0.86 | 0.27 |
| Heterotrophic Bacteria | Chromatin structure and dynamics | Transducin-like enhancer of split protein (contains WD40 repeats) | | _GAP_\|\|Groucho/TLE N-terminal Q-rich domain | Other Bacteria | 1.00 | 1.00 |
| Heterotrophic Bacteria | Replication, recombination and repair | 3-methyladenine DNA glycosidase | HhH-GPD | HhH-GPD superfamily base excision DNA repair protein | SAR11 clade | 1.00 | 1.00 |
| Heterotrophic Bacteria | RNA processing and modification | ATP-dependent RNA helicase | DEAD Helicase_C ResIII AAA_19 CMS1 UTP25 Fibrinogen_BP | DEAD/DEAH box helicase\|\|_GAP_\|\|Helicase conserved C-terminal domain\|\|_GAP_ | SAR11 clade | 1.00 | 1.00 |
| Eukaryotic Phytoplankton | RNA processing and modification | 60S ribosomal protein 15.5kD/SNU13, NHP2/L7A family (includes ribonuclease P subunit p38), involved in splicing | Ribosomal_L7Ae PELOTA_1 FlaC_arch | Ribosomal protein L7Ae/L30e/S12e/Gadd45 family | Mamiellophyceae | 0.50 | 0.27 |
| Eukaryotic Phytoplankton | Translation, ribosomal structure and biogenesis | 40s ribosomal protein s10 | S10_plectin | Plectin/S10 domain\|\|_GAP_ | Bacillariophyta | 0.68 | 0.03 |
| Eukaryotic Phytoplankton | Chromatin structure and dynamics | Beta-transducin family (WD-40 repeat) protein | WD40 ANAPC4_WD40 eIF2A LisH IKI3 Coatomer_WDAD Nbas_N | WD domain, G-beta repeat\|\|_GAP_ | Pelagophyceae | 0.73 | 0.95 |
| Eukaryotic Phytoplankton | Transcription | Activating signal cointegrator 1 | | | Mamiellophyceae | 0.69 | 0.11 |
| Eukaryotic Phytoplankton | Replication, recombination and repair | 3'-5' exonuclease | RNase_T | _GAP_\|\|Exonuclease\|\|_GAP_ | Bacillariophyta | 0.54 | 0.78 |
| Heterotrophic Eukaryotic Protists | Translation, ribosomal structure and biogenesis | 40s ribosomal protein s10 | S10_plectin | Plectin/S10 domain | Ciliophora | 0.28 | 0.96 |
| Heterotrophic Eukaryotic Protists | RNA processing and modification | 60S ribosomal protein 15.5kD/SNU13, NHP2/L7A family (includes ribonuclease P subunit p38), involved in splicing | Ribosomal_L7Ae PELOTA_1 Adenosine_kin | Ribosomal protein L7Ae/L30e/S12e/Gadd45 family | Other Eukaryota | 0.23 | 0.08 |
| Heterotrophic Eukaryotic Protists | Transcription | Activating signal cointegrator 1 | ASCH zf-C2HC5 | _GAP_\|\|Putative zinc finger motif, C2HC5-type\|\|_GAP_ | Other Eukaryota | 0.17 | 0.06 |
| Heterotrophic Eukaryotic Protists | Chromatin structure and dynamics | Beta-transducin family (WD-40 repeat) protein | WD40 ANAPC4_WD40 LisH eIF2A Nup160 Coatomer_WDAD RAB3GAP2_N WD40_like | _GAP_\|\|WD domain, G-beta repeat\|\|_GAP_ | Other Eukaryota | 0.40 | 0.28 |
| Heterotrophic Eukaryotic Protists | Replication, recombination and repair | 3'-5' exonuclease | RNase_T DNA_pol_A_exo1 | _GAP_\|\|Exonuclease\|\|_GAP_ | Other Eukaryota | 0.52 | 0.66 |
| Archaea | RNA processing and modification | RNAse L inhibitor, ABC superfamily | ABC_tran AAA_21 RLI SMC_N Fer4 AAA_23 Rad17 AAA AAA_16 Fer4_6 AAA_29 Fer4_10 Fer4_21 Fer4_7 Fer4_9 Fer4_16 Fer4_2 DO-GTPase2 Fer4_4 Fer4_8 | Possible metal-binding domain in RNase L inhibitor, RLI\|\|4Fe-4S binding domain\|\|ABC transporter\|\|_GAP_ | Other Archaea | 0.97 | 1.00 |
| Archaea | Translation, ribosomal structure and biogenesis | 40S ribosomal protein S11 | Ribosomal_S17 Ribosomal_S17_N CxC5 zf-C3H1 | _GAP_\|\|Ribosomal protein S17 | Other Archaea | 0.74 | 0.97 |
| Archaea | Transcription | RNA polymerase II, second largest subunit | RNA_pol_Rpb2_6 RNA_pol_Rpb2_1 RNA_pol_Rpb2_2 RNA_pol_Rpb2_3 RNA_pol_Rpb2_7 RNA_pol_Rpb2_4 RNA_pol_Rpb2_5 Terminase_GpA HMBD zf-RING_7 A2L_zn_ribbon gpD TF_Zn_Ribbon | RNA polymerase Rpb2, domain 2\|\|RNA polymerase beta subunit\|\|RNA polymerase Rpb2, domain 3\|\|_GAP_\|\|RNA polymerase Rpb2, domain 4 | Thermoplasmata | 0.97 | 1.00 |
| Cyanobacteria | Translation, ribosomal structure and biogenesis | 60S ribosomal protein L14/L17/L23 | Ribosomal_L14 | Ribosomal protein L14p/L23e | Cyanobacteria_X | 0.17 | 1.00 |
| Cyanobacteria | Transcription | RNA polymerase II, second largest subunit | RNA_pol_Rpb2_6 RNA_pol_Rpb2_2 RNA_pol_Rpb2_3 RNA_pol_Rpb2_7 RNA_pol_Rpb2_1 RNA_pol_Rpb2_45 RnfC_N | RNA polymerase Rpb2, domain 6 | Cyanobacteria_X | 1.00 | 1.00 |

## 4.7 Acknowledgements

CCJ acknowledges graduate student support by Scripps Institution of Oceanography.

This study was supported by National Science Foundation, California Current Ecosystem Long Term Ecological Research Grants, CCE-LTER Phase II and III (NSF-OCE-1026607 and NSF-

OCE-1637632), and NSF-OCE-1756884, NOAA (NOAA OAR Omics, CIMEC NA15OAR4320071, and ECOHAB NA19NOS4780181), and Gordon and Betty Moore Foundation grants GBMF3828 to AEA.

We would like to acknowledge former CalCOFI director David M. Checkley and Margot Bohan from the NOAA Office of Ocean Exploration and Research (OER) for their vision and guidance during the initial phase of the NCOG program and current CalCOFI director Brice X. Semmens for his continued support. We are also especially grateful to California Current Ecosystem, Long Term Ecological Research (CCE-LTER) and CalCOFI project and team members and crew who have assisted with the NCOG program 2014-present.

Chapter 4, in part is currently being prepared for submission for publication of the material. James, C. C., Barton, A. D., Allen, L. Z., Smith, S. R., Venepally, P., Lampe, R. H., Rabines, A., Schulberg, A., Zheng, H., Goericke, R., Goodwin, K. D., Allen, A. E. The dissertation author was the primary investigator and author of this paper.

# Chapter 5 - Conclusion

The purpose of this thesis is to uncover the processes that shape regional patterns of marine microbial structure, diversity, and function at the high-resolution provided by meta-omics sampling. Combined, the results presented here confirm many of the patterns and processes observed within well-studied groups, such as diatoms or cyanobacteria, while providing a breadth of new information with regards to cryptic groups that could not be identified through traditional means. Within the Southern California Current (SCC) region, cross-shore gradients in nutrient supply to the surface ocean appear to be one of the strongest structuring forces of the marine microbiome, affecting the spatial and temporal structure of microbial community composition and function.

Chapter 2 presented an examination of how combined spatial and temporal meta-omics sampling can reveal system wide relationships between regional environmental conditions and ecological structure and diversity. Through small subunit ribosomal RNA gene sequencing on the V4-V5 region of the 16S rRNA gene and V9 region of the 18S rRNA gene, known generally as metabarcoding, data described in Chapter 2 captured the majority of prokaryotic and eukaryotic protist diversity within the region. The community structure and diversity of both broad and narrow microbial taxonomic groups within the region was largely driven by the supply of nutrients to the surface ocean, captured by nitracline depth measurements. Nitracline depth, defined as the depth at which the concentration of nitrate exceeds 1 μM, is the result of both abiotic (upwelling) and biotic (biological drawdown) factors and in many cases was a better predictor of community structure and diversity than actual concentrations of nutrients or globally important variables like temperature (Sunagawa et al. 2015). Overall regional nutrient supply, as the result of the relative magnitude and duration of spring upwelling conditions, led to shifts in

the relative dominance of nearshore versus offshore community types seasonally and interannually. While these patterns have been previously observed via bulk measures such as chlorophyll-α or within select groups (Kahru and Mitchell 2001; Venrick 2009; Taylor et al. 2015), this study found that these gradients drive large changes in community structure within taxonomic groups as well, leading to divergent communities even within taxonomically related species.

The relative effects of selection and dispersal on microbial community composition has been a primary subject of microbial ecology for nearly a century (Becking 1934; Fenchel and Finlay 2004; Martiny et al. 2006; Gibbons et al. 2013; Ward et al. 2021). Within terrestrial systems, endemic soil microbes, found only in particular locations or regions, represent a majority of the observed taxonomic diversity (Talbot et al. 2014). Within the marine environment dispersal via currents is thought to lead to more cosmopolitan distributions. However, many previous studies that have measured rates of endemism and cosmopolitanism of marine microbes have done so without considering the dynamic nature of marine habitats (Malviya et al. 2016; Gimmler et al. 2016; Canals et al. 2020). Water masses, with conserved physical and chemical properties serve as the foundation for marine microbial habitats and shift across both space and time (D'Ovidio et al. 2010). Following the identification of regional water masses within the SCC, Chapter 3 aimed to identify the role of selection on marine microbes and asked the following questions 1) are marine microbes found in preferred habitats and 2) where do most marine microbial distributions fall along a spectrum from endemism to cosmopolitanism. Utilizing a subset of 445 metabarcoding surface samples within the SCC, ~60% of microbial taxa fell somewhere in between endemism and cosmopolitanism, occurring in some but not all available habitats within the SCC. In general, these species were rare and had no

affinity towards a particular habitat. Approximately 10% of prokaryotic and eukaryotic taxa were endemic to individual habitats within the region, far lower than rates of microbial endemism in terrestrial systems. Surprisingly, of the ~30% of species that were cosmopolitan, occurring in every regional habitat, the majority were over abundant in only one habitat—highlighting the importance of both selection, and spatial and temporal mass effects (dispersal), which drove regional cosmopolitan distributions (Shmida and Wilson 1985). Thus, in general it appears there are two major modes of marine microbial distributions: 1) rare, evenly distributed taxa that make up the majority of microbial diversity and 2) marine microbes that have habitat preferences and vary in abundance from rare, endemic taxa, to abundant, cosmopolitan taxa that become regionally dispersed into all habitats likely as a result of ecological mass effects.

Chapter 4 explored the metatranscriptome of the marine microbial community in the SCC and identified the abiotic and biotic processes that shape microbial activity and function across the region. In this chapter the following questions were addressed: 1) which environmental gradients lead to the greatest niche partitioning amongst active microbial members within the region and 2) how does the functional composition of the microbial community change as a result of environmental conditions and community structure? Across all taxonomic groups, both nitracline depth and temperature showed intermediate levels of variability in niche optimums coupled with strong habitat selectivity (abundances decline rapidly away from niche optimums), indicative of niche partitioning both across and within taxonomic groups. Across all taxa, salinity gradients aligned with shifts in community-wide functional composition. This pattern was largely driven by the relative abundance of genes associated with energy production and conversion and occurred across all taxa relatively evenly. In general, most functional shifts were not the result of individual taxa dominating the ecosystem but rather community-wide functional

responses to changing environmental conditions. One exception was within eukaryotic phytoplankton where most functional shifts aligned with decreases in community evenness. In general, this chapter presents a framework for tackling ecological questions, both longstanding (niche partitioning) and new (functional partitioning) within the highly complex landscape of metatranscriptome data.

Across the three studies presented in this thesis, an overarching goal was to effectively utilize the astonishing breath of ecological data generated from NCOG to tackle ecological questions about the marine microbiome. The UN has declared this decade (2021-2030) as the Decade of Ocean Science for Sustainable Development with programs like the Ocean Biomolecular Observing Network (OBON) aiming to use environmental sequencing as the basis for improving our knowledge about ocean ecosystems (Chavez et al. 2021). Meta-omic data while comprehensive, can also be complex and difficult to synthesize. The NCOG metatranscriptome data used in Chapter 4 represents nearly 2,000 prokaryotic and eukaryotic taxa and over 100,000 unique genes. While this increase in available ecological data represents an extraordinary leap forward in our ability to uncover ecological patterns and processes it also drastically changes how ecologists must conceive of hypotheses, analyze results, synthesize, and interpret observed patterns and processes. Far from answering the majority of these ecological questions, this thesis only just begins to address how, with meta-omic data we can answer longstanding questions in ecology and start to ask new questions at a scale appropriate to the resolution provided by environmental genomic sampling.

# References


Abrantes, Fatima, Pedro Cermeno, Cristina Lopes, Oscar Romero, Lélia Matos, Jolanda Van Iperen, Marta Rufino, and Vitor Magalhães. 2016. “Diatoms Si Uptake Capacity Drives Carbon Export in Coastal Upwelling Systems.” Biogeosciences 13: 4099–4109. https://doi.org/10.5194/bg-13-4099-2016.

Alexander, Harriet, Bethany D. Jenkins, Tatiana A. Rynearson, and Sonya T. Dyhrman. 2015. “Metatranscriptome Analyses Indicate Resource Partitioning between Diatoms in the Field.” Proceedings of the National Academy of Sciences of the United States of America 112 (17): E2182–90. https://doi.org/10.1073/PNAS.1421993112/-/DCSUPPLEMENTAL/PNAS.1421993112.SD01.XLSX.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” Journal of Molecular Biology 215 (3): 403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Amaral-Zettler, Linda A., Elizabeth A. McCliment, Hugh W. Ducklow, and Susan M. Huse. 2009. “A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes.” PLoS ONE 4 (7). https://doi.org/10.1371/journal.pone.0006372.

Aramaki, Takuya, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, and Hiroyuki Ogata. 2020. “KofamKOALA: KEGG Ortholog Assignment Based on Profile HMM and Adaptive Score Threshold.” Bioinformatics 36 (7). https://doi.org/10.1093/bioinformatics/btz859.

Azam, F, T Fenche, J G Field, J S Gra, L A Meyer-Rei, and F Thingstad. 1983. “The Ecological Role of Water-Column Microbes in the Sea” 10 (November 2015): 257–63. https://doi.org/10.3354/meps010257.

Bachy, Charles, and Alexandra Z. Worden. 2014. “Microbial Ecology: Finding Structure in the Rare Biosphere.” Current Biology 24 (8): R315–17. https://doi.org/10.1016/J.CUB.2014.03.029.

Barth, Alex, Ryan K. Walter, Ian Robbins, and Alexis Pasulka. 2020. “Seasonal and Interannual Variability of Phytoplankton Abundance and Community Composition on the Central Coast of California.” Marine Ecology Progress Series 637. https://doi.org/10.3354/meps13245.

Barton, Andrew D., Stephanie Dutkiewicz, Glenn Flierl, Jason Bragg, and Michael J. Follows. 2010. “Patterns of Diversity in Marine Phytoplankton.” Science 327 (5972): 1509–11. https://doi.org/10.1126/science.1184961.

Becking, Lourens Gerhard Marinus Baas. 1934. Geobiologie of Inleiding Tot de Milieukunde. WP Van Stockum \& Zoon.

Benincá, Elisa, Jef Huisman, Reinhard Heerkloss, Klaus D. Jöhnk, Pedro Branco, Egbert H. Van Nes, Marten Scheffer, and Stephen P. Ellner. 2008. “Chaos in a Long-Term Experiment with a Plankton Community.” Nature 451 (7180): 822–25. https://doi.org/10.1038/nature06512.

Boelaert, Julien, Laura Bendhaiba, Madalina Olteanu, and Nathalie Villa-Vialaneix. 2014. “SOMbrero: An R Package for Numeric and Non-Numeric Self-Organizing Maps.” Advances in Intelligent Systems and Computing 295: 219–28. https://doi.org/10.1007/978-3-319-07695-9_21.

Bograd, Steven J., and Ronald J. Lynn. 2003. “Long-Term Variability in the Southern California Current System.” Deep-Sea Research Part II: Topical Studies in Oceanography 50 (14–16): 2355–70. https://doi.org/10.1016/S0967-0645(03)00131-0.

Bograd, Steven J., Isaac D. Schroeder, and Michael G. Jacox. 2019. “A Water Mass History of the Southern California Current System.” Geophysical Research Letters 46 (12): 6690–98. https://doi.org/10.1029/2019GL082685.

Bokulich, Nicholas A., Benjamin D. Kaehler, Jai Ram Rideout, Matthew Dillon, Evan Bolyen, Rob Knight, Gavin A. Huttley, and J. Gregory Caporaso. 2018. “Optimizing Taxonomic Classification of Marker-Gene Amplicon Sequences with QIIME 2’s Q2-Feature-Classifier Plugin.” Microbiome 6 (1). https://doi.org/10.1186/s40168-018-0470-z.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” Bioinformatics 30 (15): 2114–20. https://doi.org/10.1093/BIOINFORMATICS/BTU170.

Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. “Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2.” Nature Biotechnology 37 (8): 852–57. https://doi.org/10.1038/s41587-019-0209-9.

Bowman, Jeff S., Maria T. Kavanaugh, Scott C. Doney, and Hugh W. Ducklow. 2018. “Recurrent Seascape Units Identify Key Ecological Processes along the Western Antarctic Peninsula.” Global Change Biology 24 (7): 3065–78. https://doi.org/10.1111/gcb.14161.

Bowman, Jeff S, Linda A Amaral-zettler, Jeremy J Rich, Catherine M Luria, and Hugh W Ducklow. 2017. “Bacterial Community Segmentation Facilitates the Prediction of Ecosystem Function along the Coast of the Western Antarctic Peninsula.” Nature Publishing Group 11 (6): 1460–71. https://doi.org/10.1038/ismej.2016.204.

Buchan, Alison, Gary R. LeCleir, Christopher A. Gulvik, and José M. González. 2014. “Master Recyclers: Features and Functions of Bacteria Associated with Phytoplankton Blooms.” Nature Reviews. Microbiology 12 (10): 686–98. https://doi.org/10.1038/nrmicro3326.

Calbet, Albert, and Michael R. Landry. 2004. “Phytoplankton Growth, Microzooplankton Grazing, and Carbon Cycling in Marine Systems.” Limnology and Oceanography 49 (1): 51–57. https://doi.org/10.4319/lo.2004.49.1.0051.

Callahan, Benjamin J, Paul J Mcmurdie, Michael J Rosen, Andrew W Han, and Amy Jo A. 2016. “DADA2: High Resolution Sample Inference from Illumina Amplicon Data” 13 (7): 581–83. https://doi.org/10.1038/nmeth.3869.DADA2.

Cameron, Ellen S., Philip J. Schmidt, Benjamin J.M. Tremblay, Monica B. Emelko, and Kirsten M. Müller. 2021. “Enhancing Diversity Analysis by Repeatedly Rarefying next Generation Sequencing Data Describing Microbial Communities.” Scientific Reports 2021 11:1 11 (1): 1–13. https://doi.org/10.1038/s41598-021-01636-1.

Canals, Oriol, Aleix Obiol, Imer Muhovic, Dolors Vaqué, and Ramon Massana. 2020. “Ciliate Diversity and Distribution across Horizontal and Vertical Scales in the Open Ocean.” Molecular Ecology 29 (15): 2824–39. https://doi.org/10.1111/mec.15528.

Catlett, D., D. A. Siegel, R. D. Simons, N. Guillocheau, F. Henderikx-Freitas, and C. S. Thomas. 2021. “Diagnosing Seasonal to Multi-Decadal Phytoplankton Group Dynamics in a Highly Productive Coastal Ecosystem.” Progress in Oceanography 197 (September): 102637.

Cermeño, Pedro, Stephanie Dutkiewicz, Roger P. Harris, Michael Follows, Oscar Schofield, and Paul G. Falkowski. 2008. “The Role of Nutricline Depth in Regulating the Ocean Carbon Cycle.” PNAS 105 (51): 20344–49.

Chase, Jonathan M., and Mathew A. Leibold. 2002. “Spatial Scale Dictates the Productivity-Biodiversity Relationship.” Nature 416 (6879): 427–30. https://doi.org/10.1038/416427a.

Chavez, Francisco P., Markus Min, Kathleen Pitz, Nathan Truelove, Jacoby Baker, Diana Lascala-Grunewald, Marguerite Blum, et al. 2021. “Observing Life in the Sea Using Environmental DNA.” Oceanography 34 (2). https://doi.org/10.5670/OCEANOG.2021.218.

Checkley, David M., and John A. Barth. 2009. “Patterns and Processes in the California Current System.” Progress in Oceanography 83 (1–4): 49–64. https://doi.org/10.1016/j.pocean.2009.07.028.

Chen, Shifu, Yanqing Zhou, Yaru Chen, and Jia Gu. 2018. “Fastp: An Ultra-Fast All-in-One FASTQ Preprocessor.” Bioinformatics 34 (17): i884–90. https://doi.org/10.1093/BIOINFORMATICS/BTY560.

Chenillat, Fanny, Peter J.S. Franks, and Vincent Combes. 2016. “Biogeochemical Properties of Eddies in the California Current System.” Geophysical Research Letters 43 (11): 5812–20. https://doi.org/10.1002/2016GL068945.

Chenillat, Fanny, Pascal Rivière, Xavier Capet, Peter J.S. Franks, and Bruno Blanke. 2013. “California Coastal Upwelling Onset Variability: Cross-Shore and Bottom-Up

Propagation in the Planktonic Ecosystem." PLoS ONE 8 (5). https://doi.org/10.1371/journal.pone.0062281.

Clayton, Sophie, Stephanie Dutkiewicz, Oliver Jahn, and Michael J. Follows. 2013. "Dispersal, Eddies, and the Diversity of Marine Phytoplankton." Limnology and Oceanography: Fluids and Environments 3 (1): 182–97. https://doi.org/10.1215/21573689-2373515.

Closset, Ivia, Heather M. McNair, Mark A. Brzezinski, Jeffrey W. Krause, Kimberlee Thamatrakoln, and Janice L. Jones. 2021. "Diatom Response to Alterations in Upwelling and Nutrient Dynamics Associated with Climate Forcing in the California Current System." Limnology and Oceanography, 1–16. https://doi.org/10.1002/lno.11705.

Collins, Sinéad, Björn Rost, and Tatiana A. Rynearson. 2014. "Evolutionary Potential of Marine Phytoplankton under Ocean Acidification." Evolutionary Applications 7 (1): 140–55. https://doi.org/10.1111/eva.12120.

Combes, V., F. Chenillat, E. Di Lorenzo, P. Rivière, M. D. Ohman, and S. J. Bograd. 2013. "Cross-Shore Transport Variability in the California Current: Ekman Upwelling vs. Eddy Dynamics." Progress in Oceanography 109: 78–89. https://doi.org/10.1016/j.pocean.2012.10.001.

Costello, Mark J., Peter Tsai, Pui Shan Wong, Alan Kwok Lun Cheung, Zeenatul Basher, and Chhaya Chaudhary. 2017. "Marine Biogeographic Realms and Species Endemicity." Nature Communications 8 (1): 1–9. https://doi.org/10.1038/s41467-017-01121-2.

Cragg, John G., and Russell S. Uhler. 1970. "The Demand for Automobiles." The Canadian Journal of Economics 3 (3). https://doi.org/10.2307/133656.

D'Ovidio, Francesco, Silvia De Monte, Séverine Alvain, Yves Dandonneau, and Marina Lévy. 2010. "Fluid Dynamical Niches of Phytoplankton Types." Proceedings of the National Academy of Sciences of the United States of America 107 (43): 18366–70. https://doi.org/10.1073/pnas.1004620107.

Dupont, Christopher L., John P. Mccrow, Ruben Valas, Ahmed Moustafa, Nathan Walworth, Ursula Goodenough, Robyn Roth, et al. 2015. "Genomes and Gene Expression across Light and Productivity Gradients in Eastern Subtropical Pacific Microbial Communities." ISME Journal 9 (5): 1076–92. https://doi.org/10.1038/ismej.2014.198.

Dutkiewicz, S., M. J. Follows, and J. G. Bragg. 2009. "Modeling the Coupling of Ocean Ecology and Biogeochemistry." Global Biogeochemical Cycles 23 (4): 1–15. https://doi.org/10.1029/2008GB003405.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." PLoS Computational Biology 7 (10). https://doi.org/10.1371/journal.pcbi.1002195.

Edwards, Kyle F., Elena Litchman, and Christopher A. Klausmeier. 2013. "Functional Traits Explain Phytoplankton Community Structure and Seasonal Dynamics in a Marine Ecosystem." Ecology Letters 16 (1): 56–63. https://doi.org/10.1111/ele.12012.

Edwards, Kyle F., Mridul K. Thomas, Christopher A. Klausmeier, and Elena Litchman. 2012. "Allometric Scaling and Taxonomic Variation in Nutrient Utilization Traits and Maximum Growth Rate of Phytoplankton." Limnology and Oceanography 57 (2): 554–66. https://doi.org/10.4319/lo.2012.57.2.0554.

Falkowski, Paul G., Richard T. Barber, and Victor Smetacek. 1998. "Biogeochemical Controls and Feedbacks on Ocean Primary Production." Science 281 (5374): 200–206. https://doi.org/10.1126/SCIENCE.281.5374.200/ASSET/8DA24923-585D-4624-B7D5-1D2D74007006/ASSETS/GRAPHIC/SE298667504A.JPEG.

Falkowski, Paul G., and Matthew J. Oliver. 2007. "Mix and Match: How Climate Selects Phytoplankton." Nature Reviews Microbiology 5 (10). https://doi.org/10.1038/nrmicro1751.

Falkowski, Paul G, Tom Fenchel, and Edward F Delong. 2008. "The Microbial Engines That Drive Earth's Biogeochemical Cycles." Microbial E 320. https://www.science.org.

Fenchel, Tom. 2008. "The Microbial Loop - 25 Years Later." Journal of Experimental Marine Biology and Ecology 366 (1–2): 99–103. https://doi.org/10.1016/j.jembe.2008.07.013.

Fenchel, Tom, and Bland J. Finlay. 2004. "The Ubiquity of Small Species: Patterns of Local and Global Diversity." BioScience 54 (8): 777–84. https://doi.org/10.1641/0006-3568(2004)054[0777:TUOSSP]2.0.CO;2.

Field, Christopher B., Michael J. Behrenfeld, James T. Randerson, and Paul Falkowski. 1998. "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components." Science 281 (5374): 237–40. https://doi.org/10.1126/SCIENCE.281.5374.237/SUPPL_FILE/982246E_THUMB.GIF.

Follows, Michael J., Stephanie Dutkiewicz, Scott Grant, and Sallie W. Chisholm. 2007. "Emergent Biogeography of Microbial Communities in a Model Ocean." Science 315 (5820): 1843–46. https://doi.org/10.1126/science.1138544.

Foukal, Nicholas P., and Andrew C. Thomas. 2014. "Biogeography and Phenology of Satellite-Measured Phytoplankton Seasonality in the California Current." Deep-Sea Research Part I: Oceanographic Research Papers 92: 11–25. https://doi.org/10.1016/j.dsr.2014.06.008.

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." Proceedings of the National Academy of Sciences 105 (22): 7774–78. https://doi.org/10.1073/pnas.0803070105.

Fuhrman, Jed A. 2009. "Microbial Community Structure and Its Functional Implications." Nature 459 (7244): 193–99. https://doi.org/10.1038/nature08058.

Fuhrman, Jed A., Jacob A. Cram, and David M. Needham. 2015. "Marine Microbial Community Dynamics and Their Ecological Interpretation." Nature Reviews Microbiology 13 (3): 133–46. https://doi.org/10.1038/nrmicro3417.

Gast, Christopher J. van der. 2015. “Microbial Biogeography: The End of the Ubiquitous Dispersal Hypothesis?” Environmental Microbiology 17 (3): 544–46. https://doi.org/10.1111/1462-2920.12635.

Gibbons, S. M., J. G. Caporaso, M. Pirrung, D. Field, R. Knight, and J. A. Gilbert. 2013. “Evidence for a Persistent Microbial Seed Bank throughout the Global Ocean.” Proceedings of the National Academy of Sciences 110 (12): 4651–55. https://doi.org/10.1073/pnas.1217767110.

Gilbert, Jack A., Joshua A. Steele, J. Gregory Caporaso, Lars Steinbrück, Jens Reeder, Ben Temperton, Susan Huse, et al. 2012. “Defining Seasonal Marine Microbial Community Dynamics.” ISME Journal 6 (2): 298–308. https://doi.org/10.1038/ismej.2011.107.

Gimmler, Anna, Ralf Korn, Colomban De Vargas, Stéphane Audic, and Thorsten Stoeck. 2016. “The Tara Oceans Voyage Reveals Global Diversity and Distribution Patterns of Marine Planktonic Ciliates.” Scientific Reports 6 (April): 1–13. https://doi.org/10.1038/srep33555.

Gloor, Gregory B., Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. 2016. “It’s All Relative: Analyzing Microbiome Data as Compositions.” Annals of Epidemiology. https://doi.org/10.1016/j.annepidem.2016.03.003.

Goldman, Joel C. 1993. “Potential Role of Large Oceanic Diatoms in New Primary Production.” Deep-Sea Research Part I 40 (1): 159–68. https://doi.org/10.1016/0967-0637(93)90059-C.

Grover, James P. 1990. “Resource Competition in a Variable Environment: Phytoplankton Growing According to Monod’s Model” 136 (6): 771–89.

Grzymski, Joseph J., and Alex M. Dussaq. 2012. “The Significance of Nitrogen Cost Minimization in Proteomes of Marine Microorganisms.” ISME Journal 6 (1). https://doi.org/10.1038/ismej.2011.72.

Guillou, Laure, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, et al. 2013. “The Protist Ribosomal Reference Database (PR2): A Catalog of Unicellular Eukaryote Small Sub-Unit RRNA Sequences with Curated Taxonomy.” Nucleic Acids Research 41 (D1). https://doi.org/10.1093/nar/gks1160.

Harke, Matthew J., and Christopher J. Gobler. 2015. “Daily Transcriptome Changes Reveal the Role of Nitrogen in Controlling Microcystin Synthesis and Nutrient Transport in the Toxic Cyanobacterium, Microcystis Aeruginosa.” BMC Genomics 16 (1): 1–18. https://doi.org/10.1186/S12864-015-2275-9/FIGURES/9.

Haury, L. R., John A. McGowan, and P. H. Wiebe. 1978. “Patterns and Processes in the Time-Space Scales of Plankton Distributions.” Patterns and Processes in the Time-Space Scales of Plankton Distributions.

Hayward, Thomas L., and Elizabeth L. Venrick. 1998. “Nearsurface Pattern in the California Current: Coupling between Physical and Biological Structure.” Deep-Sea Research Part II: Topical Studies in Oceanography. https://doi.org/10.1016/S0967-0645(98)80010-6.

Hsieh, Chih-hao, Christian S. Reiss, William Watson, M. James Allen, John R. Hunter, Robert N. Lea, Richard H. Rosenblatt, Paul E. Smith, and George Sugihara. 2005. “A Comparison of Long-Term Trends and Variability in Populations of Larvae of Exploited and Unexploited Fishes in the Southern California Region: A Community Approach.” Progress in Oceanography 67 (1–2): 160–85. https://doi.org/10.1016/j.pocean.2005.05.002.

Ibarbalz, Federico M, Nicolas Henry, Fabien Lombard, Chris Bowler, Lucie Zinger, Greta Busseni, and Hannah Byrne. 2019. “Global Trends in Marine Plankton Diversity across Kingdoms of Life,” 1084–97. https://doi.org/10.1016/j.cell.2019.10.008.

Istvánovics, V. 2009. “Eutrophication of Lakes and Reservoirs.” Encyclopedia of Inland Waters, January, 157–65. https://doi.org/10.1016/B978-012370626-3.00141-1.

Jacox, Michael G., Christopher A. Edwards, Elliott L. Hazen, and Steven J. Bograd. 2018. “Coastal Upwelling Revisited: Ekman, Bakun, and Improved Upwelling Indices for the U.S. West Coast.” Journal of Geophysical Research: Oceans 123 (10): 7332–50. https://doi.org/10.1029/2018JC014187.

James, Chase C., Andrew D. Barton, Lisa Zeigler Allen, Robert H. Lampe, Ariel Rabines, Anne Schulberg, Hong Zheng, Ralf Goericke, Kelly D. Goodwin, and Andrew E. Allen. 2022. “Influence of Nutrient Supply on Plankton Microbiome Biodiversity and Distribution in a Coastal Upwelling Region.” Nature Communications 2022 13:1 13 (1): 1–13. https://doi.org/10.1038/s41467-022-30139-4.

Johnson, Jerald B., and Kristian S. Omland. 2004. “Model Selection in Ecology and Evolution.” Trends in Ecology and Evolution 19 (2): 101–8. https://doi.org/10.1016/j.tree.2003.10.013.

Kahru, Mati, Michael G. Jacox, and Mark D. Ohman. 2018. “CCE1: Decrease in the Frequency of Oceanic Fronts and Surface Chlorophyll Concentration in the California Current System during the 2014–2016 Northeast Pacific Warm Anomalies.” Deep-Sea Research Part I: Oceanographic Research Papers 140 (January): 4–13. https://doi.org/10.1016/j.dsr.2018.04.007.

Kahru, Mati, and B. Greg Mitchell. 2001. “Seasonal and Nonseasonal Variability of Satellite-Derived Chlorophyll and Colored Dissolved Organic Matter Concentration in the California Current.” Journal of Geophysical Research: Oceans 106 (C2): 2517–29. https://doi.org/10.1029/1999jc000094.

Karl, David M., and Roger Lukas. 1996. “The Hawaii Ocean Time-Series (HOT) Program: Background, Rationale and Field Implementation.” Deep-Sea Research Part II: Topical Studies in Oceanography 43 (2–3): 129–56. https://doi.org/10.1016/0967-0645(96)00005-7.

Kenitz, Kasia M., Eric C. Orenstein, Paul L.D. Roberts, Peter J.S. Franks, Jules S. Jaffe, Melissa L. Carter, and Andrew D. Barton. 2020. “Environmental Drivers of Population Variability in Colony-Forming Marine Diatoms.” Limnology and Oceanography 65 (10): 2515–28. https://doi.org/10.1002/lno.11468.

Kim, Hey Jin, Arthur J. Miller, John McGowan, and Melissa L. Carter. 2009. “Coastal Phytoplankton Blooms in the Southern California Bight.” Progress in Oceanography 82 (2): 137–47. https://doi.org/10.1016/j.pocean.2009.05.002.

Kohonen, Teuvo. 1997. “Exploration of Very Large Databases by Self-Organizing Maps.” IEEE International Conference on Neural Networks - Conference Proceedings 1. https://doi.org/10.1109/ICNN.1997.611622.

Kolody, Bethany C., Matthew J. Harke, Sharon E. Hook, and Andrew E. Allen. 2022. “Transcriptomic and Metatranscriptomic Approaches in Phytoplankton: Insights and Advances.” Advances in Phytoplankton Ecology, January, 435–85. https://doi.org/10.1016/B978-0-12-822861-6.00022-4.

Kolody, Bethany C, J P McCrow, L Zeigler Allen, F O Aylward, K M Fontanez, A Moustafa, M Moniruzzaman, et al. 2019. “Diel Transcriptional Response of a California Current Plankton Microbiome to Light, Low Iron, and Enduring Viral Infection.” The ISME Journal. https://doi.org/10.1038/s41396-019-0472-2.

Krogh, Anders, Björn Larsson, Gunnar Von Heijne, and Erik L.L. Sonnhammer. 2001. “Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes.” Journal of Molecular Biology 305 (3). https://doi.org/10.1006/jmbi.2000.4315.

Landa, Marine, Andrew S. Burns, Selena J. Roth, and Mary Ann Moran. 2017. “Bacterial Transcriptome Remodeling during Sequential Co-Culture with a Marine Dinoflagellate and Diatom.” The ISME Journal 2017 11:12 11 (12): 2677–90. https://doi.org/10.1038/ismej.2017.117.

Laws, Edward A., Eurico D’Sa, and Puneeta Naik. 2011. “Simple Equations to Estimate Ratios of New or Export Production to Total Production from Satellite-Derived Estimates of Sea Surface Temperature and Primary Production.” Limnology and Oceanography: Methods 9 (DECEMBER). https://doi.org/10.4319/lom.2011.9.593.

Laws, Edward A., Paul G. Falkowski, Walker O. Smith, Hugh Ducklow, and James J. McCarthy. 2000. “Temperature Effects on Export Production in the Open Ocean.” Global Biogeochemical Cycles 14 (4). https://doi.org/10.1029/1999GB001229.

Legendre, Pierre, and Louis Legendre. 2012. Numerical Ecology.

Leibold, M. A., M. Holyoak, Nicolas Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, Robert D. Holt, et al. 2004. “The Metacommunity Concept: A Framework for Multi-Scale Community Ecology.” Ecology Letters 7 (7): 601–13. https://doi.org/10.1111/j.1461-0248.2004.00608.x.

Lennon, Jay T., and Stuart E. Jones. 2011. “Microbial Seed Banks: The Ecological and Evolutionary Implications of Dormancy.” Nature Reviews Microbiology 2011 9:2 9 (2): 119–30. https://doi.org/10.1038/nrmicro2504.

Lévy, Marina, Peter J.S. Franks, and K. Shafer Smith. 2018. “The Role of Submesoscale Currents in Structuring Marine Ecosystems.” Nature Communications 9 (1): 4758. https://doi.org/10.1038/s41467-018-07059-3.

Li, Weizhong, and Adam Godzik. 2006. “Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences.” Bioinformatics 22 (13): 1658–59. https://doi.org/10.1093/BIOINFORMATICS/BTL158.

Lilly, Laura E., and Mark D. Ohman. 2018. “CCE IV: El Niño-Related Zooplankton Variability in the Southern California Current System.” Deep-Sea Research Part I: Oceanographic Research Papers 140 (June): 36–51. https://doi.org/10.1016/j.dsr.2018.07.015.

Lindegren, Martin, David M. Checkley, Mark D. Ohman, J. Anthony Koslow, and Ralf Goericke. 2016. “Resilience and Stability of a Pelagic Marine Ecosystem.” Proceedings of the Royal Society B: Biological Sciences 283 (1822). https://doi.org/10.1098/rspb.2015.1931.

Logares, Ramiro, Stéphane Audic, David Bass, Lucie Bittner, Christophe Boutte, Richard Christen, Jean Michel Claverie, et al. 2014. “Patterns of Rare and Abundant Marine Microbial Eukaryotes.” Current Biology 24 (8): 813–21. https://doi.org/10.1016/J.CUB.2014.02.050.

Logares, Ramiro, Ina M. Deutschmann, Pedro C. Junger, Caterina R. Giner, Anders K. Krabberød, Thomas S.B. Schmidt, Laura Rubinat-Ripoll, et al. 2020. “Disentangling the Mechanisms Shaping the Surface Ocean Microbiota.” Microbiome 8 (1): 1–17. https://doi.org/10.1186/S40168-020-00827-8/FIGURES/4.

Lozupone, Catherine A., and Rob Knight. 2007. “Global Patterns in Bacterial Diversity.” Proceedings of the National Academy of Sciences of the United States of America 104 (27): 11436–40. https://doi.org/10.1073/PNAS.0611525104.

Maier, Marco J. 2014. “DirichletReg: Dirichlet Regression for Compositional Data in R.” Vienna. https://epub.wu.ac.at/4077/.

Malviya, Shruti, Eleonora Scalco, Stéphane Audic, Flora Vincent, Alaguraj Veluchamy, Julie Poulain, Patrick Wincker, et al. 2016. “Insights into Global Diatom Distribution and Diversity in the World’s Ocean.” Proceedings of the National Academy of Sciences 113 (11): E1516–25. https://doi.org/10.1073/pnas.1509523113.

Mantyla, Arnold W, Elizabeth L Venrick, and Thomas L Hayward. 1995. “Primary Production and Chlorophyll Relationships, Derived from Ten Year of CalCOFI Measurements” 36: 159–66.

Margalef, Ramon. 1978. “Life-Forms of Phytoplankton as Survival Alternatives in an Unstable Environment.” Oceanolgica Acta 1. https://doi.org/10.1007/BF00202661.

Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” EMBnet.Journal 17 (1). https://doi.org/10.14806/ej.17.1.200.

Martín, Paula Villa, Aleš Buček, Thomas Bourguignon, and Simone Pigolotti. 2020. “Ocean Currents Promote Rare Species Diversity in Protists.” Science Advances 6 (29). https://doi.org/10.1126/SCIADV.AAZ9037/SUPPL_FILE/AAZ9037_SM.PDF.

Martiny, Jennifer B.Hughes, Brendan J.M. Bohannan, James H. Brown, Robert K. Colwell, Jed A. Fuhrman, Jessica L. Green, M. Claire Horner-Devine, et al. 2006. “Microbial Biogeography: Putting Microorganisms on the Map.” Nature Reviews Microbiology 4 (2): 102–12. https://doi.org/10.1038/nrmicro1341.

Mcclatchie, Sam, Michael G. Jacox, Mark D. Ohman, Linsey M. Sala, Ralf Goericke, Mati Kahru, Bill Peterson, et al. 2016. “State of the California Current 2015-16: Comparisons with the 1997-98 El Niño.” CalCOFI Report 57.

McMurdie, Paul J., and Susan Holmes. 2014. “Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible.” PLoS Computational Biology 10 (4). https://doi.org/10.1371/journal.pcbi.1003531.

Mende, Daniel R., Jessica A. Bryant, Frank O. Aylward, John M. Eppley, Torben Nielsen, David M. Karl, and Edward F. Delong. 2017. “Environmental Drivers of a Microbial Genomic Transition Zone in the Ocean’s Interior.” Nature Microbiology 2 (10). https://doi.org/10.1038/s41564-017-0008-3.

Moisan, Tiffany A., Kay M. Rufty, John R. Moisan, and Matthew A. Linkswiler. 2017. “Satellite Observations of Phytoplankton Functional Type Spatial Distributions, Phenology, Diversity, and Ecotones.” Frontiers in Marine Science 4 (JUN): 1–24. https://doi.org/10.3389/fmars.2017.00189.

Morris, Robert M., Michael S. Rappé, Stephanie A. Connon, Kevin L. Vergin, William A. Siebold, Craig A. Carlson, and Stephen J. Giovannoni. 2002. “SAR11 Clade Dominates Ocean Surface Bacterioplankton Communities.” Nature 420 (6917). https://doi.org/10.1038/nature01240.

Mullin, Michael M. 1998. “Biomasses of Large-Celled Phytoplankton and Their Relation to the Nitricline and Grazing in the California Current System off Southern California, 1994-1996.” California Cooperative Oceanic Fisheries Investigations Reports 39: 117–23.

Needham, David M., and Jed A. Fuhrman. 2016. “Pronounced Daily Succession of Phytoplankton, Archaea and Bacteria Following a Spring Bloom.” Nature Microbiology 1. https://doi.org/10.1038/nmicrobiol.2016.5.

Nemergut, D. R., S. K. Schmidt, T. Fukami, S. P. O’Neill, T. M. Bilinski, L. F. Stanish, J. E. Knelman, et al. 2013. “Patterns and Processes of Microbial Community Assembly.”

Microbiology and Molecular Biology Reviews 77 (3): 342–56. https://doi.org/10.1128/mmbr.00051-12.

Not, Fabrice, Raffaele Siano, Wiebe H.C.F. Kooistra, Nathalie Simon, Daniel Vaulot, and Ian Probert. 2012. “Diversity and Ecology of Eukaryotic Marine Phytoplankton.” Advances in Botanical Research 64 (January): 1–53. https://doi.org/10.1016/B978-0-12-391499-6.00001-3.

Oksanen, Author Jari, F Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan Mcglinn, Peter R Minchin, et al. 2020. “Package ‘ Vegan .’”

Ottesen, Elizabeth A, Curtis R Young, John M Eppley, John P Ryan, Francisco P Chavez, Christopher A Scholin, and Edward F Delong. 2013. “Pattern and Synchrony of Gene Expression among Sympatric Marine Microbial Populations.” https://doi.org/10.1073/pnas.1222099110.

Parada, Alma E., David M. Needham, and Jed A. Fuhrman. 2016. “Every Base Matters: Assessing Small Subunit RRNA Primers for Marine Microbiomes with Mock Communities, Time Series and Global Field Samples.” Environmental Microbiology 18 (5). https://doi.org/10.1111/1462-2920.13023.

Partensky, F., J. Blanchot, and Daniel Vaulot. 1999. “Differential Distribution and Ecology of Prochlorococcus and Synechococcus in Oceanic Waters: A Review.” Bulletin-Institut Oceanographique Monaco 19: 457–75. http://cat.inist.fr/?aModele=afficheN&cpsidt=1218663.

Pedregosa, Fabian, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” Journal of Machine Learning Research 12.

Phoma, Boitumelo Sandra, and Thulani Peter Makhalanyane. 2021. “Depth-Dependent Variables Shape Community Structure and Functionality in the Prince Edward Islands.” Microbial Ecology 81 (2). https://doi.org/10.1007/s00248-020-01589-4.

Podell, Sheila, and Terry Gaasterland. 2007. “DarkHorse: A Method for Genome-Wide Prediction of Horizontal Gene Transfer.” Genome Biology 8 (2). https://doi.org/10.1186/gb-2007-8-2-r16.

Powell, Jesse R., and Mark D. Ohman. 2015a. “Changes in Zooplankton Habitat, Behavior, and Acoustic Scattering Characteristics across Glider-Resolved Fronts in the Southern California Current System.” Progress in Oceanography 134: 77–92. https://doi.org/10.1016/j.pocean.2014.12.011.

Powell, Jesse R, and Mark D Ohman. 2015b. “Covariability of Zooplankton Gradients with Glider-Detected Density Fronts in the Southern California Current System.” Deep Sea Research Part II: Topical Studies in Oceanography 112: 79–90.

Pruesse, Elmar, Christian Quast, Katrin Knittel, Bernhard M. Fuchs, Wolfgang Ludwig, Jörg Peplies, and Frank Oliver Glöckner. 2007. “SILVA: A Comprehensive Online Resource for Quality Checked and Aligned Ribosomal RNA Sequence Data Compatible with ARB.” Nucleic Acids Research 35 (21). https://doi.org/10.1093/nar/gkm864.

Rho, Mina, Haixu Tang, and Yuzhen Ye. 2010. “FragGeneScan: Predicting Genes in Short and Error-Prone Reads.” Nucleic Acids Research 38 (20). https://doi.org/10.1093/nar/gkq747.

Richardson, A. J., A. W. Walne, A. W.G. John, T. D. Jonas, J. A. Lindley, D. W. Sims, D. Stevens, and M. Witt. 2006. “Using Continuous Plankton Recorder Data.” Progress in Oceanography 68 (1): 27–74. https://doi.org/10.1016/j.pocean.2005.09.011.

Righetti, Damiano, Meike Vogt, Nicolas Gruber, Achilleas Psomas, and Niklaus E. Zimmermann. 2019. “Global Pattern of Phytoplankton Diversity Driven by Temperature and Environmental Variability.” Science Advances 5 (5): 1–11. https://doi.org/10.1126/sciadv.aau6253.

Robertson, Gordon, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D. Jackman, Karen Mungall, et al. 2010. “De Novo Assembly and Analysis of RNA-Seq Data.” Nature Methods 2010 7:11 7 (11): 909–12. https://doi.org/10.1038/nmeth.1517.

Rudnick, Daniel L., Katherine D. Zaba, Robert E. Todd, and Russ E. Davis. 2017. “A Climatology of the California Current System from a Network of Underwater Gliders.” Progress in Oceanography 154: 64–106. https://doi.org/10.1016/j.pocean.2017.03.002.

Rusch, Douglas B., Aaron L. Halpern, Granger Sutton, Karla B. Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, et al. 2007. “The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.” PLOS Biology 5 (3): e77. https://doi.org/10.1371/JOURNAL.PBIO.0050077.

Rykaczewski, Ryan R., and David M. Checkley. 2007. “Influence of Ocean Winds on the Pelagic Ecosystem in Upwelling Regions.” PNAS 105 (6): 1965–70.

Ryther, John H. 1969. “Photosynthesis and Fish Production in the Sea.” Science 166 (3901): 72–76. https://doi.org/10.1126/science.166.3901.72.

Schmieder, Robert, Yan Wei Lim, and Robert Edwards. 2012. “Identification and Removal of Ribosomal RNA Sequences from Metatranscriptomes.” Bioinformatics 28 (3): 433–35. https://doi.org/10.1093/BIOINFORMATICS/BTR669.

Ser-Giacomi, Enrico, Lucie Zinger, Shruti Malviya, Colomban De Vargas, Eric Karsenti, Chris Bowler, and Silvia De Monte. 2018. “Ubiquitous Abundance Distribution of Non-Dominant Plankton across the Global Ocean.” Nature Ecology and Evolution 2 (8): 1243–49. https://doi.org/10.1038/s41559-018-0587-2.

Shmida, Avi, and Mark V. Wilson. 1985. “Biological Determinants of Species Diversity.” Journal of Biogeography 12 (1). https://doi.org/10.2307/2845026.

Sommer, Ulrich, Herwig Stibor, Alexis Katechakis, Frank Sommer, and Thomas Hansen. 2002. "Pelagic Food Web Configurations at Different Levels of Nutrient Richness and Their Implications for the Ratio Fish Production:Primary Production." Hydrobiologia 484: 11–20. https://doi.org/10.1023/A:1021340601986.

Steinberg, Deborah K, Craig A Carlson, Nicholas R Bates, Rodney J Johnson, Anthony F Michaels, and Anthony H Knap. 2015. "Overview of the US JGOFS Bermuda Atlantic Time-Series Study ( BATS ): A Decade-Scale Look at Ocean Biology and Biogeochemistry Overview of the US JGOFS Bermuda Atlantic Time-Series Study ( BATS ): A Decade-Scale Look at Ocean Biology and Biogeochemistry" 48 (October): 1405–47. https://doi.org/10.1016/S0967-0645(00)00148-X.

Stock, Charles A., Jasmin G. John, Ryan R. Rykaczewski, Rebecca G. Asch, William W.L. Cheung, John P. Dunne, Kevin D. Friedland, Vicky W.Y. Lam, Jorge L. Sarmiento, and Reg A. Watson. 2017. "Reconciling Fisheries Catch and Ocean Productivity." Proceedings of the National Academy of Sciences of the United States of America 114 (8): E1441–49. https://doi.org/10.1073/PNAS.1610238114/SUPPL_FILE/PNAS.1610238114.SM01.MOV.

Stukel, Michael R., Lihini I. Aluwihare, Katherine A. Barbeau, Alexander M. Chekalyuk, Ralf Goericke, Arthur J. Miller, Mark D. Ohman, et al. 2017. "Mesoscale Ocean Fronts Enhance Carbon Export Due to Gravitational Sinking and Subduction." Proceedings of the National Academy of Sciences 114 (6): 1252–57. https://doi.org/10.1073/pnas.1609435114.

Sunagawa, Shinichi, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, et al. 2015. "Structure and Function of the Global Ocean Microbiome." Science 348 (6237): 1–10. https://doi.org/10.1126/science.1261359.

Talbot, Jennifer M., Thomas D. Bruns, John W. Taylor, Dylan P. Smith, Sara Branco, Sydney I. Glassman, Sonya Erlandson, et al. 2014. "Endemism and Functional Convergence across the North American Soil Mycobiome." Proceedings of the National Academy of Sciences of the United States of America 111 (17): 6341–46. https://doi.org/10.1073/pnas.1402584111.

Tatusov, Roman L., Natalie D. Fedorova, John D. Jackson, Aviva R. Jacobs, Boris Kiryutin, Eugene V. Koonin, Dmitri M. Krylov, et al. 2003. "The COG Database: An Updated Vesion Includes Eukaryotes." BMC Bioinformatics 4. https://doi.org/10.1186/1471-2105-4-41.

Taylor, Andrew G., and Michael R. Landry. 2018. "Phytoplankton Biomass and Size Structure across Trophic Gradients in the Southern California Current and Adjacent Ocean Ecosystems." Marine Ecology Progress Series 592 (March): 1–17. https://doi.org/10.3354/MEPS12526.

Taylor, Andrew G., Michael R. Landry, Karen E. Selph, and John J. Wokuluk. 2015. "Temporal and Spatial Patterns of Microbial Community Biomass and Composition in the Southern

California Current Ecosystem." Deep-Sea Research Part II: Topical Studies in Oceanography 112. https://doi.org/10.1016/j.dsr2.2014.02.006.

Thompson, Andrew R., Issac D. Schroeder, Steven J. Bograd, Elliott L. Hazen, Michael G. Jacox, Andrew Leising, Brian K Wells, et al. 2018. "State of the California Current 2017-18: Still Not Quite Normal in the North and Getting Interesting in the South" 59 (December).

Tréguer, Paul, Chris Bowler, Brivaela Moriceau, Stephanie Dutkiewicz, Marion Gehlen, Olivier Aumont, Lucie Bittner, et al. 2017. "Influence of Diatom Diversity on the Ocean Biological Carbon Pump." Nature Geoscience 2017 11:1 11 (1): 27–37. https://doi.org/10.1038/s41561-017-0028-x.

Vallina, S M, M J Follows, S Dutkiewicz, J M Montoya, P Cermeno, and M Loreau. 2014. "Global Relationship between Phytoplankton Diversity and Productivity in the Ocean." Nature Communications, 1–10. https://doi.org/10.1038/ncomms5299.

Vargas, Colomban de, Stéphane Audic, N. Henry, J. Decelle, Frederic Mahe, Ramiro Logares, E. Lara, et al. 2015. "Eukaryotic Plankton Diversity in the Sunlit Ocean." Science 348 (6237): 1261605-1/11. https://doi.org/10.1007/s13398-014-0173-7.2.

Vellend, Mark. 2010. "Conceptual Synthesis in Community Ecology." Quarterly Review of Biology 85 (2): 183–206. https://doi.org/10.1086/652373.

Venrick, E. L. 2009. "Floral Patterns in the California Current: The Coastal-Offshore Boundary Zone." Journal of Marine Research 67 (1): 89–111. https://doi.org/10.1357/002224009788597917.

Villarino, Ernesto, James R. Watson, Bror Jönsson, Josep M. Gasol, Guillem Salazar, Silvia G. Acinas, Marta Estrada, et al. 2018. "Large-Scale Ocean Connectivity and Planktonic Body Size." Nature Communications 9 (1). https://doi.org/10.1038/s41467-017-02535-8.

Vorobev, Alexey, Marion Dupouy, Quentin Carradec, Tom O. Delmont, Anita Annamalé, Patrick Wincker, and Eric Pelletier. 2020. "Transcriptome Reconstruction and Functional Analysis of Eukaryotic Marine Plankton Communities via High-Throughput Metagenomics and Metatranscriptomics." Genome Research 30 (4): 647–59. https://doi.org/10.1101/GR.253070.119.

Walker Jr, H J, Philip A Hastings, John R Hyde, Robert N Lea, Owyn E Snodgrass, and Lyall F Bellquist. 2020. "Unusual Occurrences of Fishes in the Southern California Current System during the Warm Water Period of 2014--2018." Estuarine, Coastal and Shelf Science 236: 106634.

Wang, Yongming, Jie Pan, Jun Yang, Zhichao Zhou, Yueping Pan, and Meng Li. 2020. "Patterns and Processes of Free-Living and Particle-Associated Bacterioplankton and Archaeaplankton Communities in a Subtropical River-Bay System in South China." Limnology and Oceanography 65 (S1). https://doi.org/10.1002/lno.11314.

Wang, Zhao, Doris L. Juarez, Jin-Fen Pan, Sara K. Blinebry, Jessica Gronniger, James S. Clark, Zackary I. Johnson, and Dana E. Hunt. 2019. “Microbial Communities across Nearshore to Offshore Coastal Transects Are Primarily Shaped by Distance and Temperature.” Environmental Microbiology, 1462-2920.14734. https://doi.org/10.1111/1462-2920.14734.

Ward, Ben A., B. B. Cael, Sinead Collins, and C. Robert Young. 2021. “Selective Constraints on Global Plankton Dispersal.” Proceedings of the National Academy of Sciences of the United States of America 118 (10). https://doi.org/10.1073/pnas.2007388118.

Ward, Christopher S., Cheuk Man Yung, Katherine M. Davis, Sara K. Blinebry, Tiffany C. Williams, Zackary I. Johnson, and Dana E. Hunt. 2017. “Annual Community Patterns Are Driven by Seasonal Switching between Closely Related Marine Bacteria.” ISME Journal 11 (6). https://doi.org/10.1038/ismej.2017.4.

Weber, Edward D., Toby D. Auth, Simone Baumann-Pickering, Timothy R. Baumgartner, Eric P. Bjorkstedt, Steven J. Bograd, Brian J. Burke, et al. 2021. “State of the California Current 2019–2020: Back to the Future With Marine Heatwaves?” Frontiers in Marine Science 8. https://doi.org/10.3389/fmars.2021.709454.

Wells, Brian K, Issac D. Schroeder, Jarrod A Santora, Jennifer Fisher, W.T. Peterson, Eric Bjorkstedt, Roxanne R. Robertson, et al. 2017. “State of the California Current 2016-17: Still Anything but ‘Normal’ in the North.”

Williams, Richard G., and Michael J. Follows. 2011. Ocean Dynamics and the Carbon Cycle: Principles and Mechanisms. Book.

Zaba, Katherine D., and Daniel L. Rudnick. 2016. “The 2014-2015 Warming Anomaly in the Southern California Current System Observed by Underwater Gliders.” Geophysical Research Letters 43 (3): 1241–48. https://doi.org/10.1002/2015GL067550.

Zhu, Zhi, Pingping Qu, Feixue Fu, Nancy Tennenbaum, Avery O. Tatters, and David A. Hutchins. 2017. “Understanding the Blob Bloom: Warming Increases Toxicity and Abundance of the Harmful Bloom Diatom Pseudo-Nitzschia in California Coastal Waters.” Harmful Algae 67: 36–43. https://doi.org/10.1016/j.hal.2017.06.004.