Dimensions of Microbial Biodiversity in the North Pacific Subtropical Gyre

By

Jessica A. Bryant
B.S., University of California at Santa Cruz (2005)
M.S., University of Oregon at Eugene (2008)

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

© Massachusetts Institute of Technology 2017. All rights reserved.

The author hereby grants to MIT permission to reproduce and to distribute publicly paper and electronic copies of this thesis document in whole or in part in any medium now known or hereafter created.

Signature redacted

Signature of Author

Department of Civil and Environmental Engineering
Feb 13, 2017

Signature redacted

Certified by..........................

Edward F. DeLong
Professor of Oceanography
Department of Civil and Environmental Engineering
at MIT and at University of Hawaiʻi, Mānoa

Signature redacted

Accepted by..........................

Jesse Kroll
Professor of Civil and Environmental Engineering
Chair, Graduate Program Committee
Dimensions of Microbial Biodiversity in the North Pacific Subtropical Gyre

by

Jessica A. Bryant

Submitted to the Department of Civil and Environmental Engineering
On Feb 14, 2017, in partial fulfillment the requirements
for the degree of Doctor of Philosophy

Abstract

Microorganisms support all life in the oceans and are fundamental to maintaining a habitable biosphere on Earth. However an understanding of their taxonomic and functional distributions across space and time are just beginning to emerge and numerous niches within the marine environment are still awaiting exploration. The motivation for this thesis is to improve our understanding of distributions of microbes and their metabolic potential at Station ALOHA, a long-term study site representative of the North Pacific Subtropical Gyre (NPSG). We observed changes in diversity and community composition at Station ALOHA across time, ocean depth and on plastic debris, a new anthropogenically derived niche in the NPSG. Despite surface waters only experiencing mild seasonal variation in the abiotic environment, using near monthly picoplankton samples collected across a 2-year period at 25m depth, we observed that microbial community composition correlated with solar irradiance, thereby demonstrating seasonal trends. Ocean surface microbes are known to differ fundamentally from those found in the ocean's interior, yet the nature of the transitions from shallow to deep surface water communities is not well understood. Using a high resolution depth series across twelve time points, we observed that microbial communities partitioned into four groups that consisted of all samples above the deep chlorophyll maximum (DCM), 125m samples below the DCM, all 200 m samples and all 500, 770 and 1000m samples. Our data also revealed a sharp discontinuity in genomic traits including GC%, genome size and proteome elemental composition spanning the DCM, suggesting that nitrogen limitation was key to shaping this sharp genomic transition zone across disparate clades. In contrast, we observed that plastic debris in the NPSG forms a habitat for complex microbial assemblages that have organisms, lifestyles and metabolic pathways that are distinct and potentially less nutrient limited than picoplankton in the surrounding water column. Taken together this work helps expand our understanding of spatial and temporal distributions of microorganisms at Station ALOHA and can help direct future microbial oceanography surveys, highlighting new directions for future research.

Thesis Supervisor:
Edward F. DeLong, Professor
Department of Civil and Environmental Engineering and Biological Engineering, MIT
Acknowledgements

Financial for this work was provided by the US EPA Environmental Star Fellowship program, the National Science Foundation Center for Microbial Oceanography: Research and Education (award #EF0424599) and the Gordon and Betty Moore Foundation (grant #492.01, #3777, and #3298).

This work would not have been possible without the many people who helped me during my time at MIT. I am enormously grateful for the scientific guidance, thought-provoking discussions, advice and the company I have received throughout this exciting journey.

A special thanks to:

My advisor, Edward F. DeLong, whose passion and dedication to discovery is both inspiring and infectious. I will forever be grateful for Ed’s scientific mentorship, support and constant enthusiasm, and for his rapid responses to my emails.

My committee members, Eric Alm and Martin Polz for helpful feedback.

DeLong Lab members, both past and present for being phenomenal collaborators, colleagues, mentors and surfing buddies.

Members of the Cordero Lab for being my ‘lab away from lab’ at Parsons.

This thesis is dedicated to my family: To my parents, Judy and Bill, for nurturing my curiosity of the natural world. I cannot imagine many other parents in the neighborhood I grew up in were nearly as willing to take their young daughter frog hunting. To my sister Steph, for perspective, encouragement and for accompanying many of the childhood frog hunting expeditions. To my partner Sean, for making our various Cambridge apartments all feel like home.
# Table of Contents

## Abstract ............................................................................................................. 3

## Acknowledgements ............................................................................................. 4

## Chapter 1. Introduction ......................................................................................... 7

### Overview of Work Presented in this Thesis ......................................................... 11

### References ......................................................................................................... 12

## Chapter 2. Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre ................................................................. 17

### Abstract ........................................................................................................... 18

### Introduction ...................................................................................................... 19

### Materials and Methods ..................................................................................... 20

#### Site and sample collection ............................................................................. 21

#### DNA extraction and sequencing ..................................................................... 22

#### Metagenomic sequence analysis and annotation ........................................... 23

#### Amplicon sequence analysis and annotation .................................................. 24

#### Statistical analyses ....................................................................................... 25

#### Weighted co-occurrence network analyses .................................................... 26

### Results and Discussion ..................................................................................... 29

#### Microbial communities at 25 versus 500 m in the NPSG ............................... 30

#### Wind correlates with alpha diversity in NPSG surface waters ..................... 32

#### Solar irradiance correlates with beta diversity in surface water .................... 34

#### Co-occurrence network analyses of OTUs and orthologs ............................. 38

### Conclusions ...................................................................................................... 40

### Figures .............................................................................................................. 42

### References ....................................................................................................... 49

### Acknowledgements .......................................................................................... 54

### Supplementary Materials for Chapter 2 ................................................................ 56

## Chapter 3. A genomic inflection point in the twilight zone of the ocean’s interior .................................................................................................................... 90

### Abstract ........................................................................................................... 90

### Main Text ......................................................................................................... 91

### References ....................................................................................................... 102

### Acknowledgements .......................................................................................... 105

### Supplementary materials for Chapter 3 ............................................................ 106

#### Materials and Methods .................................................................................. 110

#### Supplementary Material References ............................................................. 118

#### Supplementary Figures ................................................................................... 121

## Chapter 4. Diversity and activity of communities inhabiting plastic debris in the North Pacific Gyre .................................................................................. 138

### Abstract ........................................................................................................... 138

### Introduction ...................................................................................................... 140

### Materials and Methods .................................................................................... 142

#### Biotic Measurements ...................................................................................... 143

#### SEM ............................................................................................................ 144

#### SUPER HI-CAT Library Construction, Sequencing and Annotation ............. 144

## References ......................................................................................................... 12
Comparison to Plastic Debris from the North Atlantic Subtropical Gyre ........................................ 147

Results ......................................................................................................................................... 149
Concentration and Size Distribution of Plastic ........................................................................... 156
Biotic Activity On Microplastics ........................................................................................................... 149
Eukaryotic and Prokaryotic Organisms on Microplastics ............................................................... 150
Taxonomic and Functional Gene Comparison to Surrounding Seawater ........................................ 153
Discussion ................................................................................................................................... 154

Figures ..................................................................................................................................... 162
References ................................................................................................................................. 171
Acknowledgments ......................................................................................................................... 177
Supplementary Materials for Chapter 4 ...................................................................................... 178

Chapter 5. Successional dynamics of microbial biofilm communities on plastic surfaces .................................................................................................................................................................................. 188

Abstract ...................................................................................................................................... 188
Introduction ................................................................................................................................... 189
Methods ...................................................................................................................................... 191
  SEM Imaging .............................................................................................................................. 192
  DNA Extractions ...................................................................................................................... 193
  Sequence Processing and Statistical Analyses ............................................................................ 193
Results ...................................................................................................................................... 195
  Richness and composition of bacterial communities on plastic squares ................................. 196
  Comparison of Bacterial Communities on Plastics Versus Nearby Habitats ............................. 197
  Correspondence Between Bacteria and Photosynthetic Eukaryotes ........................................ 197
Discussion .................................................................................................................................. 197
Figures ...................................................................................................................................... 204
References ................................................................................................................................... 215
Acknowledgements ....................................................................................................................... 219

Chapter 6. Summary and Future Directions ..................................................................................... 221

Variability in Surface Planktonic Communities ............................................................................. 223
Microbial Communities Across the Depth Gradient ....................................................................... 224
Microbes on Plastic Debris .......................................................................................................... 227
References ................................................................................................................................. 230
Chapter 1. Introduction

Microorganisms, the ocean’s tiniest residence, play key roles in all biogeochemical cycles on Earth. Marine cyanobacteria significantly contribute to global primary production, approximately 50% of which occurs in the oceans (Field et al., 1998). Bacteria then remineralize a large fraction of this fixed carbon in the euphotic zone (del Giorgio and Duarte, 2002). In addition, microbial transformations control the availability of nitrogen, a key and often limiting nutrient and impact phosphorous cycling and micronutrient availability (Karl, 2014; Vraspir and Butler, 2009; Zehr and Kudela, 2011; Voss et al., 2013). In sum, marine microorganisms support all life in the oceans and are fundamental to maintaining a habitable biosphere on Earth. Recent sampling efforts have shown that marine microbial communities encompass a tremendous quantity of diversity (Rusch et al., 2007; DeLong et al., 2006). Given the fundamental role of microorganisms in the ecosystem, it is vital that we understand the structure, function and drivers of this diversity, in order to accurately predict and potentially re-engineer the impacts of natural and anthropogenic environmental changes on marine systems (Follows and Dutkiewicz, 2011; Landa et al., 2016).

In 1988, scientists from the Hawaii Ocean time-series (HOT) began monthly sampling at Station ALOHA (A Long-term Oligotrophic Habitat Assessment) located in the North Pacific Subtropical Gyre (NPSG). The aim was to help address the severe undersampling of open ocean ecosystems with a focus on improving our understanding of marine microbial processes (Karl and Lukas, 1996). This station was established under the premise that weak horizontal gradients in many physical, chemical and biological characteristics of the NPSG would allow observations at this site to scale to the larger region (Karl, 1999). The North Pacific Subtropical Gyre (NPSG) has a total area of approximately 2 x 10^7 km^2, a mean depth of nearly 5 km and is considered the largest contiguous biome on Earth. Analogous subtropical gyres are located in the South Pacific Ocean, Indian Ocean and both hemispheres in the Atlantic Ocean and together make up approximately 40% of the surface of our planet (Karl and Lukas, 1996; Karl and Church, 2014).
The habitat in the NPSG is characterized by warm surface waters, which result in year-round water column stratification, isolating the upper water column from colder waters below. Organisms in the perennially stratified surface waters experience excess solar energy but nutrient deplete conditions (Karl and Church 2014). Organisms in the dark waters below experience order-of-magnitude-higher inorganic nutrient concentrations but rely on inconsistent sinking organic matter and chemoautotrophy for energy (Karl and Church 2014). The deep chlorophyll maximum layer (DCM), a feature of most of the world's tropical and subtropical oceans, is believed to be the depth at the base of the euphotic zone where phytoplankton are able to access upward diffusing inorganic nutrients while still receiving sufficient down-welling solar radiation for primary production (Cullen, 2015; Letelier et al., 2004). At station ALOHA, the DCM it is tightly coupled with the top of the nitricline, which is the depth at which nitrate concentrations begin to rise with increasing ocean depths (Letelier et al., 2004).

The perennially stratified waters but dramatic transitions in light, nutrients and other biogeochemical properties create a microbial habitat at Station ALOHA with limited seasonal variability, but steep vertical gradients. Microbial communities have been observed to vary across this gradient in the NPSG, but the nature of the transition from shallow to deep surface water communities and potential stratification within the euphotic and aphotic zones however is not well understood (DeLong et al., 2006). In addition, despite the initial belief that the NPSG provides a relatively constant habitat, evidence continues to mount demonstrating that episodic events such as mesoscale eddies that inject nutrients into surface waters, result in a dynamic environment in the NPSG (Karl and Lukas, 1996; Sakamoto et al., 2004; Fong et al., 2008; Karl and Church, 2014). However the relative importance of episodic versus seasonal variability in shaping microbial communities is not well understood.

The majority of marine microbial genomic or metagenomic studies have focused on planktonic organisms. It was recognized early on however, that microbes partition across two distinct niches, broadly defined as free-living or planktonic and particle or surface associated (DeLong et al., 1993; Dang and Lovell, 2016; Allen et
Generally particle-associated communities are operationally defined as cells attached to material that is retained on filters with pore sizes greater than 1.2μm, while the free-living cells passes through such filters. Communities appear to be more strongly shaped by these structural habitat distinctions than environmental factors. For example, at the Pivers Island Coastal Observatory, communities from the same particle size fraction, but sampled on different dates, were more similar to each other than to communities of different size fractions that were sampled on the same day (Yung et al., 2016). Similar observations have been made when comparing particulate and free-living communities across different ocean depths in the subtropical North Pacific Gyre, different redox conditions in an oxygen minimum zone off the coast of Chile and different regions of a river outlet in the Columbia River Basin (Smith et al., 2013; Ganesh et al., 2014, 2015; Fontanez et al., 2015). The extent to which neighboring free-living and particle-associated communities differs in composition widely ranges and likely depends on physical factors including particle compositions and densities, water chemistry and turbulence and biological factors including the proportion of generalists and specialists in a community (Crump et al., 1999; Hollibaugh et al., 2000; Dang and Lovell, 2016).

In surface waters at Station ALOHA and similar oligotrophic habitats, Prochlorococcus and Pelagibacter dominate the free-living picoplanktonic communities (Campbell and Vaulot, 1994; Campbell et al., 1997; Morris et al., 2002; Eiler et al., 2009). A high diversity of other microbial clades including Synechococcus, SAR116, SAR86, SAR406 and Roseobacter are also present and believed to be active at Station ALOHA and other oligotrophic sites (Campbell et al., 1997; Aylward et al., 2015; Giovannoni and Vergin, 2012). Clades within the Rhodobacteraceae, Bacteroidetes and Gammaproteoacteria groups however are frequently surface-associated and generally considered copiotrophic compared to several of the planktonic clades (Dang and Lovell, 2016). Such copiotrophs may be at low abundances in picoplankton communities and bloom when environmental conditions become more favorable (Vergin et al., 2013; Lynch and Neufeld, 2015; Gilbert et al., 2012; Yooseph et al., 2010).
One of the goals set forth when establishing station ALOHA as a long-term study site was to better understand the link between human activity and the NPSG ecosystem (Karl and Lukas, 1996). We are now considered to be in the Anthropocene, a time-period where recent human population growth and demand for energy and natural resources is significantly shaping global physical, chemical and biological cycles (Steffen et al., 2007). Impacts on the NPSG include increasing water temperatures and nitrogen deposition, ocean acidification and expansion of oligotrophic boundaries (Doney et al., 2012; Kim et al., 2014; Polovina et al., 2008). In addition, the accumulation of plastic, a new microbial particulate niche in the marine environment, is considered a geologic indicator of the Anthropocene (Waters et al., 2016).

Scientists first reported floating plastic debris in open ocean surface waters in 1972 (Carpenter and Smith, 1972). Plastic debris has since become recognized as a persistent marine pollutant, which upon entering the ocean, fragments into small pieces coined ‘microplastics’ that become concentrated in gyres. These plastic particles have become a feature of the NPSG surface waters with several survey studies demonstrating that, microplastics can reach concentrations between $10^4$ to $10^7$ pieces per km$^2$ (Law et al., 2014; Eriksen et al., 2014). These values will likely increase as the quantity of mismanaged plastic waste likely to enter the oceans is predicted to increase by an order of magnitude by from 2010 levels by 2025 (Jambeck et al., 2015).

Plastic debris is known to impact marine organisms, including turtles, birds, mammals, fish, and invertebrates through entanglement and ingestion (Laist, 1997; Wilcox et al., 2013; Derraik, 2002). Additionally, a number of studies have clearly demonstrated that diverse biofouling organisms, such as bryozoans, settle on marine plastic debris (Winston, 1982; Goldstein et al., 2014; Reisser et al., 2014). In this regard, plastic can serve as a vector for the introduction of nonnative species into new environments (Barnes, 2002; Masó et al., 2003). Despite known impacts of plastic on higher organisms, much less is known about the interactions between marine microbiota and plastic (Oberbeckmann et al., 2015). Colonization of plastic particles by microbes was also first reported in 1972 (Carpenter et al., 1972).
Subsequent studies have shown that microbes can rapidly colonize debris and that in the Atlantic Ocean, communities on plastic are taxonomically distinct from those in the surrounding water column (Dang et al., 2008; Lobelle and Cunliffe, 2011; Zettler et al., 2013; Oberbeckmann et al., 2014; Harrison et al., 2014). However, at the time of starting research for this thesis, microbial community composition on floating plastic in the NPSG had never been characterized. More significantly, the potential for functional differences between microbes found on plastics and those in the surrounding water column had yet to be explored.

Overview of Work Presented in this Thesis

The overarching goal of this thesis is to improve our understanding of the distribution and drivers of microbial communities in the North Pacific Subtropical Gyre. To this end, we used metagenomic survey data to investigate both free-living and surface-associated communities in the NPSG across space and time. First we investigated whether temporal variability in diversity and composition of naturally occurring, free-living picoplankton communities at 25 m and 500m depths demonstrate seasonal or episodic trends that correlated with variability in biogeochemical environmental measurements (Chapter 2). Next we explored picoplankton community changes across the depth gradient from surface waters through the mesopelagic at a higher spatial resolution (Chapter 3). The dramatic environmental gradient over short distances also allowed us to investigate how different genomic properties (e.g. GC content, coding density) may be advantageous for picoplankton in euphotic low nutrient surface waters compared to those in aphotic high nutrient waters deeper in the water column. Third, we described the complex communities observed on buoyant plastic debris, a new and largely unexplored habitat in ocean gyres of anthropogenic origin (Chapter 4). To uncover potential assembly processes that structured the rich communities observed on plastic debris, we conducted experiments to explore colonization of plastic films in a
flowing seawater tank over a two-week period (Chapter 5). Taken together, this work contributes to an improved understanding of natural microbial assemblages in the NPSG and identifies directions for future research.

References

14. Derraik JGB. (2002). The pollution of the marine environment by plastic


Chapter 2. Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre

This work has been published with the following co-authors in the manuscript:

Bryant JA, Aylward FO$^{1,2}$, Eppley JM, Karl DM$^{1,2}$, Church MJ$^{1,2}$, DeLong EF$^{1,2}$. 2016. Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre. ISME J 10:1308-1322.

Affiliations:

$^1$Daniel K. Inouye Center for Microbial Oceanography, Research and Education, University of Hawaii, Manoa, Honolulu, HI, USA
$^2$Department of Oceanography, University of Hawaii, Manoa, University of Hawaii, Honolulu, HI, USA
Abstract

Few microbial time-series studies have been conducted in open ocean habitats having low seasonal variability such as the North Pacific Subtropical Gyre (NPSG), where surface waters experience comparatively mild seasonal variation. To better describe microbial seasonal variability in this habitat, we analyzed rRNA amplicon and shotgun metagenomic data over two years at the Hawaii Ocean Time-series Station ALOHA. We postulated that this relatively stable habitat might reveal different environmental factors that influence planktonic microbial community diversity than those previously observed in more seasonally dynamic habitats. Unexpectedly, the data showed that microbial diversity at 25 m was positively correlated with average wind speed 3 to 10 days prior to sampling. In addition, microbial community composition at 25 m exhibited significant correlations with solar irradiance. Many bacterial groups whose relative abundances varied with solar radiation corresponded to taxa known to exhibit strong seasonality in other oceanic regions. Network co-correlation analysis of 25 m communities showed seasonal transitions in composition, and distinct successional cohorts of co-occurring phylogenetic groups. Similar network analyses of metagenomic data also indicated distinct seasonality in genes originating from cyanophage, and several bacterial clades including SAR116 and SAR324. At 500 m, microbial community diversity and composition did not vary significantly with any measured environmental parameters. The minimal seasonal variability at 25 m in the NPSG facilitated detection of more subtle environmental influences, such as episodic wind variation, on surface water microbial diversity. Community composition in NPSG surface waters varied in response to solar irradiance, but less dramatically than reported in other ocean provinces.
Introduction

Microbial community structure and function have pivotal roles in the biogeochemical dynamics of marine ecosystems, yet the microbial ocean remains largely undersampled. Coordinated time-series studies are a key strategy for addressing this undersampling, and improve understanding of the complex interplay between environmental variability and microbial community diversity and dynamics. Several recent time-series efforts focusing on marine surface waters have observed dramatic seasonality in microbial communities, including studies in the Western English Channel (Gilbert et al., 2012), the Sargasso Sea (Morris et al., 2005; Treusch et al., 2009), coastal waters near southern California (Fuhrman et al., 2006) and coastal waters in Antarctica (Murray et al., 1998). Seasonal variability at these locations has been attributed to changes in the physical habitat, including solar irradiance, stratification and mixing. For example, Gilbert et al. (2012), observed dramatic shifts in microbial richness and community composition in the English Channel that correlated with changing day lengths that vary by as much as 8 h between seasons. Clear seasonal patterns in community composition were also observed at the oligotrophic Bermuda Atlantic Time-series Study (BATS,(Treusch et al., 2009)), where fluctuations in microbial populations varied with the annual cycle of deep convective mixing in the winter, a predictable spring bloom and late summer/early autumn stratification of the upper ocean (Giovannoni and Vergin, 2012).

Compared with other oceanic regions, the physicochemical environment of the North Pacific Subtropical Gyre (NPSG) exhibits relatively low seasonality (Bingham and Lukas, 1996). For example, at the Hawaii Ocean Time-series (HOT) Station ALOHA, a well-studied site representative of the NPSG, there is only a 3.09 h time difference between the longest and shortest days of the year, and sea surface water temperatures vary <4 °C annually (HOT Data Organization and Graphical Systems (DOGS)-see Methods). Additionally, predominantly stratified surface waters create oligotrophic conditions at Station ALOHA year round, unlike the more seasonally oligotrophic waters at BATS. Currently, it is unknown whether the milder
climatic and hydrographic seasonal variability at Station ALOHA results in
differences in microbial seasonality compared with other oceanic regions. Since the
NPSG represents the largest circulation feature on Earth and substantially impacts
major global biogeochemical cycles, better understanding its biological dynamics
remains an important endeavor (Karl and Lukas, 1996; Karl and Church, 2014).

To investigate the potential seasonality in microbial dynamics at Station
ALOHA and identify possible physical and biogeochemical drivers, we examined
changes in microbial communities at two discrete depths, 25 m and 500 m, for near-
monthly time intervals over a 2-year period. We used bacterial small subunit SSU
ribosomal RNA (SSU rRNA) amplicon and shotgun metagenomic sequences to follow
changes in microbial taxonomic and functional gene diversity and representation.
Amplicon sequencing was used to identify differences between microbial
communities by comparing bacterial small subunit ribosomal RNA gene sequences
within and between samples directly. Metagenomic shotgun sequencing was used to
capture genes from a broader array of cells from all domains, as well as their
viruses, and provide broader insight into microbial community composition and
variability. Two fundamental dimensions of biodiversity were investigated; alpha
diversity, defined as the diversity within individual time points, and beta diversity,
defined as the dissimilarity in community composition between pairs of time points.
We also used a weighted co-occurrence network analyses to identify clusters of co-
varying organisms and protein-coding genes in our samples. We postulated that
microbial community dynamics analyzed using these diversity metrics and
analytical approaches would reveal clear but potentially muted seasonal trends via
correlations with biotic and abiotic seasonal changes. We further hypothesized that
the comparatively low seasonal variability in the NPSG might reveal the influence of
different, potentially more subtle environmental factors on microbial diversity that
have not been reported in previous studies.

Materials and Methods
Site and sample collection

The HOT program has been conducting research cruises at approximately monthly intervals at Station ALOHA (22°45'N, 158°00'W) to make physical, chemical and biotic observations since 1988 (Karl and Lukas, 1996). Water samples for the current study were collected on HOT cruises between August 2007 and September 2009 (HOT cruises #194–215). Sampling dates and times are listed in Table S2.1 and S2.2. Corresponding HOT environmental observations were downloaded from the HOT-DOGS website http://hahana.soest.hawaii.edu/hot/hot-dogs/. Environmental measurement protocols are available on the HOT-DOGS site. Environmental measurements were collected during the same cruise and where possible the same day as microbial sampling. HOT program measurements sampled from between 20 and 30 m were utilized for contextual information on microbial community dynamics occurring at 25 m, with the exception of silicate, which was measured between 5 and 25 m. Temperature and salinity measurements collected between 495 and 505 m were averaged for 500 m analyses. Other 500 m environmental measurements (for example, nutrient and dissolved oxygen concentrations) sampled between 470 and 530 m were utilized for the analyses of microbial community structure at 500 m. Pigments were measured using high-performance liquid chromatography and total picoplankton cell numbers were estimated using epi-fluorescence microscopy. The depth of the deep chlorophyll maximum was identified by visual inspection of water column fluorescence data. Mixed layer depth values were based on a 0.125-unit potential density criterion.

In addition, we downloaded daily mean sea level height anomaly data from the Integrated Climate Data Center (http://icdc.zmaw.de/ssh_aviso.html?&L=1). We obtained wind velocity and solar radiation data (incident light energy from 0.28 to 2.8 μm wavelengths in watts m⁻²) collected by the Upper Oceans Processes Group at the Woods Hole Oceanographic Institution with the WHOTS buoy located at Station ALOHA (retrieved from http://uop.whoi.edu/projects/WHOTS/whots.html). For comparisons between wind velocity and surface chlorophyll concentrations across a longer time scale (1989–2009), wind velocity measurements were retrieved from
the NOAA National Data Buoy #51001 located 450 km away from Station ALOHA (retrieved from http://www.ndbc.noaa.gov/station_page.php?station=51001). Theoretical hours of daylight were calculated using the Lammi’s Online-Photoperiod Calculator V1.94L (http://www.sci.fi/~benefon/sol.html).

Microbial cells were sampled during HOT cruises by filtering 20 L of seawater, collected with a CTD rosette sampler, through an inline 47 mm diameter, 1.6 μm pore-size GF/A pre-filter (Whatman, Piscataway, NJ, USA) followed by collection on a 0.22 μm pore-size Sterivex GV filter (Millipore, Billerica, MA, USA) using a peristaltic pump. Immediately after filtering was completed, 2 ml of sterile DNA storage buffer (50 mm Tris-HCl, 40 mm EDTA and 0.75 M sucrose) was added to the Sterivex cartridges, and the filters were flash frozen in liquid nitrogen and stored at −80 °C until DNA extraction.

DNA extraction and sequencing

Cells were lysed directly in Sterivex filter units and DNA in the crude lysate was purified on a Quick-Gene 6101 system (Fujifilm, Tokyo, Japan) using DNA Tissue Kit L (Autogen, Holliston, MA, USA). Modifications made to the manufacturer’s cell lysis protocol are described by (Sharma et al., 2013). Shotgun pyrosequencing was performed using either FLX or Titanium series chemistry (Table S2.1 and S2.2) on a Roche Genome Sequencer FLX instrument according to manufacturer’s recommendations (Roche, Indianapolis, IN, USA). FLX or Titanium Rapid Library Preparation protocols were used for library construction. Libraries were quantified using the Titanium Slingshot kit (Fluidigm, San Francisco, CA, USA) and added to emulsion PCR reactions at 0.1 molecules per bead. Bacterial amplicon libraries targeting the V1-V3 region of bacterial SSU rRNA genes were generated with 27F (5’ AGAGTTTGATCCTGGCTCAG 3’) and 534R (5’ ATTACCGCGGCTGCTGG 3’) primers using the PCR amplification protocol established for the Human Microbiome Project (Jumpstart Consortium Human Microbiome Project Data Generation Working Group, 2012). To increase yield while keeping the number of amplification cycles low, triplicate PCR reactions using 20
amplification cycles were run for each sample then pooled. Amplicons were purified using the QiaQuick PCR Clean-Up kit (Qiagen, Valencia, CA, USA) and their size was verified with agarose gel electrophoresis. Following this step, equal quantities of all PCR reactions from the same ocean depth were pooled together, and sequenced using a 454 Genome Sequencer. Sequencing of amplicons was carried out using the Titanium Rapid Library Preparation protocol. We empirically determined the optimal conditions for library preparation since the amplicon DNA fragments were shorter than the fragment length targeted by the library preparation kit. The manufacturer's protocol was followed except adaptor-ligated libraries were not diluted before size selection with AMPure XP beads and 1/4 of the recommended volume of amplification primers was used in emulsion PCR reactions. All metagenomic and amplicon sequencing data is available in the NCBI SRA database (accession numbers in Table S2.1, S2.1 and S2.3).

**Metagenomic sequence analysis and annotation**

Before analysis of sequencing data, duplicate identical DNA sequences, which were likely 454 sequencing protocol artifacts were removed from metagenomic data sets using previously described computational methods (Stewart et al., 2010). BLASTX searches were conducted against an in-house database comprised of NCBI RefSeq plant and microbe peptide databases (release 51) combined with peptide sequences from marine taxa whose genomes were sequenced using single-cell sequencing (Swan et al., 2011b, 2013). Reads were assigned to the taxonomy of their best match, provided the match had a bit score of 50 or greater. Reads matching multiple hits equally well were assigned to the lowest common ancestor of all equally scored top hits.

SSU rRNA genes were identified in the metagenomic data using BLASTN searches against the ARB-Silva non-redundant SSU rRNA reference database with a minimum bit score cutoff of 50 ((Pruesse et al., 2007), release 102). Reads identified as SSU genes were assigned to their top hit in the database and corresponding leaf on the ARB-Silva non-redundant SSU reference guide phylogeny. By using the ARB
software program, the reference phylogeny was then pruned to only include sequences that matched SSU sequences in our metagenomic libraries (Ludwig et al., 2004). This phylogeny was used for subsequent phylogenetic analyses.

Metagenomic reads were also binned into de novo protein-coding ‘functional gene’ clusters by first using the gene-finding program MetaGene to identify amino acid sequences within reads (Noguchi et al., 2006). Amino acid sequences originating from the same ocean depth were then pooled, clustered first to a 90% identity threshold, followed by 60% identity threshold, both with a 70% minimum overlap, using the program cd-hit (Li and Godzik, 2006).

**Amplicon sequence analysis and annotation**

Amplicons were analyzed within the software package QIIME (Caporaso et al., 2010b). First PCR and 454 sequencing artifacts were removed using the AmpliconNoise and Perseus algorithms (Quince et al., 2011). Next, de-multiplexed sequences were binned into de novo operational taxonomic units (OTUs) at 97% identity using the UCLUST algorithm (Edgar, 2010). Reference sequences from each OTU were aligned using PyNAST, uninformative base positions based on the default lane mask were removed and a phylogeny was constructed using the FastTree algorithm (Price et al., 2009; Caporaso et al., 2010a). This phylogeny, with OTUs as leaves was used to calculate subsequent amplicon diversity metrics.

Aggressive taxonomic assignments were made outside of the QIIME package, by comparing the OTU reference sequences to the ARB-Silva SSU database using BLASTN (release 108). Reads were assigned to the lowest common ancestor of all database hits with a score within 5% of the top score, provided that bit scores were 50 or higher and the database sequences spanned at least 95% of the amplicon sequence. On occasion, all but a small number of the top hits were to the same taxonomic assignment and the few incongruous hits had low pintail values indicating a high probability the hits were chimeras. In such cases the taxonomy of the majority was assigned to the reference sequence.
Statistical analyses

We estimated alpha diversity with SSU rRNA genes from both amplicon and metagenomic samples, using the metric termed phylogenetic diversity (PD), which is similar to taxonomic richness, but incorporates the phylogenetic relatedness of organisms (Faith, 1992). We estimated beta diversity within SSU rRNA genes using the UniFrac Metric, which is similar to the Jaccard Index that quantifies dissimilarity in taxonomic composition between pairs of samples but unifrac also incorporates the phylogenetic relatedness of organisms (Lozupone and Knight, 2005). We used the protein-coding gene clusters to calculate functional alpha diversity with richness, and to calculate functional beta diversity with the Jaccard Index, parallel to our rRNA phylogenetic metrics PD and Unifrac. Metrics described thus far all depend on the presence, but not abundance of the organisms in the samples. To investigate the additional influence of varying taxon abundances, we also calculated parallel abundance-based metrics (see Chapter 2 Supplementary Material).

Alpha diversity within the amplicon data and SSU reads extracted from the metagenomic libraries were calculated in R using the vegan and picante packages (Faith, 1992; Allen et al., 2009; Kembel et al., 2010; Oksanen et al., 2015). PD rarefaction curves were generated in QIIME, as were Unifrac distances (Caporaso et al., 2010b). Functional alpha and beta diversity measures were calculated using Vegan and the Python package SciPy (Jones et al., 2001; Oksanen et al., 2015). To account for differing sequencing depths between samples, rarefaction re-sampling was conducted for all indices by averaging the diversity values generated from 100 random subsamples of each community. Communities were subsampled to the read depth of the smallest of the samples being compared. See figure captions for subsampling levels used for each analysis. Spearman’s correlation coefficients and corresponding two-sided P-values were calculated between alpha diversity and environmental measurements using the R function cor.test. Mantel tests using Spearman’s coefficient implemented in the vegan package were used to test the significance of correlations between beta diversity measures and Euclidean
distances between environmental measurements. \( P \)-values were adjusted to account for multiple tests using the Benjamini–Hochberg (BH) procedure.

Because of cruise schedules and the nature of solar radiation at HOT, most of our microbial sampling took place when Station ALOHA was experiencing annual extremes in incoming solar irradiance. Therefore to identify microorganisms whose relative abundances may change with variation in incoming solar irradiance, we divided samples into high- and low-light samples. Samples collected in October 2007, February 2008, October 2008 and April 2009 (cruises 196, 200, 205 and 210) that did not fall during annual extremes in incoming shortwave solar radiation were removed.

A DESeq2 enrichment analysis was used to test the null hypothesis that the Log2 fold change in the number reads mapping to a given taxonomic group or OTU between high- and low-light times of the year is zero (Love et al., 2014). A detailed description is available elsewhere (Love et al., 2014), but in brief, DESeq2, using count data, tests for differential abundances using negative binomial generalized linear models and estimates size and dispersion factors to control for differences in sequencing depth between libraries and dispersion between taxa. \( P \)-values are adjusted to account for multiple tests using the BH procedure after an independent filtering criterion is applied to remove tests that have little chance of showing significance. We included taxa that DESeq2 flagged as potential outliers, because they appeared biologically reasonable. To test for differential abundance in clades within the amplicon data, we collapsed OTU counts into clade counts based on Arb-Silva taxonomic assignments (see above). We choose clades that corresponded to roughly the family level and had relative abundances that summed >1% across all samples. The cladogram in Figure 2.5 displaying these results was built using the program GraPhlAn (https://bitbucket.org/nsegata/graphlan/).

**Weighted co-occurrence network analyses**

The R package WGCNA (Langfelder and Horvath, 2008) was used for weighted co-occurrence network analyses of both SSU amplicon and community
metagenomic data. For the 25 and 500 m SSU amplicon data sets, count matrices of OTUs were constructed whereby counts represented the number of sequences identified as belonging to a particular OTU with the total number of counts in each sample used for normalization. Low abundance OTUs (<50 counts across all time points or 0 values for at least 15 time points) were excluded on the grounds that they were not likely to yield robust patterns. To corroborate seasonal patterns identified in the 25 m OTU data, count matrices were also constructed for the 25 m community metagenomic data by mapping sequence reads onto ortholog clusters using methods described previously (Aylward et al., 2015). Briefly, metagenomic reads were first mapped against RefSeq v. 62 using LAST (Frith et al., 2010) with parameters ‘-b 1 -x 15 -y 7 -z 25 -e 80 -F 15 -u 2 -Q 0’, with bit scores calculated afterwards and only hits having scores > 50 retained. Ortholog clusters were constructed for phylogenetic groups of interest (Prochlorococcus, SAR11, SAR116, SAR86, SAR324, SAR406, Roseobacter, pelagiphage and cyanophage) with ProteinOrtho (Lechner et al., 2011) using select genomes that had a high number of reads mapping across metagenomes. For annotations a representative protein was selected from each ortholog cluster (the longest protein in the cluster, or a randomly selected protein in case of ties) and queried against the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000) using LAST (default parameters, bit score cutoff of 50). Count tables in which columns represented metagenomes and rows represented ortholog clusters were then constructed, in which counts were given based on the number of reads mapping to all of the proteins comprising each of the ortholog clusters. Ortholog clusters having <50 total reads mapping were excluded from subsequent analysis. Ortholog cluster count tables were normalized using the DESeq function ‘varianceStabilizingTransformation’.

Pairwise Pearson correlations were calculated for all OTUs or ortholog clusters in the count tables and a matrix of pairwise adjacency scores was subsequently generated using the equation:

\[ a_{ij} = \left| s_{ij}^{\beta} \right| \]
where $a$ is the adjacency, $s$ is the Pearson correlation, $i$ and $j$ are the rows and columns of the correlation and adjacency matrix, respectively, and $\beta$ is the soft threshold calculated using the scale-free topology index with the guideline that a soft threshold yielding an $R^2 > 0.8$ for scale-free topology fit is suitable (Zhang and Horvath, 2005). The 'blockwiseModules' command in WGCNA was used to identify subgroups of OTUs (termed 'modules') with co-varying temporal profiles (minimum group size $>5$ specified for OTUs, $30$ for ortholog clusters). Clustering was performed using the average linkage hierarchical clustering algorithm, and the Dynamic Tree Cut package (Langfelder and Horvath, 2008) was used to determine module delineations. The first principle components ('eigengenes') of modules were calculated using the 'moduleEigengenes' command in WGCNA, with default parameters. OTUs or ortholog clusters were classified as 'unassigned' if their correlation to the eigengene of the module to which they were assigned was $<0.3$. Modules having eigengene dissimilarities $<0.25$ were subsequently merged. Because these methods do not distinguish between positive and negative correlations, we separated modules post facto by calculating the Pearson correlation of each OTU or ortholog cluster's temporal profile with the eigengene of the subgroup to which it belonged; positively and negatively correlated OTUs or ortholog clusters were then denoted with 'pos' or 'neg'.

To further examine co-correlation patterns and corroborate weighted network analysis findings, we also employed a more commonly used unweighted network analysis to examine seasonal patterns in the 25 m data set. In this approach, significantly correlated OTUs were identified independently of weighted co-occurrence analyses using the 'CorAndPvalue' function in the WGCNA package, which uses Student's $t$-test to calculate $P$-values from pairwise Pearson values while accounting for the total number of observations. OTUs found to be significantly correlated (BH adjusted $P<0.05$) were then used to construct unweighted networks in which edges were drawn between OTUs with significantly correlated or anticorrelated temporal profiles (Figure 2.6). Network construction was performed using the R package igraph (Csardi and Nepusz, 2006).
Results and Discussion

Microbial plankton samples were collected at roughly monthly intervals from August 2007 through September 2009, at 25 and 500 m depths during HOT program cruises to Station ALOHA in the NPSG. As is typical for this region, we observed a persistently stratified upper ocean and only mild seasonal variation in surface waters (Figure 2.1). At the ocean’s surface, mixing depths ranged from 13 to 111 m and water temperatures and bacterial cell numbers tended to peak in the late summer and fall (Figure 2.1 and Supplementary Figure S2.1). In contrast, there was little evidence of seasonality in physical or biogeochemical dynamics at 500 m, with conditions being relatively homogenous throughout the year, relative to near-surface waters (Figure 2.1). For example, maximum and minimum temperature values differed by 3.8 versus 1.4 °C, salinity values differed by 0.52 versus 0.08 (Practical Salinity Scale 1978) and dissolved organic carbon differed by 12.0 versus 7.4 μmol kg⁻¹ at 25 and 500 m, respectively. In addition, nitrate+nitrite and phosphate concentrations were all at least an order of magnitude higher at 500 m compared with 25 m.

We characterized microbial samples using both shotgun metagenomic pyrosequencing of whole community DNA (hereafter referred to as metagenomic data sets or samples), and bacterial SSU rDNA amplicon sequencing. Amplicon-based methods are the most commonly used method for studying microbial diversity. However, shotgun metagenomic approaches are becoming more common as sequencing costs decrease (Logares et al., 2014; Bryant et al., 2012). Comparing functional gene (protein-coding gene) diversity with SSU rRNA-based measures is also useful and relevant, since gene composition ultimately dictates how microbes can potentially interact with each other and their environment (Gilbert et al., 2010; Raes et al., 2011; Barberán et al., 2012; Fierer et al., 2012). For each of the 21 discrete time points and two depths studied, we generated metagenomic libraries with an average of 1.1 million reads of an average length of 360 base pairs (bps) (Supplementary Table S2.1). We generated 21 amplicon libraries at 25 m and 19 at
500 m, each with an average of 12,000 reads of ~350 bps in length (Supplementary Table S2.2).

As anticipated, *Prochlorococcus* and SAR11, the dominant microorganisms in surface waters year round at Station ALOHA, comprised roughly 50% and 10% of metagenomic reads across all samples, respectively (Supplementary Figure S2.2). This was consistent with the corresponding amplicon data, as well as with previous reports in this same oceanic region (Supplementary Figure S2.3; Schmidt *et al.*, 1991; Campbell and Vaulot, 1993; Eiler *et al.*, 2011). At 500 m, SAR11, SAR324, SAR406, *Nitrospina*, SAR202 and *Thaumarchaeota* were consistently the most abundant taxa (Supplementary Figures S2.4 and S2.5).

We compared SSU rRNA alpha and beta diversity metrics generated by the two different sequencing approaches (SSU rRNA amplicons versus shotgun metagenomic sequencing) and also compared diversity metrics generated from amplicon SSU rRNA to diversity metrics generated from metagenomic functional genes to investigate whether the different methods were consistent with one another (Chapter 2 Supplementary Material and Supplementary Figures S2.6-S2.9). At 25 m, all the diversity metrics were generally well correlated. In contrast, diversity metrics at 500 m were not consistently correlated. We propose that lower correspondence between amplicon versus metagenomic data sets in 500 m samples likely reflects lower variability between samples at 500 m (see next section).

Detailed results and additional discussion are available in the chapter 2 supplementary material.

**Microbial communities at 25 versus 500 m in the NPSG**

An emerging pattern in pelagic systems is an increase in alpha diversity at intermediate water depths compared with surface waters (Treusch *et al.*, 2009; Jing *et al.*, 2013; Sunagawa *et al.*, 2015). Discrepancies in this richness pattern can usually be explained by the taxonomic resolution of the study (Brown *et al.*, 2009; Kembel *et al.*, 2011). Consistent with these findings, we observed a broad array of dominant taxa at 500 m, which yielded higher phylogenetic and functional richness
compared with that found at 25 m (Figure 2.2, Supplementary Figures S2.10 and S2.11). Explanations for higher richness in the mesopelagic are not immediately obvious, since productivity is highest in the photic zone. The distance from sunlight-driven energy and productivity, however, likely causes intense resource competition at depth. This may promote a more diverse and even community containing a broader range of heterotrophic and chemolithotrophic lifestyles and taxa compared with shallower waters. Evidence for more prevalent chemolithotrophic associated activities, including ammonia oxidation, sulfur oxidation and CO$_2$ fixation have all been reported in the mesopelagic at the NPSG (Karl et al., 1984; Ingalls et al., 2006; Hansman et al., 2009; Swan et al., 2011a; Giovannoni and Vergin, 2012).

Previous studies have demonstrated that in the NPSG, steep physical and chemical gradients and perennial stratification are reflected in strong vertical structure in microbial community composition (DeLong et al., 2006). Moreover, previous studies have observed that microbial communities in the near-surface ocean tend to be more variable in time compared with communities in deeper water, likely because the physical environment below the euphotic zone is more stable (for example, (Lee and Fuhrman, 1991; Treusch et al., 2009; Rich et al., 2011; Ghiglione et al., 2012). Results derived from our metagenomic and amplicon data were consistent with these previously described depth-dependent trends.

When combining all the amplicon data from both 25 and 500 m together, we only observed five OTUs with relative abundances >0.1% at both depths studied (Figure 2.2). These OTUs mapped to SAR11 and SAR406. Other shared OTUs that were abundant at one depth were at least an order of magnitude less abundant at the other depth.

Variance between time points was significantly higher among 25 m samples than among 500 m samples, for the majority of alpha and beta diversity metrics including Unifrac (Levene’s test $P<0.05$, Figure 2.2 and Supplementary Figure S2.10). In addition, we observed that microbial communities at 500 m had lower average beta diversity indices compared with 25 m, indicating that the composition of microbial communities was more similar across time points at 500 m (Figure 2.2,
and Supplementary Figure S2.10). However, when incorporating abundance information into Unifrac values (weighted-Unifrac, (Lozupone et al., 2007)), values were higher at 500 m compared with 25 m. This is likely because two groups, Prochlorococcus and Pelagibacter, dominated all the 25 m time points.

At 500 m, no consistent correlations between microbial alpha and beta diversity metrics and environmental measurements were found (BH adjusted P<0.10). Also weighted gene co-occurrence network analyses revealed that only a minority of 500 m OTUs could be clustered into modules and these modules showed no significant correlations with environmental parameters (Supplementary Figure S2.16). Mesopelagic communities may of course have responded to environmental variables we did not measure. The minimal variability in 500 m samples also suggests that at this depth the temporal frequency we sampled was not optimal for detecting variation in microbial communities. Our 500 m results are consistent with previous work at Station ALOHA, and at global scales, that showed that in the bathypelagic, individual environmental variables appear to have only a small effect on free-living communities (Eiler et al., 2011; Salazar et al., 2016). Microbes in the mesopelagic have been shown to be very active on particulate matter, which is rapidly transported from surface water to the ocean floor (Karl et al., 1984, 2012). Including particulate associated microbial communities in mesopelagic studies, as well as incorporating longer sampling periods might reveal more variability and perhaps seasonality in mesopelagic microbial communities than we could observe with this data set.

**Wind correlates with alpha diversity in NPSG surface waters**

To reveal potential environmental factors influencing variability in alpha diversity across our 25 m time series, we looked for correlates between the alpha diversity measures and the physical, chemical and biotic environmental parameters measured during HOT cruises and at the nearby WHOTS meteorological buoy. Unlike studies in other regions of the ocean, we did not observe significant correlations between alpha diversity and seasonally driven environmental
parameters such as temperature and mixed layer depth (Gilbert et al., 2012; Ladau et al., 2013). Instead all alpha diversity metrics were consistently most strongly correlated with the average wind speed of the days leading up to sample collection (Figure 2.3, Supplementary Table S2.4). The correlation between wind and alpha diversity was statistically significant for amplicon phylogenetic diversity (PD) and functional richness (Spearman \( r=0.7 \), BH adjusted \( P<0.10 \)). This trend held when averaging the wind speeds from 3 to 10 days before the sampling date (Supplementary Figure S2.12). The alpha diversity metric phylogenetic entropy, which incorporates taxon abundance information, is less influenced by rare taxa than PD. The lower correlation between wind and phylogenetic entropy (Spearman \( r<0.50 \)) combined with an OTU abundance distribution demonstrates that low abundance taxa largely contributed to the wind-associated increase in alpha diversity (Supplementary Table S2.4, Supplementary Figure S2.13).

Chlorophyll a concentrations (chl a) also correlated with amplicon PD (BH adjusted \( P<0.10 \)) and chl a was reciprocally correlated with average wind speeds (Spearman \( r=0.46, P<0.05 \), Figure 2.3). The relationship between wind speed and chlorophyll concentrations at HOT appears to be a long-term trend, consistent with findings from a global-scale study (Kahru et al., 2010). We compared 20 years of HOT near-surface ocean chl a measurements from 1989 to 2009 to wind speed measurements collected from the NOAA Nation Data Buoy #51001. Chl a was significantly correlated with wind speeds over the 20-year period (Spearman \( r=0.26, P<0.001 \)). Surface chlorophyll concentrations are also impacted by phytoplankton photoadaption, where per cell pigment concentrations rapidly adjust to current light levels (Olman et al., 1982; Letelier et al., 1993; Winn et al., 1995). The wind-chl a relationship across the 20-year period remained significant after statistically accounting for phytoplankton photoadaption by modeling chl a and day length (Supplementary Figure S2.14, linear regression \( P<0.001 \), \( r^2=0.13 \)).

Strong wind-driven mixing events and wind/eddy interactions can spur phytoplankton blooms in the open ocean (Winn et al., 1992; Letelier et al., 2000; McGillicuddy et al., 2007). Wind may also be fertilizing surface waters by
transporting nutrients including fixed nitrogen, phosphorus and iron through both wet and dry dust deposition (DiTullio and Laws, 1991; Young et al., 1991; Karl and Tien, 1997; Jickells et al., 2005; Fitzsimmons et al., 2014). We did not observe significant increases in autotrophic biomass or cyanobacteria cell concentrations with increasing wind speeds (Supplementary Figure S2.14). These data likely reflect a wind-driven increase in chlorophyll per cell, as opposed to increased cell numbers, suggesting that alpha diversity is not a product of wind-driven increases in oxygenic photoautotroph abundance.

Photoautotrophs deeper in the water column have higher per cell chlorophyll concentrations to adjust for lower light levels (Winn et al., 1995). Thus the increases in microbial alpha diversity and chlorophyll may also result from entrainment of microorganisms located deeper in the water column via wind-driven mixing. In addition, wind-driven mixing increases the rate at which organisms cycle through the extreme light conditions at the top of the mixed layer and poorly lit conditions at the bottom of the mixed layer, and so may influence microbial communities and chl a concentrations by modifying light exposure. Incoming solar irradiance and wind speed were not significantly correlated (Spearman $r=-0.23$, $P=0.32$), albeit the three windiest sampling periods occurred in the winter. Additional studies with high-resolution sampling of strong wind events in the NPSG will help improve the mechanistic understanding of our observations.

**Solar irradiance correlates with beta diversity in surface water**

To uncover potential environmental factors influencing beta diversity across our 25 m time series, we looked for correlates among the beta diversity measures and changes in physical, chemical and biotic environmental properties (Supplementary Table S2.5). Unlike alpha diversity, the beta diversity measures demonstrated a significant seasonal trend. All the non-abundance-based beta diversity indices consistently correlated with the average solar irradiance reaching the ocean's surface derived from light energy between 0.28 to 2.8 μm wavelengths (Crescenti et al., 1989), Figure 2.4, Mantel test, BH adjusted $P<0.05$, Supplementary
The lower correlation between solar irradiance and the abundance-based measures demonstrates that lower abundance taxa largely contributed to this pattern, similar to our alpha diversity observations. The seasonal variables temperature and mixed layer depth were also correlated with some beta diversity measures, but the relationship was not as strong, nor as consistent as with solar irradiance (Supplementary Table S2.5). Chl $a$ also correlated with most beta diversity measures, presumably reflecting the phytoplankton response to changing light levels via photoadaptation. Our results are consistent with recent global-scale studies that found seasonally fluctuating variables including light and temperature strongly impact marine surface communities (Raes et al., 2011; Sunagawa et al., 2015). The stronger impact of light compared with temperature we observed at Station ALOHA contrasted with findings from the Tara Oceans Expedition, and may reflect differing impacts of these variables at different spatial scales (Sunagawa et al., 2015).

Serial sampling date did weakly correlate with some of the beta diversity indices at 25 m suggesting there may have been some autocorrelation among our samples. Ocean currents and water masses in the open ocean are very dynamic and therefore it is unlikely that any such autocorrelation we observed comes from sampling the same microbial population each month. Instead, regional-scale processes likely drive the observed temporal autocorrelation. Seasonal trends, including solar irradiance fall under this description. Regardless, the significant relationships between incoming solar irradiance and compositional dissimilarity were still significant after taking the time separating sampling dates into account (partial Mantel tests, $P<0.05$).

We compared beta diversity measures to the average incoming solar radiation using 24 h measurements across the 30 days prior to sampling. Therefore our solar irradiance values incorporate the influence of hours of daylight, cloud cover and seasonal variability in light intensity. Community composition significantly varied with hours of daylight alone as well, but the correlation was not as strong as that derived from the daily average over the preceding 30 days (Supplementary Table S2.5). This suggests that the quantity of incoming solar
irradiance is impacting microbial community composition rather than members of the community exhibiting photoperiodism.

The observed seasonal changes in microbial communities at Station ALOHA are more subtle than what has been previously reported at other marine microbial time-series stations. This likely reflects the less extreme environmental variability at Station ALOHA compared with coastal sites (Murray et al., 1998; Morris et al., 2005; Fuhrman et al., 2006; Gilbert et al., 2012). For example, *Rickettsiales* and *Rhodobacterales* alternate being the most abundant clade in surface water in the Western English Channel depending on season (Gilbert et al., 2012). At BATS, *Prochlorococcus* cell densities are an order of magnitude higher than other picophytoplankton until the spring water column mixing when *Synechococcus* bloom, *Prochlorococcus* cells decline and cell densities of these two groups of cyanobacteria become comparable (DuRand et al., 2001). At Station ALOHA, changes in community composition are less pronounced. *Prochlorococcus* and *Pelagibacter* remain the dominant organisms year round (Supplementary Figure S2.2 and S2.3). This is consistent with a 5-year study showing weak and inconsistent seasonality in total *Prochlorococcus* cell counts at Station ALOHA (Malmstrom et al., 2010). However these dominant groups (for example, *Prochlorococcus*, *Pelagibacter*) undoubtedly harbor microdiverse subpopulations that may also fluctuate with light (Kashtan et al., 2014).

The changes we observed in the relative abundance of some taxa at Station ALOHA between annual high and low solar irradiance periods, however, were consistent with seasonal studies in other oceanic regions (Figure 2.5). The clades that were more abundant in the amplicon data during high irradiance periods, SAR86, SAR116 and Rhodobacteraceae, were also reported as being more abundant during low nutrient, peak summer water stratification periods at BATS and/or were more abundant in the summer than winter in the Western English Channel (Figure 2.5, (Mary et al., 2006; Treusch et al., 2009; Gilbert et al., 2012; Giovannoni and Vergin, 2012)). This was consistent with the genomes we identified in the metagenomic data whose abundance varied with solar radiation when annotating metagenomic reads with the NCBI RefSeq database (Supplementary Figure S2.15).
We also observed a small but statistically significant increase in the relative abundance of some *Pelagibacter* strains during high solar irradiance periods (Supplementary Figure S2.15). Similarly, (Carlson et al., 2009) found that SAR11 cell densities distinctly increase when the mixed layer depth shoals at BATS in the spring.

The clades we observed to be more abundant during low-light periods of the year, including OCS116, SAR324, SAR202, *Synechococcus* and SAR406, were either reported as more abundant in surface water at BATS during spring mixing periods or as having peak abundances deeper in the water column at Station ALOHA ((Gordon and Giovannoni, 1996; Giovannoni et al., 1996; Wright et al., 1997; Treusch et al., 2009; Giovannoni and Vergin, 2012); Figure 2.5 and Supplementary Figure S2.15)). Consistent with our observations, previous studies have reported that *Synechococcus* cell counts at Station ALOHA peak in the winter (Malmstrom et al., 2010). We also observed that low light adapted *Prochlorococcus* marinus ecotypes NATL1A and NATL2A reached maximal abundances in our study at 25 m during low-light sampling periods (Supplementary Figure S2.15). *Prochlorococcus* strains NATL1A and NATL2A reach their highest abundances deeper in the water column, although they can tolerate short periods of more intense light exposure (Moore et al., 1998; Johnson et al., 2006; Kettler et al., 2007; Zinser et al., 2007). Similar seasonal trends have also been observed in NATL and other *Prochlorococcus* ecotypes (Zinser et al., 2007; Malmstrom et al., 2010; Kashtan et al., 2014).

Solar radiation can influence microbes in a variety of ways including direct impacts on cell physiology, indirect impacts through food web dynamics and chemical transformation of organic matter (Moran and Miller, 2007; Ruiz-González et al., 2013). Incident solar radiation also impacts ocean hydrology. For example, the deepening of the mixed layer in the winter may increase transport of microbes from the lower photic zone to 25 m. Our data indicated however, that solar irradiance was more strongly correlated with beta diversity than mixed layer depth. Solar radiation and temperature also exert top-down controls on microbial community structure, through changes in bacterivory and viral infection (Tsai et al., 2009, 2012; Ruiz-
González et al., 2013). Over the 2-year period examined, we observed an increase in the relative abundance of several cyanophage genomes in low-light samples (Supplementary Figure S2.15). Similarly, Parsons et al. observed that viral particle concentrations in the mixed layer at BATS are lowest during periods of summer water stratification (Parsons et al., 2012). Our observations, however, likely reflect an increase either in phage intracellular production or extracellular attachment, rather than an increase in planktonic phage particles, since free viral particles should not be efficiently retained on the sampling filters we used.

Co-occurrence network analyses of OTUs and orthologs

A weighted co-occurrence network analysis (Langfelder and Horvath, 2008) was used to identify potential subgroups of co-varying organisms in our samples that might be responding to seasonal or other measured environmental parameters not captured by the community level analysis. Taxa within co-varying subgroups may be directly or indirectly interacting or sharing a similar niche space. Using the 25 m amplicon data, we identified six clusters of co-correlated OTUs (‘modules’) that contained over five OTUs (Figure 2.6 and Supplementary Figure S2.16). Consistent with our previous findings, the first principal component (‘eigengene’; (Langfelder and Horvath, 2008)) of these modules most strongly correlated with seasonally fluctuating environmental measurements or average wind speed (Supplementary Table S2.6). The eigengenes of the modules containing the largest number of OTUs (96 and 23 total) significantly correlated with the seasonally fluctuating measurements solar irradiance and dissolved oxygen concentrations (Spearman $r$, BH adjusted $P<0.1$, Figure 2.6, Supplementary Table S2.6).

Interestingly the peak of module 1 and 2’s eigengenes are slightly offset, suggesting a potential seasonal succession. Several phylogenetic groups had OTUs with both positive and negative correlations to individual modules, or OTUs that were present in multiple modules, suggesting that there are distinct sets of ecotypes from different clades that appear to be co-varying. For example, there are OTUs belonging to Flavobacteriaceae and clade SAR406 that either positively or
negatively correlated with module 1, which peaked in early summer. OTUs belonging to clade SAR86 showed a similar pattern in module 3, which most strongly co-varied with water temperature (Figure 2.6).

These co-correlation patterns were not unique to the weighted network analysis, as an unweighted network analysis yielded similar clustering patterns for the 25 m data (Figure 2.6c). Particularly noteworthy in the unweighted network analysis was the tight clustering of OTUs assigned to module 1 of the weighted network (both correlated and anticorrelated to the eigengene of this module), demonstrating that the patterns of seasonal succession identified in this module are particularly prominent.

To corroborate the patterns at 25 m and uncover additional information as to how the genomic content of taxonomic groups may vary over time, we used weighted co-occurrence network analysis to explore the temporal patterns of protein-coding ortholog clusters within the abundant taxa in our 25 m metagenomic samples (Supplementary Figure S2.17). As expected, we observed that ortholog clusters from the same taxonomic group tended to cluster together into the same module, consistent with read abundance for ortholog clusters being influenced mainly by microbial abundances. Some ortholog clusters, however, were placed in different modules than the majority of clusters from the same taxa, suggesting the presence of microdiversity within these clades that might not co-vary uniformly with clade abundance. These orthologs may belong to genomic islands present only in certain ecotypes within a clade, or to mobile elements present only in some microbial groups. Confirmation of these observations awaits more complete reference data sets (from single-cell genomes or larger metagenomes), to allow mapping of these genes on specific ecotypes over time.

Analysis of the ortholog cluster temporal profiles at 25 m corroborated and extended our other analytical approaches. For example, the eigengenes of modules containing most cyanophage and clade SAR324 ortholog clusters, co-varied significantly with seasonally variable parameters (Supplementary Table S2.7 and Supplementary Figure S2.17). SAR324 in particular was positively associated with a deeper mixed layer, reflecting its greater predominance in deeper
waters. In addition, eigengenes of some of the smaller modules significantly correlated with total autotrophic biomass, nitrate+nitrite and particulate carbon concentrations. Module 1, containing the largest number of ortholog clusters, did not significantly correlate with any measured environmental parameters, being primarily composed of core genes in phylogenetic groups that are abundant throughout the year (Prochlorococcus, SAR11) (see Supplementary Table S2.7). The majority of SAR116 genes were found in modules 1 and 6 suggesting that although many SAR116 types are present throughout the year, some of its variants change in abundance. Several SAR116 ortholog clusters present in module 6 were associated with cofactor and vitamin metabolism, flagellar biosynthesis, several transporters and many housekeeping genes (most positively correlated with module 6 and peaking in abundance in spring and summer months). Similarly, module 10 was positively associated with solar irradiance and contained a number of SAR11 ortholog clusters annotated as transporters and genes involved in central carbon and amino acid metabolism (all positively correlated with module 10 and peaking in abundance in the summer). Interestingly, module 10 also included two Prochlorococcus ortholog clusters annotated as PhoB and PhoR, which are involved in phosphate acquisition during phosphate-limiting conditions, and were more abundant in the summer.

**Conclusions**

The results of this study support the hypothesis that the dampened seasonal variation in the physical environment at the NPSG may provide greater sensitivity for the detection of microbial community responses to non-seasonal phenomena. Research at Station ALOHA conducted by the HOT program over the past 25 years has shown that episodic (for example, mesoscale eddies) and inter-annual-scale (Pacific Decadal Oscillation and El Niño-Southern Oscillation), as well as seasonal processes, all may be important in carbon cycling and plankton dynamics (Letelier *et al.*, 2000; Fong *et al.*, 2008; Bidigare *et al.*, 2009; Karl *et al.*, 2012). Over a 2-year
period, in the absence of strong interannual variation we observed that episodic variation in wind speed was the predominant correlate of variability in microbial alpha diversity, but that seasonal variation, in particular solar irradiation, was key to microbial beta diversity. The former feature may be unique to the NPSG, but the latter appears to be a common feature in other marine surface water systems around the globe.

Future work that leverages long-term time-series analyses on a variety of spatial and temporal scales has significant potential to further enhance our understanding of the temporal dynamics and variability of microbial community diversity. Advances in technologies that increase the number of available reference genomes from indigenous ecotypes (Swan et al., 2013), that allow greater sequencing depth to facilitate more complete metagenomic assemblies (Sharon and Banfield, 2013), and that enable more automated, highly resolved sampling over a broader array of nested temporal scales (Robidart et al., 2014), all promise to improve our ability to observe the microbial ocean over increasingly resolved and relevant scales of space and time.
Figure 2.1. Environmental variation at Station ALOHA during our study period. (a) Orange ovals display the 24-hour daily average solar irradiance (light energy in watts from 0.28 to 2.8 μm wavelengths (shortwave) per m²) at the ocean's surface measured at the nearby WHOI buoy. The color gradient in (b) signifies water temperature across time and depth. The white line outlines the bottom of the mixed layer measured during HOT cruises. Dots mark sampling points.
Figure 2.2. Comparison of microbial communities at 25 and 500 m. (a) Venn diagram displaying the number of OTUs from each depth with relative abundances >0.1 % at either one or both depths. All amplicon data within each depth were combined to calculate relative abundances. (b) Boxplots comparing the distribution of phylogenetic diversity and Unifrac values between 25 and 500 m samples based on amplicon data. Whiskers delineate the full range of values. The variance of Unifrac values was significantly smaller in 500 m samples compared with 25 m samples (Levene's test, P<0.05). In b, amplicon data was resampled to 6906 reads per sample.
Figure 2.3. Relationship between wind speed and phylogenetic diversity at 25 m. Inset in a shows the average wind speed (m s⁻¹) of the 4 days before sample collection across time. (a) Relationship between average wind speed and phylogenetic diversity of samples characterized using amplicon data (BH adjusted P<0.01, Spearman r=0.7). The relationship was consistent when considering other alpha diversity measures and average wind speeds derived from a narrower or broader range of days leading up to sampling (Supplementary Table S2.8, Supplementary Figure S2.12). Each amplicon data set was resampled to 9661 reads (see Methods). (b) Relationship between average wind speed and chlorophyll a concentrations (P<0.05, Spearman r=0.46).
Figure 2.4. Nonmetric multidimensional scaling plot to visualize Unifrac distances between amplicon data sets at 25 m. Stress=0.15. Inset shows the solar irradiance (Wm⁻²) averaged across the 30 days leading up to sample collection. Red indicates high-incident solar irradiance data points that were >260 W m⁻². Blue indicates low-incident solar irradiance data points that were <175 W m⁻². Colors in inset correspond to colors in main figure. Each amplicon data set was resampled to 9661 reads (see Methods).
Figure 2.5. Cladogram highlighting bacterial OTUs and clades with differential relative abundances between samples experiencing high- and low-incident solar irradiance at 25 m (as defined in Figure 2.4). Shaded areas of branches delineate broad taxonomic groups. The innermost ring (ring #1) delineates clades that are more abundant during high light (>260 W m⁻² average solar irradiance of the preceding 30 days; yellow) or low light (<200 W m⁻² average solar irradiance of the preceding 30 days; light blue) sampling points (DESeq2 with BH correction, P<0.1). Letters correspond to listed clade names. The next two rings display the average relative abundance to within an order of magnitude of each OTU during high (ring # 2) and low light (ring # 3) sampling points. Red and blue bars in the outermost ring (ring # 4) mark individual OTUs that are significantly more abundant during high (red) and low (dark blue) light sampling points (DESeq2 with BH correction, P<0.1). Only OTUs with relative abundances across all samples summing to greater than 0.01% are shown. Abbreviations: Alpha, Gamma and Delta, denote corresponding Proteobacteria classes; Verruco, Verrucomicrobia; Actino: Actinobacteria.
Figure 2.6. (previous page) Weighted co-occurrence network analysis at 25 m. (a) OTU trends in the three largest modules with module eigengenes (gray solid lines, left axis) overlaid with their most strongly correlated environmental measurement (dashed lines, right axis). n and r correspond to the total number of OTUs in the module and the Spearman’s correlation between the module and overlaid environmental measurement, respectively. P-values (p) are shown where significant (BH adjusted P<0.1). Note that modules contain both positively and negatively correlating OTUs, resulting in sets of OTUs whose relative abundances either positively or negatively correlate with eigengene trends. The bar charts (b) display the number of OTUs that are positively (red) or negatively (blue) correlated with the eigengene trend displayed to the left (a). (c) An unweighted network of OTUs identified in the 25 m data set. Nodes represent OTUs that are colored according to the module to which they were assigned in the weighted network analysis, with ‘a’ and ‘b’ signifiers representing OTUs that are either positively or negatively associated with the eigengene of that module, respectively (gray nodes were either assigned to another module or not assigned to any module). Edges are colored based on whether the OTUs were found to be correlated (blue) or anticorrelated (red). Note that although the network is unweighted similar patterns identified in the weighted network analysis are recapitulated, such as tight clustering of correlated and anticorrelated nodes assigned to module 1 in the weighted network analysis.
References


**Acknowledgements**

We thank Dr. Robert R. Bidigare, Dr. Yoshimi M. Rii and members of the DeLong Lab for valuable discussions, Craig Nosse for assistance with WHOTS Buoy data, Lance Fujieki for assistance with HOT environmental data, Rachel Barry and Tsultrim Palden for preparing samples for pyrosequencing and Ramunas Stepanauskas for providing single-cell genomes before publication. We are indebted to the HOT scientists and staff and to the captains and crews of the research vessel R/V Kilo Moana and R/V KOK and for logistical support. This work was supported by grants from the Gordon and Betty Moore Foundation #492.01 and 3777 (EFD), #3794 (DMK), and by the US Environmental Protection Agency STAR Fellowship (JAB). In addition, We acknowledge NSF for support of the HOT program (including the most recent OCE1260164 to MJC and DMK) the Center for Microbial Oceanography: Research and Education (C-MORE; EF0424599 to DMK and EFD) and the
Simons Collaboration on Ocean Processes and Ecology (SCOPE; #329108 to DMK, EFD). Data from the WHOTS surface mooring are gratefully acknowledged; the NOAA Climate Observation Division provides funding to Robert A Weller and Albert J Plueddemann at WHOI to support the long-term deployment of the surface mooring. This work is a contribution of the Center for Microbial Oceanography: Research and Education, and the Simons Collaboration on Ocean Processes and Ecology.
Supplementary Materials for Chapter 2

Table S2.1. Twenty-five meter shotgun metagenomic library information

<table>
<thead>
<tr>
<th>sample name</th>
<th>depth (m)</th>
<th>Sample Date GMT (M/D/Y)</th>
<th>Sampling Time GMT (H:M)</th>
<th>chemistry</th>
<th>non-duplicated reads</th>
<th>16s reads</th>
<th>average read length</th>
<th>NCBI Accession #</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOT194_25m</td>
<td>25</td>
<td>8/6/07</td>
<td>02:34</td>
<td>flex</td>
<td>530175</td>
<td>929</td>
<td>256.3</td>
<td>SRX556047</td>
</tr>
<tr>
<td>HOT195_25m</td>
<td>25</td>
<td>9/5/07</td>
<td>00:09</td>
<td>flex</td>
<td>458155</td>
<td>968</td>
<td>253.7</td>
<td>SRX556048</td>
</tr>
<tr>
<td>HOT196_25m</td>
<td>25</td>
<td>10/2/07</td>
<td>12:00</td>
<td>flex</td>
<td>529406</td>
<td>892</td>
<td>244.7</td>
<td>SRX556049</td>
</tr>
<tr>
<td>HOT197_25m</td>
<td>25</td>
<td>12/1/07</td>
<td>10:51</td>
<td>flex</td>
<td>601796</td>
<td>1015</td>
<td>239.3</td>
<td>SRX556050</td>
</tr>
<tr>
<td>HOT198_25m</td>
<td>25</td>
<td>12/20/07</td>
<td>09:38</td>
<td>flex</td>
<td>517483</td>
<td>930</td>
<td>214.2</td>
<td>SRX556051</td>
</tr>
<tr>
<td>HOT199_25m</td>
<td>25</td>
<td>1/29/08</td>
<td>11:33</td>
<td>titanium</td>
<td>1378412</td>
<td>2709</td>
<td>392.1</td>
<td>SRX556052</td>
</tr>
<tr>
<td>HOT200_25m</td>
<td>25</td>
<td>2/23/08</td>
<td>11:15</td>
<td>titanium</td>
<td>1287149</td>
<td>2750</td>
<td>339.4</td>
<td>SRX556053</td>
</tr>
<tr>
<td>HOT201_25m</td>
<td>25</td>
<td>5/27/08</td>
<td>10:34</td>
<td>titanium</td>
<td>1284251</td>
<td>2599</td>
<td>389.1</td>
<td>SRX556054</td>
</tr>
<tr>
<td>HOT202_25m</td>
<td>25</td>
<td>6/25/08</td>
<td>10:46</td>
<td>titanium</td>
<td>1481716</td>
<td>3008</td>
<td>334.3</td>
<td>SRX556055</td>
</tr>
<tr>
<td>HOT203_25m</td>
<td>25</td>
<td>7/26/08</td>
<td>11:59</td>
<td>titanium</td>
<td>1400027</td>
<td>2918</td>
<td>288</td>
<td>SRX556056</td>
</tr>
<tr>
<td>HOT204_25m</td>
<td>25</td>
<td>8/16/08</td>
<td>10:25</td>
<td>titanium</td>
<td>1282945</td>
<td>2984</td>
<td>345.7</td>
<td>SRX556057</td>
</tr>
<tr>
<td>HOT205_25m</td>
<td>25</td>
<td>10/10/08</td>
<td>10:34</td>
<td>titanium</td>
<td>1394074</td>
<td>3170</td>
<td>409.8</td>
<td>SRX556058</td>
</tr>
<tr>
<td>HOT206_25m</td>
<td>25</td>
<td>11/30/08</td>
<td>12:01</td>
<td>titanium (half plate)</td>
<td>793958</td>
<td>1865</td>
<td>360.8</td>
<td>SRX556059</td>
</tr>
<tr>
<td>HOT208_25m</td>
<td>25</td>
<td>1/20/09</td>
<td>11:35</td>
<td>titanium</td>
<td>1333536</td>
<td>2841</td>
<td>420.9</td>
<td>SRX556060</td>
</tr>
<tr>
<td>HOT209_25m</td>
<td>25</td>
<td>2/17/09</td>
<td>12:54</td>
<td>titanium</td>
<td>1239004</td>
<td>2725</td>
<td>413.8</td>
<td>SRX556061</td>
</tr>
<tr>
<td>HOT210_25m</td>
<td>25</td>
<td>4/28/09</td>
<td>10:42</td>
<td>titanium</td>
<td>1236790</td>
<td>2899</td>
<td>403.4</td>
<td>SRX556062</td>
</tr>
<tr>
<td>HOT211_25m</td>
<td>25</td>
<td>5/28/09</td>
<td>13:53</td>
<td>titanium (half plate)</td>
<td>729212</td>
<td>1717</td>
<td>434.9</td>
<td>SRX556063</td>
</tr>
<tr>
<td>HOT212_25m</td>
<td>25</td>
<td>7/3/09</td>
<td>11:37</td>
<td>titanium</td>
<td>1273388</td>
<td>2842</td>
<td>386.9</td>
<td>SRX556064</td>
</tr>
<tr>
<td>HOT213_25m</td>
<td>25</td>
<td>7/24/09</td>
<td>12:02</td>
<td>titanium</td>
<td>1318710</td>
<td>3634</td>
<td>413.7</td>
<td>SRX556065</td>
</tr>
<tr>
<td>HOT214_25m</td>
<td>25</td>
<td>8/18/09</td>
<td>10:51</td>
<td>titanium</td>
<td>1188282</td>
<td>2726</td>
<td>415.5</td>
<td>SRX556066</td>
</tr>
<tr>
<td>HOT215_25m</td>
<td>25</td>
<td>9/24/09</td>
<td>12:42</td>
<td>titanium</td>
<td>1132896</td>
<td>2855</td>
<td>360.5</td>
<td>SRX556067</td>
</tr>
</tbody>
</table>
Table S2.2. Five-hundred meter shotgun metagenomic library information

<table>
<thead>
<tr>
<th>sample name</th>
<th>depth (m)</th>
<th>Sample Date (M/D/Y)</th>
<th>Sampling Time GMT (H:M)</th>
<th>chemistry</th>
<th>non-duplicated reads</th>
<th>16s reads</th>
<th>average read length</th>
<th>NCBI Accession #</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOT194_500m</td>
<td>500</td>
<td>8/5/07</td>
<td>03:36</td>
<td>flex</td>
<td>384324</td>
<td>695</td>
<td>245.2</td>
<td>SRX556068</td>
</tr>
<tr>
<td>HOT195_500m</td>
<td>500</td>
<td>9/3/07</td>
<td>21:45</td>
<td>titanium</td>
<td>1498998</td>
<td>2506</td>
<td>366</td>
<td>SRX556069</td>
</tr>
<tr>
<td>HOT196_500m</td>
<td>500</td>
<td>10/3/07</td>
<td>21:33</td>
<td>titanium</td>
<td>1358148</td>
<td>2346</td>
<td>287.9</td>
<td>SRX556070</td>
</tr>
<tr>
<td>HOT197_500m</td>
<td>500</td>
<td>12/2/07</td>
<td>21:40</td>
<td>titanium</td>
<td>1441382</td>
<td>2723</td>
<td>424.9</td>
<td>SRX556071</td>
</tr>
<tr>
<td>HOT198_500m</td>
<td>500</td>
<td>12/21/07</td>
<td>21:48</td>
<td>titanium</td>
<td>1466456</td>
<td>2615</td>
<td>329.9</td>
<td>SRX556072</td>
</tr>
<tr>
<td>HOT199_500m</td>
<td>500</td>
<td>1/30/08</td>
<td>21:40</td>
<td>titanium</td>
<td>1398928</td>
<td>2931</td>
<td>440.9</td>
<td>SRX556073</td>
</tr>
<tr>
<td>HOT200_500m</td>
<td>500</td>
<td>2/24/08</td>
<td>09:43</td>
<td>titanium</td>
<td>1408054</td>
<td>2423</td>
<td>408.1</td>
<td>SRX556074</td>
</tr>
<tr>
<td>HOT201_500m</td>
<td>500</td>
<td>5/28/08</td>
<td>09:46</td>
<td>titanium</td>
<td>1369491</td>
<td>2936</td>
<td>392.6</td>
<td>SRX556075</td>
</tr>
<tr>
<td>HOT202_500m</td>
<td>500</td>
<td>6/26/08</td>
<td>09:41</td>
<td>titanium</td>
<td>1082386</td>
<td>2191</td>
<td>425.7</td>
<td>SRX556076</td>
</tr>
<tr>
<td>HOT203_500m</td>
<td>500</td>
<td>7/27/08</td>
<td>09:50</td>
<td>titanium</td>
<td>1609628</td>
<td>3100</td>
<td>355.8</td>
<td>SRX556077</td>
</tr>
<tr>
<td>HOT204_500m</td>
<td>500</td>
<td>8/17/08</td>
<td>09:48</td>
<td>titanium</td>
<td>1534212</td>
<td>3094</td>
<td>380.8</td>
<td>SRX556078</td>
</tr>
<tr>
<td>HOT205_500m</td>
<td>500</td>
<td>10/11/08</td>
<td>09:52</td>
<td>titanium</td>
<td>1392115</td>
<td>2950</td>
<td>437.9</td>
<td>SRX556079</td>
</tr>
<tr>
<td>HOT206_500m</td>
<td>500</td>
<td>12/2/08</td>
<td>09:43</td>
<td>titanium (half plate)</td>
<td>822821</td>
<td>1507</td>
<td>362.3</td>
<td>SRX556080</td>
</tr>
<tr>
<td>HOT208_500m</td>
<td>500</td>
<td>1/21/09</td>
<td>07:25</td>
<td>titanium</td>
<td>1332923</td>
<td>2571</td>
<td>419.6</td>
<td>SRX556081</td>
</tr>
<tr>
<td>HOT209_500m</td>
<td>500</td>
<td>2/18/09</td>
<td>10:06</td>
<td>titanium</td>
<td>933997</td>
<td>1771</td>
<td>416</td>
<td>SRX556082</td>
</tr>
<tr>
<td>HOT210_500m</td>
<td>500</td>
<td>4/29/09</td>
<td>09:44</td>
<td>titanium</td>
<td>1301060</td>
<td>2566</td>
<td>435.2</td>
<td>SRX556083</td>
</tr>
<tr>
<td>HOT211_500m</td>
<td>500</td>
<td>5/28/09</td>
<td>09:44</td>
<td>titanium (half plate)</td>
<td>741152</td>
<td>1475</td>
<td>420.9</td>
<td>SRX556084</td>
</tr>
<tr>
<td>HOT212_500m</td>
<td>500</td>
<td>7/4/09</td>
<td>09:54</td>
<td>titanium</td>
<td>1257921</td>
<td>2211</td>
<td>349.2</td>
<td>SRX556085</td>
</tr>
<tr>
<td>HOT213_500m</td>
<td>500</td>
<td>7/26/09</td>
<td>06:48</td>
<td>titanium</td>
<td>1226981</td>
<td>2583</td>
<td>371.3</td>
<td>SRX556086</td>
</tr>
<tr>
<td>HOT214_500m</td>
<td>500</td>
<td>8/19/09</td>
<td>10:15</td>
<td>titanium</td>
<td>1125152</td>
<td>2127</td>
<td>411.4</td>
<td>SRX556087</td>
</tr>
<tr>
<td>HOT215_500m</td>
<td>500</td>
<td>9/25/09</td>
<td>09:44</td>
<td>titanium</td>
<td>1346960</td>
<td>2735</td>
<td>365.4</td>
<td>SRX556088</td>
</tr>
</tbody>
</table>
## Table S2.3. Bacterial SSU rDNA amplicon library information

<table>
<thead>
<tr>
<th>sample name</th>
<th>depth (m)</th>
<th>Sample Date GMT (M/D/Y)</th>
<th>Sampling Time GMT (H:M)</th>
<th># of de-noised reads</th>
<th>average length of de-noised reads (bp)</th>
<th>barcode(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOT194_25m</td>
<td>25</td>
<td>8/6/07 02:34</td>
<td>13552</td>
<td>339.1</td>
<td>ATACCAC</td>
<td></td>
</tr>
<tr>
<td>HOT195_25m</td>
<td>25</td>
<td>9/5/07 00:09</td>
<td>10444</td>
<td>346</td>
<td>TCCGCTC</td>
<td></td>
</tr>
<tr>
<td>HOT196_25m</td>
<td>25</td>
<td>10/2/07 12:00</td>
<td>14210</td>
<td>349.6</td>
<td>ATCTTAC</td>
<td></td>
</tr>
<tr>
<td>HOT197_25m</td>
<td>25</td>
<td>12/1/07 10:51</td>
<td>11062</td>
<td>350.1</td>
<td>AACCAGC</td>
<td></td>
</tr>
<tr>
<td>HOT198_25m</td>
<td>25</td>
<td>12/20/07 09:38</td>
<td>11991</td>
<td>334.2</td>
<td>TTCGAGC</td>
<td></td>
</tr>
<tr>
<td>HOT199_25m</td>
<td>25</td>
<td>1/29/08 11:33</td>
<td>9661</td>
<td>342.1</td>
<td>TCACCTC</td>
<td></td>
</tr>
<tr>
<td>HOT200_25m</td>
<td>25</td>
<td>2/23/08 11:15</td>
<td>12661</td>
<td>362.7</td>
<td>CAAGAAC, CAGGC, TATTCGTC</td>
<td></td>
</tr>
<tr>
<td>HOT201_25m</td>
<td>25</td>
<td>5/27/08 10:34</td>
<td>11315</td>
<td>352.6</td>
<td>CAAGAAC, CAGGC, TATTCGTC</td>
<td></td>
</tr>
<tr>
<td>HOT202_25m</td>
<td>25</td>
<td>6/25/08 10:46</td>
<td>12476</td>
<td>351.9</td>
<td>ACGGCCTC</td>
<td></td>
</tr>
<tr>
<td>HOT203_25m</td>
<td>25</td>
<td>7/26/08 11:59</td>
<td>10570</td>
<td>357.3</td>
<td>CGTGAC, TGAAGC, TATCAAC</td>
<td></td>
</tr>
<tr>
<td>HOT204_25m</td>
<td>25</td>
<td>8/16/08 10:25</td>
<td>14050</td>
<td>358.2</td>
<td>ACTCCTC</td>
<td></td>
</tr>
<tr>
<td>HOT205_25m</td>
<td>25</td>
<td>10/10/08 10:34</td>
<td>13966</td>
<td>363.3</td>
<td>AGCTTC</td>
<td></td>
</tr>
<tr>
<td>HOT206_25m</td>
<td>25</td>
<td>11/30/08 12:01</td>
<td>15018</td>
<td>352.1</td>
<td>ACTCAC</td>
<td></td>
</tr>
<tr>
<td>HOT208_25m</td>
<td>25</td>
<td>1/20/09 11:35</td>
<td>11449</td>
<td>342.3</td>
<td>CGCAAC</td>
<td></td>
</tr>
<tr>
<td>HOT209_25m</td>
<td>25</td>
<td>2/17/09 12:54</td>
<td>10843</td>
<td>357</td>
<td>CGGTATC</td>
<td></td>
</tr>
<tr>
<td>HOT210_25m</td>
<td>25</td>
<td>4/28/09 10:42</td>
<td>11429</td>
<td>354.2</td>
<td>TAATCTC</td>
<td></td>
</tr>
<tr>
<td>HOT211_25m</td>
<td>25</td>
<td>5/28/09 13:53</td>
<td>11107</td>
<td>346.7</td>
<td>ACAAGGC</td>
<td></td>
</tr>
<tr>
<td>HOT212_25m</td>
<td>25</td>
<td>7/3/09 11:37</td>
<td>11248</td>
<td>346.7</td>
<td>AAAGTGC</td>
<td></td>
</tr>
<tr>
<td>HOT213_25m</td>
<td>25</td>
<td>7/24/09 12:02</td>
<td>12593</td>
<td>348.8</td>
<td>CCAGGAC</td>
<td></td>
</tr>
<tr>
<td>HOT214_25m</td>
<td>25</td>
<td>8/18/09 10:51</td>
<td>13612</td>
<td>367.3</td>
<td>ACTTGC</td>
<td></td>
</tr>
<tr>
<td>HOT215_25m</td>
<td>25</td>
<td>9/24/09 12:42</td>
<td>15310</td>
<td>355</td>
<td>TGAGCAG</td>
<td></td>
</tr>
<tr>
<td>HOT194_500m</td>
<td>500</td>
<td>8/5/07 03:36</td>
<td>18845</td>
<td>343.6</td>
<td>ATACCAC</td>
<td></td>
</tr>
<tr>
<td>HOT195_500m</td>
<td>500</td>
<td>9/3/07 21:45</td>
<td>15567</td>
<td>346.7</td>
<td>TCCGCTC</td>
<td></td>
</tr>
<tr>
<td>HOT196_500m</td>
<td>500</td>
<td>10/3/07 21:33</td>
<td>14961</td>
<td>350.8</td>
<td>ATCTTAC</td>
<td></td>
</tr>
<tr>
<td>HOT197_500m</td>
<td>500</td>
<td>12/2/07 21:40</td>
<td>7158</td>
<td>348.5</td>
<td>AACCAGC</td>
<td></td>
</tr>
<tr>
<td>HOT198_500m</td>
<td>500</td>
<td>12/21/07 21:48</td>
<td>14396</td>
<td>351.4</td>
<td>TTCGAGC</td>
<td></td>
</tr>
<tr>
<td>HOT199_500m</td>
<td>500</td>
<td>1/30/08 21:40</td>
<td>6908</td>
<td>344.1</td>
<td>TCCGAC, ACGGC</td>
<td></td>
</tr>
<tr>
<td>HOT201_500m</td>
<td>500</td>
<td>5/28/08 09:46</td>
<td>7951</td>
<td>346.7</td>
<td>CAAGAAC, CAGGC, TATTCGTC</td>
<td></td>
</tr>
<tr>
<td>HOT202_500m</td>
<td>500</td>
<td>6/26/08 09:41</td>
<td>14585</td>
<td>346.7</td>
<td>ACGGCCTC</td>
<td></td>
</tr>
<tr>
<td>HOT203_500m</td>
<td>500</td>
<td>7/27/08 09:50</td>
<td>15367</td>
<td>348</td>
<td>CGTGAC, TGAAGC, TATCAAC</td>
<td></td>
</tr>
<tr>
<td>HOT204_500m</td>
<td>500</td>
<td>8/17/08 09:48</td>
<td>13729</td>
<td>345.8</td>
<td>ACTCCTC</td>
<td></td>
</tr>
<tr>
<td>HOT205_500m</td>
<td>500</td>
<td>10/11/08</td>
<td>09:52</td>
<td>13987</td>
<td>350.8</td>
<td>AGCTTC</td>
</tr>
<tr>
<td>HOT206_500m</td>
<td>500</td>
<td>12/2/08</td>
<td>09:43</td>
<td>15766</td>
<td>349</td>
<td>ACTCAG</td>
</tr>
<tr>
<td>HOT208_500m</td>
<td>500</td>
<td>1/21/09</td>
<td>07:25</td>
<td>4221</td>
<td>358.5</td>
<td>CGCAAC</td>
</tr>
<tr>
<td>HOT209_500m</td>
<td>500</td>
<td>2/18/09</td>
<td>10:06</td>
<td>11733</td>
<td>353.6</td>
<td>CGGTATC</td>
</tr>
<tr>
<td>HOT210_500m</td>
<td>500</td>
<td>4/29/09</td>
<td>09:44</td>
<td>12788</td>
<td>346.6</td>
<td>TAATCTC</td>
</tr>
<tr>
<td>HOT211_500m</td>
<td>500</td>
<td>5/28/09</td>
<td>09:44</td>
<td>13623</td>
<td>351.4</td>
<td>ACAAGGC</td>
</tr>
<tr>
<td>HOT212_500m</td>
<td>500</td>
<td>7/4/09</td>
<td>09:54</td>
<td>11478</td>
<td>348.7</td>
<td>AAGGTGC</td>
</tr>
<tr>
<td>HOT214_500m</td>
<td>500</td>
<td>8/19/09</td>
<td>10:15</td>
<td>12271</td>
<td>346.6</td>
<td>ACTTGC</td>
</tr>
<tr>
<td>HOT215_500m</td>
<td>500</td>
<td>9/25/09</td>
<td>09:44</td>
<td>16258</td>
<td>347.8</td>
<td>TGACGAC</td>
</tr>
</tbody>
</table>
**Table S2.4.** Correlations between microbial community alpha diversity and environmental properties at 25 m.  
All diversity indices were resampled using rarefaction. Read depths as follows: Amplicon reads: 9661, Protein-coding reads from metagenomic data: 476661. Symbols: § averaged over previous 4 days, † averaged over previous 30 days, *significant below P< 0.1 after applying the Benjamini-Hochberg procedure to account for multiple tests (n=24, hours of day light not included due to similarity with shortwave radiation).

<table>
<thead>
<tr>
<th></th>
<th>Phylogenetic diversity using amplicon data</th>
<th>Phylogenetic entropy using amplicon data</th>
<th>Richness using functional gene clusters</th>
<th>Shannon index using functional gene clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Spearman r</td>
<td>single test p-value</td>
<td>Spearman r</td>
<td>single test p-value</td>
</tr>
<tr>
<td>Serial sampling day</td>
<td>-0.19</td>
<td>0.42</td>
<td>-0.17</td>
<td>0.46</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.05</td>
<td>0.83</td>
<td>0.3</td>
<td>0.19</td>
</tr>
<tr>
<td>Salinity</td>
<td>-0.38</td>
<td>0.09</td>
<td>-0.01</td>
<td>0.96</td>
</tr>
<tr>
<td>Wind speed§</td>
<td>0.7</td>
<td>0.0006*</td>
<td>0.44</td>
<td>0.05</td>
</tr>
<tr>
<td>Shortwave (solar) irradiance†</td>
<td>-0.26</td>
<td>0.24</td>
<td>0.15</td>
<td>0.51</td>
</tr>
<tr>
<td>Daily mean sea level anomaly</td>
<td>0.49</td>
<td>0.02</td>
<td>0.02</td>
<td>0.92</td>
</tr>
<tr>
<td>Mixed layer depth</td>
<td>0.37</td>
<td>0.10</td>
<td>-0.07</td>
<td>0.77</td>
</tr>
<tr>
<td>Deep chlorophyll maximum depth</td>
<td>-0.26</td>
<td>0.26</td>
<td>-0.17</td>
<td>0.46</td>
</tr>
<tr>
<td>Dissolved oxygen</td>
<td>-0.33</td>
<td>0.14</td>
<td>-0.29</td>
<td>0.20</td>
</tr>
<tr>
<td>Dissolved inorganic carbon</td>
<td>-0.23</td>
<td>0.32</td>
<td>-0.16</td>
<td>0.48</td>
</tr>
<tr>
<td>Dissolved organic carbon</td>
<td>0.19</td>
<td>0.42</td>
<td>0.21</td>
<td>0.35</td>
</tr>
<tr>
<td>Particulate carbon</td>
<td>-0.15</td>
<td>0.52</td>
<td>0.1</td>
<td>0.68</td>
</tr>
<tr>
<td>Low-level nitrate + nitrite</td>
<td>0.25</td>
<td>0.28</td>
<td>-0.11</td>
<td>0.62</td>
</tr>
<tr>
<td>Particulate nitrogen</td>
<td>-0.08</td>
<td>0.75</td>
<td>0.05</td>
<td>0.83</td>
</tr>
<tr>
<td>Particulate phosphorus</td>
<td>0.15</td>
<td>0.52</td>
<td>0.01</td>
<td>0.95</td>
</tr>
<tr>
<td>Low-level phosphorus</td>
<td>0.09</td>
<td>0.69</td>
<td>-0.02</td>
<td>0.93</td>
</tr>
<tr>
<td>Particulate silica</td>
<td>0.34</td>
<td>0.13</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Silicate</td>
<td>Heterotrophic bacteria cell counts</td>
<td>Prochlorococcus cell counts</td>
<td>Synechococcus cell counts</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------</td>
<td>-----------------------------------</td>
<td>-----------------------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td></td>
<td>-0.22</td>
<td>0.33</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>Heterotrophic bacteria</td>
<td>-0.02</td>
<td>0.92</td>
<td>0.21</td>
<td>0.37</td>
</tr>
<tr>
<td>cell counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prochlorococcus cell</td>
<td>-0.2</td>
<td>0.38</td>
<td>0.04</td>
<td>0.87</td>
</tr>
<tr>
<td>counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synechococcus cell counts</td>
<td>-0.06</td>
<td>0.79</td>
<td>-0.01</td>
<td>0.96</td>
</tr>
<tr>
<td>Picoeukaryote cell counts</td>
<td>0.05</td>
<td>0.83</td>
<td>0.01</td>
<td>0.97</td>
</tr>
<tr>
<td>Chlorophyll a</td>
<td>0.56</td>
<td>0.008*</td>
<td>-0.04</td>
<td>0.87</td>
</tr>
<tr>
<td>Total autotroph biomass</td>
<td>0.16</td>
<td>0.47</td>
<td>0.10</td>
<td>0.68</td>
</tr>
<tr>
<td>Hours of day light</td>
<td>-0.31</td>
<td>0.18</td>
<td>-0.04</td>
<td>0.87</td>
</tr>
</tbody>
</table>
Table S2.5. Mantel tests comparing microbial community beta diversity and environmental properties at 25 m. All diversity indices were resampled using rarefaction. Read depths as follows: Amplicon reads: 9661, Protein-coding reads from metagenomic data: 476666. The Spearman correlation coefficient was used for mantel tests. Symbols: § averaged over previous 4 days, ¶ averaged over previous 30 days, *significant below P< 0.1 after applying the Benjamini-Hochberg Procedure to account for multiple tests (n=24, hours of day light not included due to similarity with shortwave radiation).

<table>
<thead>
<tr>
<th></th>
<th>Unifrac using amplicon data</th>
<th>Weighted Unifrac using amplicon data</th>
<th>Jaccard index using functional gene clusters</th>
<th>Bray Curtis index using functional gene clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mantel r</td>
<td>single test p-value</td>
<td>mantel r</td>
<td>single test p-value</td>
</tr>
<tr>
<td>Serial sampling day</td>
<td>-0.01</td>
<td>0.49</td>
<td>0.17</td>
<td>0.02</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.31</td>
<td>0.002*</td>
<td>0.04</td>
<td>0.31</td>
</tr>
<tr>
<td>Salinity</td>
<td>0.00</td>
<td>0.44</td>
<td>-0.08</td>
<td>0.80</td>
</tr>
<tr>
<td>Wind speed§</td>
<td>0.20</td>
<td>0.04</td>
<td>0.02</td>
<td>0.41</td>
</tr>
<tr>
<td>Shortwave (solar)</td>
<td>0.57</td>
<td>0.0001*</td>
<td>0.26</td>
<td>0.01</td>
</tr>
<tr>
<td>irradiance¶</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Daily mean sea level</td>
<td>0.04</td>
<td>0.34</td>
<td>-0.15</td>
<td>0.52</td>
</tr>
<tr>
<td>anomaly</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mixed layer depth</td>
<td>0.44</td>
<td>0.0002*</td>
<td>0.08</td>
<td>0.27</td>
</tr>
<tr>
<td>Deep chlorophyll</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maximum depth</td>
<td>-0.04</td>
<td>0.68</td>
<td>-0.16</td>
<td>0.99</td>
</tr>
<tr>
<td>Oxygen</td>
<td>0.02</td>
<td>0.39</td>
<td>-0.10</td>
<td>0.84</td>
</tr>
<tr>
<td>Dissolved inorganic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>carbon</td>
<td>-0.01</td>
<td>0.51</td>
<td>-0.10</td>
<td>0.88</td>
</tr>
<tr>
<td>Dissolved organic</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>carbon</td>
<td>0.13</td>
<td>0.13</td>
<td>0.16</td>
<td>0.10</td>
</tr>
<tr>
<td>Particulate carbon</td>
<td>0.05</td>
<td>0.28</td>
<td>-0.09</td>
<td>0.74</td>
</tr>
<tr>
<td>Particulate nitrogen</td>
<td>-0.04</td>
<td>0.66</td>
<td>-0.15</td>
<td>0.92</td>
</tr>
<tr>
<td>Particulate phosphorus</td>
<td>-0.04</td>
<td>0.61</td>
<td>-0.16</td>
<td>0.93</td>
</tr>
<tr>
<td>Low-level phosphorus</td>
<td>0.00</td>
<td>0.48</td>
<td>-0.14</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>0.07</td>
<td>0.22</td>
<td>0.08</td>
<td>0.28</td>
</tr>
<tr>
<td>-------------------------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>Low-level nitrogen</td>
<td>0.10</td>
<td>0.14</td>
<td>-0.14</td>
<td>0.94</td>
</tr>
<tr>
<td>Particulate silica</td>
<td>0.12</td>
<td>0.10</td>
<td>0.11</td>
<td>0.12</td>
</tr>
<tr>
<td>Silicate</td>
<td>-0.04</td>
<td>0.64</td>
<td>-0.11</td>
<td>0.84</td>
</tr>
<tr>
<td>Heterotrophic bacteria</td>
<td>0.02</td>
<td>0.41</td>
<td>-0.02</td>
<td>0.49</td>
</tr>
<tr>
<td>cell counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prochlorococcus cell</td>
<td>0.21</td>
<td>0.07</td>
<td>0.23</td>
<td>0.06</td>
</tr>
<tr>
<td>counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synechococcus cell</td>
<td>0.23</td>
<td>0.03</td>
<td>0.24</td>
<td>0.06</td>
</tr>
<tr>
<td>counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Picoeukaryote cell</td>
<td>0.27</td>
<td>0.013*</td>
<td>0.30</td>
<td>0.03</td>
</tr>
<tr>
<td>counts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chlorophyll a</td>
<td>-0.09</td>
<td>0.78</td>
<td>-0.06</td>
<td>0.65</td>
</tr>
<tr>
<td>Total autotroph biomass</td>
<td>0.44</td>
<td>0.001</td>
<td>0.22</td>
<td>0.02</td>
</tr>
<tr>
<td>Hours of day light</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table S2.6. Correlations between OTU module eigengenes and the most strongly correlated environmental measurements at 25 m. Modules were calculated using the 193 most abundant OTUs in the amplicon datasets (see methods). Both positively and negatively correlated OTUs were included in the same module. All environmental parameters that significantly correlated (BH corrected $P < 0.1$) with ME1 and ME2 are shown. For modules 3-6 only the environmental parameter with the strongest correlation is shown, as no environmental parameters significantly correlated with these modules. P-values were adjusted using the Benjamini-Hochberg procedure to account for multiple tests ($n=144$, 24 environmental measurements (as in table S8) x 6 modules.) § averaged over previous 4 days, ¶ averaged over previous 30 days.

<table>
<thead>
<tr>
<th>Module</th>
<th>Environmental measurement</th>
<th># OTUs in module</th>
<th>Correlation (Spearman r)</th>
<th>Single $p$-value</th>
<th>Adjusted $p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME1</td>
<td>Shortwave (solar) irradiance ¶</td>
<td>96</td>
<td>0.82</td>
<td>2.93E-06</td>
<td>0.0002</td>
</tr>
<tr>
<td>ME1</td>
<td>Mixed layer depth</td>
<td>96</td>
<td>-0.76</td>
<td>0.0001</td>
<td>0.003</td>
</tr>
<tr>
<td>ME1</td>
<td>Chlorophyll a</td>
<td>96</td>
<td>-0.73</td>
<td>0.0002</td>
<td>0.005</td>
</tr>
<tr>
<td>ME1</td>
<td>Temperature</td>
<td>96</td>
<td>0.68</td>
<td>0.002</td>
<td>0.03</td>
</tr>
<tr>
<td>ME1</td>
<td>Picoeukaryote cell counts</td>
<td>96</td>
<td>-0.61</td>
<td>0.003</td>
<td>0.06</td>
</tr>
<tr>
<td>ME2</td>
<td>Dissolved oxygen §</td>
<td>23</td>
<td>0.86</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ME2</td>
<td>Temperature</td>
<td>23</td>
<td>-0.78</td>
<td>4.00E-05</td>
<td>0.002</td>
</tr>
<tr>
<td>ME2</td>
<td>Dissolved organic carbon</td>
<td>23</td>
<td>-0.76</td>
<td>9.00E-05</td>
<td>0.003</td>
</tr>
<tr>
<td>ME2</td>
<td>Picoeukaryote cell counts</td>
<td>23</td>
<td>0.72</td>
<td>0.0002</td>
<td>0.006</td>
</tr>
<tr>
<td>ME2</td>
<td>Dissolved inorganic carbon</td>
<td>23</td>
<td>0.59</td>
<td>0.005</td>
<td>0.09</td>
</tr>
<tr>
<td>ME3</td>
<td>Temperature</td>
<td>19</td>
<td>0.54</td>
<td>0.013</td>
<td>0.16</td>
</tr>
<tr>
<td>ME4</td>
<td>Wind speed §</td>
<td>13</td>
<td>0.49</td>
<td>0.026</td>
<td>0.27</td>
</tr>
<tr>
<td>ME5</td>
<td>Dissolved oxygen</td>
<td>10</td>
<td>-0.57</td>
<td>0.008</td>
<td>0.12</td>
</tr>
<tr>
<td>ME6</td>
<td>Mixed layer depth</td>
<td>10</td>
<td>-0.33</td>
<td>0.15</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Table S2.7. Correlations between protein ortholog module eigengenes and strongly correlated environmental measurements at 25 m. Both positively and negatively correlated OTUs were included in the same module. All environmental parameters that significantly correlated with modules are shown or the environmental parameter with the strongest correlation if no parameter significantly correlated. P-values were adjusted using the BH procedure to account for multiple tests (n=576, 24 environmental measurements (as in table S8) x 24 modules.)

<table>
<thead>
<tr>
<th>Module</th>
<th>Environmental measurement</th>
<th># Protein clusters in module</th>
<th>Correlation (Spearman r)</th>
<th>Single p-value</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Shortwave (solar) irradiance</td>
<td>3854</td>
<td>0.47</td>
<td>0.0335</td>
<td>0.30</td>
</tr>
<tr>
<td>2</td>
<td>Serial sampling day</td>
<td>913</td>
<td>0.74</td>
<td>0.0002</td>
<td>0.02</td>
</tr>
<tr>
<td>2</td>
<td>Total autotroph biomass</td>
<td>913</td>
<td>-0.61</td>
<td>0.0030</td>
<td>0.08</td>
</tr>
<tr>
<td>3</td>
<td>Prochlorococcus cell counts</td>
<td>800</td>
<td>0.47</td>
<td>0.0314</td>
<td>0.30</td>
</tr>
<tr>
<td>4</td>
<td>Picoeukaryote cell counts</td>
<td>753</td>
<td>0.47</td>
<td>0.0309</td>
<td>0.30</td>
</tr>
<tr>
<td>5</td>
<td>Chlorophyll a</td>
<td>646</td>
<td>0.64</td>
<td>0.0016</td>
<td>0.07</td>
</tr>
<tr>
<td>6</td>
<td>Chlorophyll a</td>
<td>459</td>
<td>-0.69</td>
<td>0.0006</td>
<td>0.04</td>
</tr>
<tr>
<td>6</td>
<td>Shortwave (solar) irradiance</td>
<td>459</td>
<td>0.65</td>
<td>0.0019</td>
<td>0.07</td>
</tr>
<tr>
<td>7</td>
<td>Serial sampling day</td>
<td>408</td>
<td>0.67</td>
<td>0.0011</td>
<td>0.06</td>
</tr>
<tr>
<td>8</td>
<td>Mixed layer depth</td>
<td>348</td>
<td>0.81</td>
<td>0.0000</td>
<td>0.04</td>
</tr>
<tr>
<td>8</td>
<td>Chlorophyll a</td>
<td>348</td>
<td>0.70</td>
<td>0.0004</td>
<td>0.03</td>
</tr>
<tr>
<td>9</td>
<td>Total autotroph biomass</td>
<td>258</td>
<td>0.63</td>
<td>0.0023</td>
<td>0.07</td>
</tr>
<tr>
<td>10</td>
<td>Shortwave (solar) irradiance</td>
<td>232</td>
<td>0.64</td>
<td>0.0022</td>
<td>0.07</td>
</tr>
<tr>
<td>11</td>
<td>Silicate</td>
<td>196</td>
<td>0.52</td>
<td>0.0165</td>
<td>0.21</td>
</tr>
<tr>
<td>12</td>
<td>Particulate phosphorus</td>
<td>194</td>
<td>-0.40</td>
<td>0.0688</td>
<td>0.40</td>
</tr>
<tr>
<td>13</td>
<td>Serial sampling day</td>
<td>165</td>
<td>0.78</td>
<td>0.0000</td>
<td>0.01</td>
</tr>
<tr>
<td>14</td>
<td>Chlorophyll a</td>
<td>149</td>
<td>-0.51</td>
<td>0.0188</td>
<td>0.22</td>
</tr>
<tr>
<td>15</td>
<td>Shortwave (solar) irradiance</td>
<td>132</td>
<td>0.71</td>
<td>0.0004</td>
<td>0.03</td>
</tr>
<tr>
<td>15</td>
<td>Chlorophyll a</td>
<td>132</td>
<td>-0.64</td>
<td>0.0017</td>
<td>0.07</td>
</tr>
<tr>
<td>16</td>
<td>Serial sampling day</td>
<td>114</td>
<td>0.41</td>
<td>0.0667</td>
<td>0.40</td>
</tr>
<tr>
<td>17</td>
<td>Low-level nitrate + nitrite</td>
<td>112</td>
<td>0.50</td>
<td>0.0216</td>
<td>0.23</td>
</tr>
<tr>
<td>18</td>
<td>Chlorophyll a</td>
<td>96</td>
<td>0.66</td>
<td>0.0012</td>
<td>0.06</td>
</tr>
<tr>
<td>18</td>
<td>Low-level nitrate + nitrite</td>
<td>96</td>
<td>0.62</td>
<td>0.0026</td>
<td>0.08</td>
</tr>
<tr>
<td>19</td>
<td>Low-level nitrate + nitrite</td>
<td>96</td>
<td>0.71</td>
<td>0.0003</td>
<td>0.03</td>
</tr>
<tr>
<td>20</td>
<td>Chlorophyll a</td>
<td>84</td>
<td>-0.56</td>
<td>0.0076</td>
<td>0.15</td>
</tr>
<tr>
<td>21</td>
<td>Mixed layer depth</td>
<td>63</td>
<td>0.63</td>
<td>0.0028</td>
<td>0.08</td>
</tr>
<tr>
<td>22</td>
<td>Salinity</td>
<td>61</td>
<td>0.47</td>
<td>0.0324</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>Deep chlorophyll maximum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Deep chlorophyll maximum</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>Temperature</td>
<td>47</td>
<td>-0.74</td>
<td>0.0002</td>
<td>0.02</td>
</tr>
<tr>
<td>24</td>
<td>Total autotroph biomass</td>
<td>47</td>
<td>0.73</td>
<td>0.0002</td>
<td>0.02</td>
</tr>
<tr>
<td>24</td>
<td>Particulate carbon</td>
<td>47</td>
<td>-0.64</td>
<td>0.0017</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Figure S2.1. Prochlorococcus (A) and non-pigmented (B) cell densities at Station ALOHA during our study period. The white line outlines the bottom of the mixed layer measured during HOT cruises. Gray dots mark sampling points.
Relative abundance

- 50% Prochlorococcus (49.9/9.7)
- 10% Candidatus Pelagibacter (10.3/3.3)
- 5% SAR116 cluster (2.2/0.8)
- 2.5% SAR86 cluster (2.1/0.9)
- 1.25% unclassified Myoviridae (1.7/0.8)
- 1.25% Roseobacter (1.1/0.3)
- 1.25% alpha proteobacterium HIMB59 (1.0/0.5)
- 1.25% Synechococcus (1.0/0.3)
- 1.25% Candidatus Puniceispirillum (0.8/0.3)
- 1.25% Calothrix (0.8/0.2)
- 1.25% unidentified bacteria (0.7/0.2)
- 1.25% actinobacterium SCGC AAA015-D07 (0.7/0.2)
- 1.25% SAR324 cluster bacterium JCVI-SC AAA005 (0.7/0.2)
- 1.25% uncultured Opitutae bacterium (0.6/0.4)
- 1.25% alpha proteobacterium HIMB114 (0.6/0.2)
- 1.25% SAR406 cluster bacterium SCGC AAA298-D23 (0.5/0.2)
- 1.25% Micromonas (0.2/0.4)

**Figure S2.2.** Area plot displaying relative abundances of the dominant taxonomic groups across time at 25 m in metagenomic libraries. Groups shown comprised at least 1% of one sample. Numbers in parenthesis next to taxa names are the average percent abundance and standard deviation of each group across all 25 m samples. Abundances were normalized to the number of reads in each library with a significant database hit. Taxonomic groups were binned to the genus level when a genus designation was available in NCBI.
Relative abundance

- 50% Cyanobacteria subsection (60.0/10.0)
- 10% SAR11 clade (10.4/3.7)
- 10% SAR406 clade (4.6/2.6)
- 5% OCS155 marine group (2.9/1.6)
- 5% Chloroplast (2.2/1.1)
- 2.5% SAR116 clade (1.8/0.9)
- 2.5% SAR324 clade (1.6/0.7)
- 2.5% Rhodobacteraceae (1.5/0.8)
- 2.5% SAR86 clade (1.3/0.5)
- 2.5% Flavobacteria NS5 marine group (1.2/0.7)
- 2.5% Opitutae MB11C04 marine group (1.1/0.8)
- 2.5% Rickettsiales S25-593 (1.0/0.5)
- 2.5% Flavobacteria NS9 marine group (1.0/0.6)

Figure S2.3. Area plot displaying relative abundances of the dominant taxonomic groups across time at 25 m in the amplicon libraries. Groups shown comprised at least 2% of one sample. Numbers in parenthesis next to taxa names are the average percent abundance and standard deviation of each group across all 25 m samples. Abundances were normalized to total number of reads per sample. Annotations are based on Arb-Silva taxonomy.
**Figure S2.4.** Area plot displaying relative abundances of the dominant taxonomic groups across time at 500 m in metagenomic libraries. Groups shown comprised at least 1% of one sample. Numbers in parenthesis next to taxa names are the average percent abundance and standard deviation of each group across all 500 m samples. Taxa abundances were normalized to the number of reads with significant database hits. Taxonomic groups were binned to the genus level when a genus designation in NCBI was available. SAR406 and SAR324 genomes were condensed into one category for figure.
Figure S2.5. Area plot displaying relative abundances of the dominant taxonomic groups across time at 500 m in amplicon libraries. Groups shown comprised at least 2% of one sample. Numbers in parenthesis next to taxa names are the average percent abundance and standard deviation of each group across all 500 m samples. Abundances were normalized to total number of reads per sample. Annotations are based on Arb-Silva taxonomy.
Comparing Alpha and Beta Diversity using Amplicon versus Metagenomic Sequencing

SSU rRNA amplicon sequencing is among the most commonly used methods for studying microbial diversity in the environment. However, it is becoming more common to use fragmented SSU sequences extracted from shotgun metagenomic data (Caporaso et al., 2010; Bryant et al. 2012; Logares et al., 2013; Sunagawa et al., 2015). We compared alpha and beta diversity measures calculated using SSU rRNA reads generated by these two sequencing approaches to ensure they captured consistent trends despite their methodological differences. We quantified alpha diversity within SSU rRNA genes using the metric termed Phylogenetic Diversity (PD), which is similar to taxonomic richness, but incorporates the phylogenetic relatedness of organisms (Faith 1992). We quantified beta diversity within SSU rRNA genes using the UniFrac Metric, which is similar to the Jaccard Index that quantifies dissimilarity in taxonomic composition between pairs of samples, but also incorporates the phylogenetic relatedness of organisms (Lozupone & Knight 2005).

Comparing functional gene (protein-coding gene) diversity with SSU rRNA-based measures is also useful and relevant, since gene composition ultimately dictates how microbes can potentially interact with each other and their environment (Gilbert 2010; Raes et al. 2011; Fierer et al. 2012; Bryant et al. 2012; Barberán et al. 2012). SSU-based methods have risen to prominence in part because it is hypothesized that phylogenetic relationships between microbial taxa reflect similarities in their genotypic and phenotypic properties, although redundancy at the genome level or horizontal gene transfer could decouple SSU-based measures and functional genes (Olsen et al., 1986, Lozupone and Knight 2005, Zaneveld et al., 2010). Functional (protein-coding sequence) diversity across samples was estimated by binning metagenomic peptide sequences into protein clusters using a de novo clustering approach. We then used these protein clusters to calculate functional alpha diversity with richness, and to calculate functional beta diversity with the Jaccard Index, parallel to our rRNA phylogenetic metrics PD and Unifrac. While multiple protein clusters might sometimes be derived from longer genes that are not spanned by a single read, homologous sequences will still be binned into the same clusters in our analyses. These are consistent with our goal of capturing the relative differences in diversity and composition of protein coding regions rather than absolute numbers of genes in every organism.

PD, Unifrac and the parallel richness and Jaccard Index all depend on the presence, but not abundance of the organisms in the samples. At 25 m, these alpha and beta diversity metrics calculated from amplicon data correlated with the parallel diversity metrics derived from bacterial SSU rRNA metagenomic reads, as well as the functional metagenomic metrics (all comparisons: P<0.01, r>0.46 except amplicon PD vs. functional richness: P< 0.1, Figure S2.6 and S2.7). Stronger correlations resulted when metrics incorporating abundance were used (P<0.001, r>0.55, Figure S2.6 and S2.7). The alpha diversity metric phylogenetic entropy and beta diversity metric weighted Unifrac, both of which incorporate taxa abundance information, were applied to the amplicon and metagenomic bacterial SSU data (Lozupone et al., 2007; Allen et al., 2009). Similarly, the alpha diversity metric
Shannon Index and the beta diversity metric Bray-Curtis distance were applied to functional gene data.

In contrast to 25 m, SSU rRNA diversity metrics derived from 500 m amplicon datasets were not consistently correlated with SSU rRNA or functional diversity measures derived from metagenomic datasets (Figure S2.8 and S2.9). For example, bacterial SSU rRNA amplicon and metagenomic alpha diversity measures were significantly correlated (P<0.05, r=0.60, Figure S2.8). However neither the amplicon versus metagenomic unweighted SSU Unifrac comparison or the majority of the amplicon and functional diversity comparisons were well correlated (Figure S2.8 and S2.9). Both 500 m abundance-based beta diversity comparisons (amplicon versus metagenomic weighted Unifrac and amplicon weighted Unifrac versus functional Bray-Curtis dissimilarity) were at least marginally correlated, suggesting weighted beta diversity measures are less affected by noise within the datasets, than were unweighted beta diversity measures.

The overall lower correlations between amplicon verses metagenomic datasets in the 500 m samples could be a result of greater PCR biases amongst 500 m samples. They also likely reflect a low signal to noise ratio, meaning true variability between samples was being obscured by sequencing noise because variability between microbial communities at different time points at this depth is low. Microbial communities at 500 m were more diverse, and there was less variation between time points at 500 m compared to 25 m (Figure S2.10). This may explain why the correspondence between amplicon and metagenomic indices was greater in the 25 m dataset. Notably, Fierer et al. (2012) found higher correlations between amplicon and metagenomic soil datasets spanning multiple biomes, compared to those observed in our 25 m samples, presumably because there was more variation across soil from different biomes than across our 25 m samples.

Citations:


Figure S2.6. Comparison of bacterial alpha and beta diversity at 25 m, determined from bacterial SSU rDNA amplicon data (X-axes) verse metagenomic reads that mapped to bacterial SSU rDNA genes (Y-axes). Each amplicon dataset was resampled to 9661 reads (see methods). Each metagenomic bacterial SSU rDNA dataset was resampled to 815 reads (see methods). T-tests were used to test for significant Pearson correlations between phylogenetic diversity (A) and phylogenetic entropy (B) values. Mantel tests with Pearson correlation coefficients were used to test for significant correlations between Unifrac (C) and weighted Unifrac (D) values.
Figure S2.7. Comparison of phylogenetic (X-axes) verse functional (Y-axes), alpha and beta diversity measures at 25 m. Phylogenetic diversity values were derived from amplicon data. Functional diversity values were derived from metagenomic reads that mapped to bacterial taxa. Each amplicon and functional gene dataset was resampled to 9661 and 275,525 reads, respectively (see methods). T-tests were used to test for significant Pearson correlations between alpha diversity values (A & B). Mantel tests with Pearson correlation coefficients were used to test for significant correlations between Unifrac (C) and weighted Unifrac (D) values.
Figure S2.8. Comparison of bacterial alpha and beta diversity at 500 m, determined from bacterial SSU rDNA amplicon data (X-axes) verse metagenomic reads that mapped to bacterial SSU rDNA genes (Y-axes). Each amplicon and metagenomic dataset was resampled to 6,909 and 1,182 reads, respectively (see methods). T-tests were used to test for significant Pearson correlations between phylogenetic diversity (A) and phylogenetic entropy (B) values. Mantel tests with Pearson correlation coefficients were used to test for significant correlations between Unifrac (C) and weighted Unifrac (D) values. Graphs with P and r values in the bottom right depict significant or marginally significant relationships. Note, one amplicon and one metagenomic sample were removed from this analysis due to shallow sequencing depths. When these two samples were included and each amplicon and metagenomic dataset was resampled to 548 and 4217 reads respectively, only the weighted Unifrac comparison resulted in a significant correlation.
**Figure S2.9.** Comparison of phylogenetic (X-axes) verse functional (Y-axes), alpha and beta diversity at 500 m. Phylogenetic diversity values were derived from amplicon data, resampled to 6,909 reads per samples (see methods). Functional diversity values were derived from metagenomic reads that mapped to bacterial taxa and were resampled to 515,411 reads (see methods). T-tests were used to test for significant Pearson correlations between alpha diversity values (A & B). Mantel tests with Pearson correlation coefficients were used to test for significant correlations between beta diversity values (C & D). Note, one amplicon and one metagenomic sample were removed from this analysis due to shallow sequencing depths. When these two samples were included and each amplicon and metagenomic dataset was resampled to 4,217 and 187,195 reads respectively, no comparisons resulted in significant correlations.
Figure S2.10. Boxplots comparing the distribution of alpha (A,B,E,F) and beta (C,D,G,H) diversity values between 25 and 500 m samples. Top row displays presence/absence-based metrics. Bottom row displays abundance-based metrics. The boxes outline the interquartile range with internal black line designating the median. Whiskers delineate the full range of values. All amplicon data were resampled to 6906 reads and all metagenomic functional gene data were resampled to 393,321 reads (see methods). P-values indicate that 500 m samples have a significantly lower variance than 25 m samples based on the Levene's test (R lawstat package).
Figure S2.11. Rarefaction curves comparing phylogenetic diversity between 25 m (in blue) and 500 m (in orange) samples at different sequencing depths. Curves were generated using bacteria SSU amplicon data.
Figure S2.12. The variance of 25 m alpha diversity across samples explained by average wind speeds while including differing numbers of days before sample collection to calculate average wind speeds (x-axes). Each line displays the coefficient of determination ($r^2$) for simple linear regression models of alpha diversity measures with wind speed: red solid line: phylogenetic diversity of metagenomic reads that mapped to SSU rDNA genes (rarefaction resampled to 892 reads), blue dotted line: phylogenetic diversity from amplicon data (resampled to 9,661 reads), orange long dash - short dash: functional richness from metagenomic data (resampled to 476,661 reads).
Figure S2.13. OTU abundance distribution of 25 m samples collected during high and low wind speed periods. Bars show the number of OTUs (y-axis) with a specific read abundance (x-axis) in bacterial SSU amplicon libraries. Libraries were repeatedly resampled to 9,661 reads and the number of OTUs (y-axis) in each abundance bin (x-axis) in each re-sample was averaged. Orange and blue bars show the average of the four time points with the highest and lowest wind speeds, respectively. Black bars indicate standard deviation of four samples.
Figure S2.14. The relationship between chlorophyll $a$ and average wind speeds. A) The relationship between average wind speed of the 4 days prior to sampling and chlorophyll $a$ concentrations at 25 m during the time span of our study. The influence of photoadaption was controlled for by first using linear regression to model chl $a$ concentrations with the average solar radiation of the previous 30 days (y-axis). B) The relationship between average wind speed of the 4 days prior to sampling and chlorophyll $a$ concentrations at 25 m during HOT cruises from 1989-2009. The influence of photoadaption was controlled for
by first using linear regression to model chl $a$ concentrations and day length as an estimate for incoming light. C) The relationship between average wind speed of the 4 days prior to sampling and the sum of *Prochlorococcus* and *Synechococcus* cell concentrations at 25 m during the time span of our study. D) Average wind speed of the 4 days prior to sampling and the sum of *Prochlorococcus* and *Synechococcus* cell concentrations at 25 m during HOT cruises from 1990-2009 (linear regression, $r^2=0.02$, $P=0.09$). E) Average wind speed of the 4 days prior to sampling and total autotrophic biomass at 25 m during the time span of our study (linear regression, $r^2=0.05$, $P=0.31$). F) Average wind speed of the 4 days prior to sampling and total autotrophic biomass at 25 m during HOT cruises from 2004-2009 (linear regression, $r^2=0.051$, $P=0.54$).
Figure S2.15. (previous page) A heat map displaying taxa at 25 m with differential relative abundances between high and low incident solar irradiance samples (see inset in Figure 2.4) based on metagenomic data (DESeq2 with BH correction, P<0.1). Colors show clade relative abundance standardized to the distance from the clade mean. Only taxa with relative abundances across all samples summing to greater than 1% are shown. Log2 fold change followed by the average relative abundance across all samples is shown in parentheses. Negative Log2 fold change values indicate a taxon is more abundant during low solar irradiance time points.
Figure S2.16. (Previous Page) A) Eigengenes (1st principle component) for the 6 modules identified in the weighted network analysis of OTUs identified in the 25 m samples. Of the 193 OTUs analyzed in the 25 m datasets, 171 (89%) could be assigned to modules. B) Eigengenes for the 10 modules identified in the weighted network analysis of OTUs identified in the 500 m samples. Of the 289 total OTUs analyzed from the 500 m samples, only 121 (42%) could be assigned to modules.
Figure S2.17. (Previous Page) A) Eigengenes (1st principle component) for the 24 modules identified from weighted network analysis of ortholog clusters from the 25 m metagenomic datasets. The module number and the number of ortholog clusters assigned to that module are given in the titles. Sampling dates for the HOT cruises are given on the x-axis. B) Bar chart showing the phylogenetic composition of the 24 modules identified in the weighted network analysis. Bar '0' contains protein clusters not assigned to any module.
Chapter 3. A genomic inflection point in the twilight zone of the ocean’s interior

A version of this work is in review for publication at the journal Nature with the following co-authors, as of Jan 21, 2017:

Daniel R. Mendel1 †, Jessica A. Bryant1,2 †, Frank O. Aylward1 †, John M. Eppley1, Torben Nielsen1,3, David M. Karl1, Edward F. DeLong1,2

†These authors contributed equally to the work.

Affiliations:
1 Center for Microbial Oceanography: Research and Education (C-MORE), University of Hawaii, Honolulu, HI, 96822, USA.
2 Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
3 Current address: DOE Joint Genome Institute, Walnut Creek, California 94598, USA.

Abstract
Ocean surface microbiomes differ fundamentally from those found in the ocean’s interior, yet the nature of the transition of organisms and evolutionary forces across vertical ocean gradients are still not well understood. Using a high-resolution depth series replicated across 12 time points, we report that microbial communities partitioned into clear clusters consisting of all 25, 75 and 125 m samples collected above the DCM, 125 m samples below the DCM, all 200 m samples and all 500, 770 and 1000 m samples. The transition across the deep chlorophyll maximum (~ 125 m depth) into the mesopelagic was punctuated with a sharp discontinuity in genomic and proteomic properties. Across this narrow depth horizon, reflective of the interplay between upwelling nutrients and down welling light, phylogenetically diverse microbial clades trended towards larger genome sizes, higher genomic GC content and encoded proteins that have higher nitrogen but lower carbon content. These features changed in tandem with changing characteristics of the surrounding physicochemical environment. In total, these data reflect multispecies, community wide shifts in genomic evolutionary modalities that appear to be driven by changes in the availability of the surrounding energy and nutrient fields.
Main Text

Extensive physical, geochemical and biological differences between the ocean's sunlit layer and the darker, colder waters below have been studied since the very beginnings of ocean exploration (Forbes 1856). Major physicochemical features that distinguish the ocean’s twilight zone (mesopelagic zone, 200-1000 m) from well-lit surface layers include lower light levels, temperature, and oxygen, together with higher hydrostatic pressure and macronutrient concentrations, and diminished seasonal variability. Correspondingly, deep-water microorganisms tend towards chemoorganoheterotrophic and chemolithoautotrophic lifestyles, adaptation to low temperature and high pressure, and higher community taxonomic diversity (DeLong et al., 2006; Swan et al., 2011; Konstantinidis et al., 2009; Bryant et al., 2012; Mizuno et al., 2016). While some differences between surface and deeper water microbes are now predictable (DeLong et al., 2006; Swan et al., 2013; Biller et al., 2015; Giovannoni et al., 2014), the range and extent of adaptive features along the vertical depth continuum is less clear, as are the evolutionary forces that drive them. It is currently unknown for example, whether genotypic and phenotypic features that distinguish surface from deeper water microbial communities accumulate gradually with depth, or whether they might appear in a more punctuated fashion at specific depth strata.

To gain better understanding of evolutionary and ecological forces that shape ocean microbiome genomic and phenotypic characteristics, we conducted a time series survey of picoplankton from the ocean's surface to a depth of 1000 m in the North Pacific Subtropical Gyre at Station ALOHA (22° 45’ N, 158° 00’ W) (Karl and Lukas 1996). Samples were collected at approximately monthly intervals over a 1.5-year sampling period and metagenomic DNA was extracted, sequenced, individually assembled, and annotated using custom workflows (Fig. S3.1, Table S3.1). Individual genes from all samples were consolidated into a non-redundant gene catalog consisting of 8.9 million genes. The Station ALOHA gene catalog was then used to explore the properties, variability and distributional patterns of
genomic properties and gene functions of Station ALOHA oligotrophic ocean microbiomes. Taxonomic distributions over time and space were investigated using sequence clusters of a universally conserved, single copy protein coding marker gene referred to as metagenomic OTUs (mOTUs) as previously described (Sunagawa et al., 2013; Sunagawa et al., 2015).

Planktonic ocean microbiomes clustered primarily by depth, consistent with previous reports of stratified microbial community depth distributions at Station ALOHA (Bryant et al., 2016; DeLong et al., 2006) (Fig. 3.1; Fig. S3.2). This was evident in hierarchical clustering of both mOTUs and aggregate gene abundances (Fig 3.1A; Fig. S3.2), and non-metric multidimensional scaling analysis (NMDS) of SSU rRNA gene sequences (Fig. 3.1C), all of which were highly congruent in their overall patterns. Strikingly, these analyses revealed a marked transition whereby all samples from above the deep chlorophyll maximum (DCM; all 25m and 75m, and three 125m samples), clustered together (Fig 3.1A, C; Fig. S3.2). Samples obtained from deeper waters included three additional clusters that consisted of all those 125m samples found below the DCM, all the 200m samples, and all samples from 500m, 770m and 1000m.

Microbial taxonomic beta diversity reflected a sharp discontinuity at and below the base of the euphotic zone (defined here as the region between the surface and the depth having 1% of the surface 475 nm blue light (Laws et al., 2014)). The largest differences in microbial communities were found at the 75m-125m, 125-200m, and 200-500m depth horizons (Fig. 3.1D). Similar changes were observed in community richness, with the lowest number of different mOTUs found close to the surface, a peak at 125m and 200m and a subsequent drop with increasing depth (Fig. 3.1E, Fig. S3.4). These distinct shifts in microbial diversity coincided with changes in a variety of physical and biogeochemical parameters (Fig 3.1B), including macronutrient availability, salinity, temperature, particulate matter concentration (Hebel and Karl 2001) and light availability (Letelier et al., 2004). Community transitions across the 75-200m depth interval were also accompanied by major changes in genome architecture. The depth stratum extending across the deep chlorophyll maximum (DCM) and into the upper mesopelagic, referred to
herein as the genomic transition zone, GTZ, enveloped fundamental community wide changes in genomic and proteomic properties of the corresponding microbiomes. One fundamental change in genome architecture observed across the GTZ was a sharp discontinuity in the aggregate microbiome GC content. All samples collected above the DCM had lower bulk GC content than all samples collected below the DCM, with a clear transition occurring at approximately 35% GC content, in the zone between 75m to 200m (Fig. 3.2A). This GC trend was partially driven by a distinct preference for low GC codons in surface samples, as evidenced by lower codon diversity (Fig. S3.5). The increased GC content below the DCM was also evident for genes in our universal single-copy gene set (Fig. S3.6) and in SSU rRNA gene fragments, although this was less pronounced. Additional changes in genome architectures in the GTZ included clear shifts in average genome size (Fig. S3.8), the average size of intergenic regions (Fig. S3.9), and the abundance of genes indicative of large genomes (Fig. S3.10) across the 75m to 200m depth horizon. This is consistent with prior reports that many ocean surface water bacterioplankton isolates (e.g. Prochlorococcus (Rocap et al., 2003); Pelagibacter (Giovannoni 2005); SAR86 (Dupont et al., 2012)) have small, low GC content genomes.

*Prochlorococcus* isolates from different depths are known to have a genomic GC contents that range from roughly 30% to 55% (Biller et al., 2015). We observed a steady increase in *Prochlorococcus* gene GC content across the GTZ that was consistent with known trends in *Prochlorococcus* isolates originating from different depths (Biller et al., 2015). Station ALOHA *Prochlorococcus* genes found below 200 m occurred at very low abundance and mapped to abundant surface water populations (mOTUs), indicating they likely originated from sinking euphotic zone cells (or disaggregating larger particles containing *Prochlorococcus*). Our data also revealed similar depth-dependent genome GC content trends in other diverse bacterial and archaeal clades, including *Roseobacter*, Marinimicrobia, SAR116 clades, and across broader taxonomic levels such as Thaumarchaeota (Fig. 3.2b, Fig. S3.11).
In contrast to GC trends observed for other taxa, SAR11 populations appeared to be an exception. Although surface water SAR11 populations had low GC content (<33%), this shifted by less than 2% throughout the GTZ and at greater depths (Fig. S3.11). The genomic properties we observed between deep and shallow water SAR11 populations were consistent with previous studies which found little to no variability in GC content across both marine and freshwater environments (Thrash et al., 2014; Konstantinidis and DeLong 2008; Luo et al., 2015). This suggests that SAR11 populations do not respond to the same selective pressures that drive depth-specific GC content differences in other bacterial clades.

Previous whole genome analyses have indicated that genomic GC content can influence the corresponding elemental composition (carbon and nitrogen content) of the encoded proteome (Bragg and Hyder 2004; Peggy Baudouin-Cornu et al., 2004). In particular, low GC genomes tend to encode proteomes having a lower nitrogen and higher carbon content, while high GC genomes exhibit the opposite trend. In Station ALOHA depth profiles, GC changes observed among aggregate microbial genes across the GTZ corresponded to similar shifts in the elemental composition of encoded proteins (Fig. 3.2C,D, Fig. S3.12, S3.13). Specifically, the average number of nitrogen and carbon atoms per amino acid residue side chains (N-ARSC, C-ARSC) shifted in opposing directions across the GTZ, with high and low C-ARSC values partitioning above and below the DCM, respectively (Figure 3.2C,D). Relative to surface water microbiomes, microbial communities below the DCM at Station ALOHA showed reduced carbon but increased nitrogen in amino acid side chains of the proteome. Similar GC, N-ARSC and C-ARSC trends were also observed from the surface to deep mesopelagic layers Tara Ocean Project samples from the North and South Pacific and Atlantic Oceans (Fig. S3.14), demonstrating that these trends occur throughout the global ocean.

To identify changes in microbiome composition that occur concomitant with whole-genome transitions across the GTZ, we performed a weighted gene correlation network analysis using mOTU abundances. This analysis revealed six primary sets of correlated mOTUs, termed modules, which contain the majority of taxa identified across all depths (Fig. 3.3A). The modules displayed abundance
profiles that reflected transitions across depth similar to those observed in other microbiome features (Fig. 3.1; Fig 3.3A). Module 1 and Module 6 contained mOTUs predominantly abundant in surface waters and deep mesopelagic respectively, while the other modules showed well-defined abundance peaks between the 125-500m depth horizons. Taxonomic composition within these modules captured the well-known transition from high-light to low-light Prochlorococcus between 25m-125m (Rocap et al., 2003) as well as an increased abundance in Thaumarchaeota beginning at 125m (Fig. 3.3B) (Karner, DeLong, and Karl 2001). The different modules also contained distinct populations of SAR11, reflecting finer-scale niche partitioning in this clade across the depth horizon (Field et al., 1997). The presence of SAR324 and Thaumarchaeota mOTUs across different modules also suggested these groups contain specific populations residing in specific depth regimes. Overall, the network module variation and composition was consistent with the particularly distinct changes in whole-community GC content across the GTZ.

The region that spans 125m at Station ALOHA represents a dynamic layer, where mixing, internal hydrodynamics, and seasonal light flux influence the inorganic nitrogen availability (Letelier et al., 2004). Consequently, over the course our time series, ambient nitrogen concentrations across the 125m samples varied considerably. Microbial community composition, aggregate microbiome GC content, N-ARSC and C-ARSC varied in tandem with changing inorganic nitrogen at 125m (Fig. 3.4A, Fig. 3.16). Inorganic nitrogen availability was highly correlated with Module 3 of the mOTU network analyses, consisting mainly of Thaumarchaeota, SAR11 and SAR324 ($r^2 = 0.86$; Fig 3.4B). This covariation was visible even for small-scale shifts, reflecting a tight correspondence of the taxon abundances to macronutrient availability. Furthermore, inorganic nitrogen concentrations at 125m depth was correlated (albeit less strongly) with community proteomic nitrogen content (N-ARSC) accounting for a large fraction of N-ARSC variation ($r^2 = 0.46$), suggestive of a direct influence of ambient nitrate concentration on providing an advantage for organisms with certain genomic properties (Fig. 3.4C). The abundance of key genes in nitrogen metabolic pathways also changed markedly across the GTZ. For example, inorganic nitrogen transporters and nitrilases were
over-represented in surface water samples above the GTZ. In contrast, deeper samples within and below the GTZ with higher ambient inorganic nitrogen concentrations exhibited a greater abundance of ammonia and nitrite oxidation genes involved in energy generation. Taken together, these observations implicate nitrogen availability as a key driver of changes in whole-community genomic signatures and community composition at 125m depth, the entry point into the GTZ.

Elucidating the ecological and evolutionary factors that shape microbial genome GC content, size, structure and gene content has emerged as a central theme in comparative microbial genomics (Andersson and Andersson 1999; Tamas et al., 2002; Ochman and Davalos 2006). Multiple studies have shown that habitat effects the genomic GC content of microbial communities (Foerstner et al., 2005; Reichenberger et al., 2015; Grzymski and Dussaq 2012). A number of potential environmental pressures have been proposed to govern GC content, but no universal mechanisms have yet emerged (Agashe and Shankar 2014; Batut et al., 2014). A puzzling recent observation is that some microbial species having strikingly similar global genomic properties, including genome size or GC content, but differ dramatically in their physiologies, life history strategies, population sizes, and habitats (Giovannoni 2005; Batut et al., 2014; Viklund, Ettema, and Andersson 2012). For example, free-living marine cyanobacteria in open ocean surface waters share remarkably similar genomic characteristics with obligate bacterial symbionts of aphids, such as reduced genome size, AT enrichment, and gene loss relative to their most recent common ancestor, and rapid sequence evolution (Batut et al., 2014; Biller et al., 2015). Several different hypotheses have been proposed as drivers of observed microbial genome reduction and its associated features (Batut et al., 2014). These include Muller's Ratchet (Moran, McCutcheon, and Nakabachi 2008), adaptation to nutrient limitation (Grzymski and Dussaq 2012; P. Baudouin-Cornu et al., 2001), high mutation rate (Partensky and Garczarek 2010; Viklund et al., 2012), and the 'Black Queen Hypothesis' (Morris et al., 2012).

A recent analysis comparing the various forces that might drive genomic evolutionary trends in disparate microbial species concluded that nutrient limitation was currently the most parsimonious explanation for genomic trends.
observed in *Prochlorococcus* isolates (Batut et al., 2014). Over evolutionary time scales, microbial genomic and proteomic features (GC content, N-ARSC, C-ARSC) are likely to become “tuned” to specific niches with respect to nitrogen and other factors including energy availability. Such trends could be further reinforced by the absence and presence of specific proofreading genes (Dufresne et al., 2005; Garcia-Gonzalez et al., 2012).

In aggregate, our observations of indigenous microbiome genomic properties at Station ALOHA support the idea that the presences of individual microbial lineages ‘tuned’ to different levels of nitrogen availability fluctuate at 125 m depending on the depth of the DCM and associated nitrocline. In oligotrophic ocean surface waters, relative to waters just tens of meters deeper, nitrogen sparing evolutionary strategies seem to be a major driver of genomic GC content and amino acid sequence compositional trends among a variety of different microbial clades. The average genome size differences we observed across the GTZ would account for ~40% savings in nitrogen used for DNA replication, and alterations in GC content could further add 1.2% to these nitrogen savings. Nitrogen savings due to N-ARSC reduction in the proteome are expected to be sizeable, given the nitrogen content of the cellular proteinaceous biomass, compared to the nitrogen content found in chromosome. Quantifying the relative nitrogen sparing in the proteome is more difficult however, since this depends in part on the magnitude and variability of protein expression.

In nutrient poor habitats, microbial species employ a variety of strategies to minimize cellular demand for limiting nutrients. For example, in phosphate poor habitats phospholipids are replaced with non-phosphorus lipid substitutes (Van Mooy et al., 2009), and iron scarcity can result in iron-sparing changes in protein expression replacement of ferredoxin for flavodoxin (Mackey et al., 2015). The relatively high spatial and temporal resolution of our study has revealed that over just a few hundred meters in depth, microbial communities have optimized their elemental stoichiometry in response to prevailing environmental conditions, resulting in fundamental changes in whole community genomic properties. It is
likely that similar genomic and proteomic elemental adjustments occur in other environmental contexts, and may shed light on universal adaptive features that shape genome evolution in a variety of habitats.

**Figures** (beginning on following page)
Figure 3.1. Quantitative relationships of microbiome genes and taxa and as a function of depth, time and environmental variables at Station ALOHA. A) Bray Curtis dendrogram of mOTU abundances across the time series. Dendrogram representing the hierarchical clustering of 83 microbial communities from 12 different HOT Cruises during 2010-2011, sampled at seven different depths between 25m and 1000m. Dendrogram terminal branch nodes are colored by sampling depth. B) Environmental data represented by heatmaps ranging from blue (low) and yellow (high). Sampling depth displayed in the same colors as in A). C) Non-metric multidimensional scaling (NMDS) plot of OTUs based on small subunit ribosomal RNA gene (miTAG) abundance profiles. Samples colored according to depth as in A). D) Bray-Curtis distances based on mOTU abundances between microbial communities sampled at neighboring depths (25m and 75m; 75m and 125m; 125m and 200m; 200m and 500m; 500m and 770m; 770m and 1000m). Boxplot colors represent the compared depths (Box and border colors). E) mOTU richness calculated from downsampled read counts. Boxplots representing samples from 7 different depths between 25m and 1000m. Colored according to depth as in A).
Figure 3.2. Microbiome GC content, N-ARSC, and C-ARSC versus depth and taxa at Station ALOHA. A) Average GC content of assembled genes in bulk microbial communities. Red triangles indicate 125 m samples collected during periods when the DCM was located below 125 m depth. The vertical gray line highlights the partitioning in GC values in samples located above (left) and below (right) the DCM. B) Average difference in GC content of mOTUs mapping to select taxa at specific depths from the overall mOTU taxa mean across all sample. The three 125 m samples located above the DCM (depicted as triangles in Figure 3.2a were removed to emphasize trends. B) N-ARSC and C) C-ARSC values of all Station ALOHA genes as a function of their corresponding GC content across all samples. Samples to the left of the vertical gray line were collected above, and to the right were collected below the deep chlorophyll maximum (DCM), respectively. Red triangles indicate those 125 m samples that were located below the DCM during at the time of sampling. Samples colored by depth of origin as in Figure 3.1.
Figure 3.3. Distribution of taxon modules above, within and below the genomic transition zone. A) Eigengenes, or first principle components, representing the relative abundance of the six most abundant mOTU modules identified in weighted network analyses. B) Prominent microbial clades represented by the mOTUs in the six modules.
Figure 3.4. Microbiome N-ARSC and taxon abundance versus ambient inorganic nitrogen concentration over time at 125m at Station ALOHA. A) Trendlines for NO$_3^{-}$+NO$_2^{-}$ concentrations, N-ARSC, and mOTU module 3 abundance in the 125m samples showing broad congruence between the three. Samples colored orange on the x-axis correspond to 125m samples occurring above the DCM. B) Regression plots of both N-ARSC and the module 3 eigengene vs. [NO$_2^{-}$+NO$_3^{-}$] demonstrating correspondence between these variables.

References


Acknowledgements

We thank the Captain and crew of the R/V Kilo Moana, and the Hawaii Ocean Time-series marine operations team, for their expert assistance in sample collection and oceanographic data acquisition and analyses at sea. We also thank Tsultrim Palden, Anna Romano, and Paul Den Uyl for their able assistance in DNA library preparation and DNA sequencing. We thank Benedetto Barone and Lance Fujieki for expert advice and assistance in accessing and displaying HOT oceanographic datasets. This research was supported by the Simons Foundation (SCOPE Award ID 329108 to E.F.D. and D.M.K.), the Gordon and Betty Moore Foundation (through Grants GBMF 3777 to E.F.D. and GBMF3794 to D.M.K.), and NSF for support of the HOT program (including the most recent OCE1260164), as well as support to D.R.M. from EMBO (ALTF 721-2015), and the European Commission (LTFCOFUND2013, GA-2013-609409) and support to J.A.B through the US EPA Science to Achieve Results Fellowship. This work is a contribution of the Simons Collaboration on Ocean Processes and Ecology, and the Center for Microbial Oceanography: Research and Education.
Supplementary materials for Chapter 3

Table S3.1. Sequencing and assembly information for each metagenomic sequencing sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sequencer</th>
<th>Number of contigs</th>
<th>Max coverage (total)</th>
<th>Assembly Statistics</th>
<th>Prodigal Gene Calls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N50 contig size</td>
<td>Num. reads</td>
</tr>
<tr>
<td>HOT224_1 0025m</td>
<td>NextSeq</td>
<td>780,250</td>
<td>444</td>
<td>850</td>
<td>68,795,546</td>
</tr>
<tr>
<td>HOT224_1 0075m</td>
<td>NextSeq</td>
<td>675,786</td>
<td>411</td>
<td>860</td>
<td>62,434,378</td>
</tr>
<tr>
<td>HOT224_1 0125m</td>
<td>NextSeq</td>
<td>327,780</td>
<td>449</td>
<td>1083</td>
<td>33,557,304</td>
</tr>
<tr>
<td>HOT224_1 0200m</td>
<td>NextSeq</td>
<td>164,297</td>
<td>134</td>
<td>506</td>
<td>27,557,654</td>
</tr>
<tr>
<td>HOT224_1 0500m</td>
<td>NextSeq</td>
<td>329,910</td>
<td>231</td>
<td>857</td>
<td>39,143,776</td>
</tr>
<tr>
<td>HOT224_1 1000m</td>
<td>NextSeq</td>
<td>579,167</td>
<td>437</td>
<td>962</td>
<td>50,099,672</td>
</tr>
<tr>
<td>HOT226_1 0025m</td>
<td>NextSeq</td>
<td>246,479</td>
<td>163</td>
<td>601</td>
<td>37,740,015</td>
</tr>
<tr>
<td>HOT226_1 0075m</td>
<td>NextSeq</td>
<td>242,771</td>
<td>115</td>
<td>573</td>
<td>32,974,908</td>
</tr>
<tr>
<td>HOT226_1 0125m</td>
<td>NextSeq</td>
<td>592,008</td>
<td>227</td>
<td>943</td>
<td>61,257,440</td>
</tr>
<tr>
<td>HOT226_1 0200m</td>
<td>NextSeq</td>
<td>379,836</td>
<td>164</td>
<td>988</td>
<td>37,462,326</td>
</tr>
<tr>
<td>HOT226_1 0500m</td>
<td>NextSeq</td>
<td>151,367</td>
<td>1047</td>
<td>441</td>
<td>29,131,958</td>
</tr>
<tr>
<td>HOT226_1 0770m</td>
<td>NextSeq</td>
<td>198,519</td>
<td>256</td>
<td>576</td>
<td>27,763,940</td>
</tr>
<tr>
<td>HOT227_1 0025m</td>
<td>NextSeq</td>
<td>179,291</td>
<td>125</td>
<td>599</td>
<td>23,025,498</td>
</tr>
<tr>
<td>HOT227_1 0075m</td>
<td>NextSeq</td>
<td>285,546</td>
<td>207</td>
<td>634</td>
<td>34,099,068</td>
</tr>
<tr>
<td>HOT227_1 0125m</td>
<td>NextSeq</td>
<td>160,109</td>
<td>1413</td>
<td>1760</td>
<td>31,015,782</td>
</tr>
<tr>
<td>HOT227_1 0200m</td>
<td>NextSeq</td>
<td>104,941</td>
<td>318</td>
<td>1612</td>
<td>20,863,908</td>
</tr>
<tr>
<td>HOT227_1 0500m</td>
<td>NextSeq</td>
<td>563,141</td>
<td>286</td>
<td>996</td>
<td>58,673,416</td>
</tr>
<tr>
<td>Sample ID</td>
<td>Flowcell Type</td>
<td>Flowcell ID</td>
<td>Flowcell Size</td>
<td>Flowcell Type</td>
<td>Flowcell ID</td>
</tr>
<tr>
<td>------------</td>
<td>---------------</td>
<td>-------------</td>
<td>---------------</td>
<td>---------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Sample ID</td>
<td>Flowcell</td>
<td>Run Time</td>
<td>Read Count</td>
<td>Basecall Count</td>
<td>Processing Count</td>
</tr>
<tr>
<td>-----------</td>
<td>----------</td>
<td>----------</td>
<td>------------</td>
<td>----------------</td>
<td>------------------</td>
</tr>
<tr>
<td>HOT233_1c_0500m</td>
<td>MiSeq</td>
<td>7,681</td>
<td>23</td>
<td>2119</td>
<td>1,289,392</td>
</tr>
<tr>
<td>HOT233_1c_0770m</td>
<td>NextSeq</td>
<td>538,999</td>
<td>179</td>
<td>901</td>
<td>60,375,417</td>
</tr>
<tr>
<td>HOT233_1c_0125m</td>
<td>NextSeq</td>
<td>505,293</td>
<td>453</td>
<td>932</td>
<td>54,516,104</td>
</tr>
<tr>
<td>HOT233_1c_0200m</td>
<td>NextSeq</td>
<td>410,387</td>
<td>213</td>
<td>564</td>
<td>53,184,084</td>
</tr>
<tr>
<td>HOT233_1 1000m</td>
<td>NextSeq</td>
<td>725,780</td>
<td>391</td>
<td>881</td>
<td>75,101,586</td>
</tr>
<tr>
<td>HOT233_1c_0125m</td>
<td>MiSeq</td>
<td>63,467</td>
<td>214</td>
<td>2826</td>
<td>5,813,396</td>
</tr>
<tr>
<td>HOT233_1c_0770m</td>
<td>MiSeq</td>
<td>68,922</td>
<td>230</td>
<td>2691</td>
<td>6,142,242</td>
</tr>
<tr>
<td>HOT233_2 1000m</td>
<td>MiSeq</td>
<td>631,574</td>
<td>696</td>
<td>992</td>
<td>60,360,120</td>
</tr>
<tr>
<td>HOT233_1 0200m</td>
<td>NextSeq</td>
<td>60,375,417</td>
<td>23,155,136</td>
<td>20,245,470</td>
<td>321,189</td>
</tr>
<tr>
<td>HOT233_1 0500m</td>
<td>NextSeq</td>
<td>570,637</td>
<td>290</td>
<td>875</td>
<td>66,557,784</td>
</tr>
<tr>
<td>HOT233_1 0770m</td>
<td>NextSeq</td>
<td>497,390</td>
<td>231</td>
<td>793</td>
<td>59,799,400</td>
</tr>
<tr>
<td>HOT233_2 1000m</td>
<td>MiSeq</td>
<td>51,322</td>
<td>100</td>
<td>1378</td>
<td>7,099,726</td>
</tr>
<tr>
<td>HOT233_1c_1000m</td>
<td>MiSeq</td>
<td>1,320,004</td>
<td>353</td>
<td>112,435,135</td>
<td>33,593,610</td>
</tr>
<tr>
<td>HOT233_2 0025m</td>
<td>MiSeq</td>
<td>63,467</td>
<td>214</td>
<td>2826</td>
<td>5,813,396</td>
</tr>
<tr>
<td>HOT233_2 0075m</td>
<td>MiSeq</td>
<td>68,922</td>
<td>230</td>
<td>2691</td>
<td>6,142,242</td>
</tr>
<tr>
<td>HOT233_2 0125m</td>
<td>NextSeq</td>
<td>459,728</td>
<td>821</td>
<td>1378</td>
<td>7,099,726</td>
</tr>
<tr>
<td>HOT233_2 0200m</td>
<td>NextSeq</td>
<td>51,322</td>
<td>100</td>
<td>1378</td>
<td>7,099,726</td>
</tr>
<tr>
<td>HOT233_3 0025m</td>
<td>MiSeq</td>
<td>732,288</td>
<td>715</td>
<td>993</td>
<td>69,587,030</td>
</tr>
<tr>
<td>HOT233_1 0025m</td>
<td>NextSeq</td>
<td>477,890</td>
<td>649</td>
<td>930</td>
<td>54,113,650</td>
</tr>
<tr>
<td>HOT233_1 0075m</td>
<td>NextSeq</td>
<td>1,320,004</td>
<td>972</td>
<td>553</td>
<td>112,435,135</td>
</tr>
<tr>
<td>HOT233_1 0125m</td>
<td>NextSeq</td>
<td>738,206</td>
<td>274</td>
<td>563</td>
<td>85,336,087</td>
</tr>
<tr>
<td>HOT233_1 0200m</td>
<td>NextSeq</td>
<td>1,025,242</td>
<td>574</td>
<td>746</td>
<td>117,076,925</td>
</tr>
<tr>
<td>HOT233_1 0500m</td>
<td>NextSeq</td>
<td>751,276</td>
<td>942</td>
<td>617</td>
<td>154,947,358</td>
</tr>
<tr>
<td>HOT233_1 1000m</td>
<td>NextSeq</td>
<td>472,905</td>
<td>1509</td>
<td>709</td>
<td>64,887,236</td>
</tr>
<tr>
<td>HOT233_2 0500m</td>
<td>NextSeq</td>
<td>1,171,819</td>
<td>2530</td>
<td>906</td>
<td>140,048,230</td>
</tr>
<tr>
<td>HOT233_2 0770m</td>
<td>NextSeq</td>
<td>1,236,465</td>
<td>513</td>
<td>985</td>
<td>128,566,144</td>
</tr>
<tr>
<td>HOT234_0500m</td>
<td>MiSeq</td>
<td>1,325,540</td>
<td>234</td>
<td>3159</td>
<td>86,545,714</td>
</tr>
<tr>
<td>HOT234_1c_0025m</td>
<td>MiSeq</td>
<td>61,651</td>
<td>185</td>
<td>2348</td>
<td>7,543,668</td>
</tr>
<tr>
<td>HOT234_1c_0075m</td>
<td>MiSeq</td>
<td>67,563</td>
<td>237</td>
<td>2545</td>
<td>6,314,039</td>
</tr>
<tr>
<td>HOT234_1 0200m</td>
<td>NextSeq</td>
<td>664,892</td>
<td>585</td>
<td>581</td>
<td>78,611,890</td>
</tr>
<tr>
<td>HOT234_1 0500m</td>
<td>MiSeq</td>
<td>614,789</td>
<td>145</td>
<td>2553</td>
<td>43,359,446</td>
</tr>
<tr>
<td>HOT234_1 0770m</td>
<td>NextSeq</td>
<td>805,648</td>
<td>436</td>
<td>1025</td>
<td>78,187,908</td>
</tr>
<tr>
<td>HOT234_1 1000m</td>
<td>NextSeq</td>
<td>1,122,954</td>
<td>2302</td>
<td>926</td>
<td>129,459,766</td>
</tr>
<tr>
<td>Location</td>
<td>Platform</td>
<td>Run Time</td>
<td>Reads</td>
<td>Mean Reads per Sample</td>
<td>Median Reads per Sample</td>
</tr>
<tr>
<td>------------</td>
<td>-----------</td>
<td>----------</td>
<td>-------</td>
<td>-----------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>HOT237_2_0075m</td>
<td>NextSeq</td>
<td>783,099</td>
<td>604</td>
<td>888</td>
<td>23,216,502</td>
</tr>
<tr>
<td>HOT237_2_0500m</td>
<td>MiSeq</td>
<td>610,173</td>
<td>148</td>
<td>2531</td>
<td>17,810,441</td>
</tr>
<tr>
<td>HOT237_2_0770m</td>
<td>MiSeq</td>
<td>216,918</td>
<td>75</td>
<td>2150</td>
<td>5,813,016</td>
</tr>
<tr>
<td>HOT237_2_1000m</td>
<td>MiSeq</td>
<td>12,148</td>
<td>196</td>
<td>1478</td>
<td>2,708,834</td>
</tr>
<tr>
<td>HOT237_3_0025m</td>
<td>NextSeq</td>
<td>877,747</td>
<td>656</td>
<td>879</td>
<td>26,663,298</td>
</tr>
<tr>
<td>HOT237_3_0125m</td>
<td>NextSeq</td>
<td>441,547</td>
<td>341</td>
<td>593</td>
<td>14,459,094</td>
</tr>
<tr>
<td>HOT237_3_0500m</td>
<td>NextSeq</td>
<td>447,696</td>
<td>167</td>
<td>806</td>
<td>18,510,148</td>
</tr>
<tr>
<td>HOT238_1_0025m</td>
<td>NextSeq</td>
<td>489,829</td>
<td>480</td>
<td>773</td>
<td>11,846,887</td>
</tr>
<tr>
<td>HOT238_1_0075m</td>
<td>NextSeq</td>
<td>450,292</td>
<td>472</td>
<td>973</td>
<td>12,214,800</td>
</tr>
<tr>
<td>HOT238_1c_0200m</td>
<td>NextSeq</td>
<td>601,297</td>
<td>185</td>
<td>643</td>
<td>19,820,102</td>
</tr>
<tr>
<td>HOT238_1c_0125m</td>
<td>NextSeq</td>
<td>815,738</td>
<td>839</td>
<td>837</td>
<td>35,157,992</td>
</tr>
<tr>
<td>HOT238_1_0500m</td>
<td>NextSeq</td>
<td>552,739</td>
<td>352</td>
<td>600</td>
<td>24,976,398</td>
</tr>
<tr>
<td>HOT238_1_1000m</td>
<td>NextSeq</td>
<td>362,416</td>
<td>2319</td>
<td>631</td>
<td>87,225,699</td>
</tr>
<tr>
<td>HOT238_2_0500m</td>
<td>NextSeq</td>
<td>935,636</td>
<td>257</td>
<td>877</td>
<td>40,440,501</td>
</tr>
<tr>
<td>HOT238_2_0770m</td>
<td>NextSeq</td>
<td>1,169,738</td>
<td>205</td>
<td>941</td>
<td>72,266,263</td>
</tr>
<tr>
<td>HOT238_2_1000m</td>
<td>MiSeq</td>
<td>1,206</td>
<td>152</td>
<td>884</td>
<td>18,599</td>
</tr>
<tr>
<td>Median Values</td>
<td></td>
<td>505,293</td>
<td>352</td>
<td>888</td>
<td>60,375,417</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>56,599,337</td>
<td>6,2E+09</td>
<td>2.1E+09</td>
<td>4.1E+09</td>
</tr>
</tbody>
</table>
Materials and Methods

Data Acquisition

Metagenomic sequencing data was generated from 83 samples obtained between August 2010 and December 2011 at 7 depths from 25m to 1000m at Station ALOHA on 12 HOT cruises of the Hawaii Ocean Time-series (HOT). Physiochemical data for all cruises is available on the website for the Hawaii Ocean Time-series Data Organization and Graphical System (HOT-DOGS) at http://hahana.soest.hawaii.edu/hot/hot-dogs/.

DNA samples for metagenomic sequencing and physicochemical data were obtained from different hydrocasts. Hence, we matched data from different casts using their potential density to account for internal waves. In short, we interpolated measurements for all physicochemical data. DNA samples obtained in the mixed layer (mixed layer depth calculated using HOT-DOGS applet) were matched to physicochemical data using the sampling depth. Samples from below the mixed layer were matched by using their potential density (instead of depth).

Similarly, sample collection for metagenomes and chlorophyll determinations were performed on separate hydrocasts. In order to determine the position of the metagenomic samples relative to the deep chlorophyll maximum (DCM), we used the seawater potential density of metagenomic samples from each cruise, and calculated which density the chlorophyll cast the sample corresponded to. Next, we determined whether the metagenomic sample occurred above or below the chlorophyll maximum.

Meteorological data including surface wind speed, rain and solar irradiance collected by the Upper Oceans Processes Group at the Woods Hole Oceanographic Institution with the WHOTS buoy located at Station ALOHA were retrieved from http://uop.whoi.edu/projects/WHOTS/whots.html (accessed Feb 2, 2016). Measurements during the 5 and 30 days leading up to sampling, all made at regular intervals, were averaged for subsequent analyses.

All samples used to generate metagenomic data were collected using the following procedure: Twenty liters of seawater were collected and subsequently...
pre-filtered with a 1.6 um, 42.5 mm Whatman GFA filter (Cat. No. 1820-042, Whatman). The filtrate was collected on 0.22um sterivex GV filter (Cat No: SVGV01015, Millipore). Cells were stored in 2 ml sucrose storage buffer (40mM EDTA, 50mM Tris (pH8.3), 0.75 M Sucrose) at -80°C. To lyse cells for DNA extraction, 2 mg/ml of lysozyme was added to sterivex filters and filters were incubated at 37°C for 30 minutes. Final concentrations of 1% SDS and 0.75 mg/ml Proteinase K were subsequently added and the solution was incubated for 2 hours at 55°C. Final DNA purification was performed using the FujiFilm Quick Gene instrument with the QuickGene DNA Tissue Kit (Cat. No DT-L Life Science). Libraries were created using the Illumina TruSeq LT Nano kit set A (PN: FC-121-4001). Sequencing data was generated using an Illumina MiSeq and NextSeq 500 systems (Table S3.1).

Assembly

Raw sequencing data was quality filtered using MIRA v. 4.9.5_2 with the 'qc' and 'pec' options and standard parameters to retain a "high confidence region" (HCR) of every read. This step also includes the removal of contamination by phiX (Chevreux, B., Wetter, T. and Suhai, S., 1999). MIRA was further used to assemble the sequencing data of each sequencing run into contigs using the standard workflow for accurate de novo genome assembly. Assembly quality statistics were summarized in Table S3.2.

ALOHA gene catalog

Genes were predicted from the assembled contigs using Prodigal (Hyatt et al., 2010) and only genes that were predicted to be complete were retained. This yielded a total of 39,436,252 protein-coding genes. We generated a non-redundant gene catalog by clustering this set of genes using CD-HIT (95% nucleotide ID and 90% overlap of the shorter gene) resulting in the ALOHA gene catalog which encompasses 8,966,703 non-redundant gene clusters, each with a single representative sequence used for downstream annotation (Fu et al., 2012; Li and Godzik, 2006).
**Functional and taxonomic annotation of the gene catalog**

We annotated the non-redundant gene catalog using multiple databases. For taxonomic annotations we employed an augmented version of RefSeq release 75(Tatusova *et al.*, 2014), which was amended by a number of high quality SAGs from marine environments. Functional annotations were generated using the KEGG(Tatusova *et al.*, 2014; Kanehisa *et al.*, 2016) and eggNOG databases(Huerta-Cepas *et al.*, 2016c).

For both RefSeq and KEGG annotations, all genes were aligned to the respective databases using LAST version 756(Kielbasa *et al.*, 2011), with scoring parameters "-b 1 -x 15 -y 7 -z 25". Each gene was assigned to the most specific taxon common (lowest common ancestor approach) to all RefSeq hits scoring within 1% of the best hit. Genes were assigned to the KEGG orthologous group or groups represented by all KEGG genes scoring within 5% of the best hit. The HMM-based eggnog-mapper tool(Huerta-Cepas *et al.*, 2016a) was used to obtain eggNOG annotations.

**Clade designations**

Reference genomes in NCBI were classified to well-known clades of marine microbial groups to assist with phylogenetic annotations of genes in the ALOHA catalog. Some previous studies have provided clade-level classifications of sequenced genomes(Swan *et al.*, 2013, 2011), but for many publicly-available genomes their affiliation with previously-established taxonomic groups was uncertain. To resolve these taxon affiliations in reference genomes, a concatenated phylogeny was constructed from a custom reference genome dataset comprised of 480 genomes found to be highly represented in the gene catalog homology search annotations. We extracted the complete set of 40 universal, single-copy marker genes(Ciccarelli *et al.*, 2006; Sorek *et al.*, 2007) from these genomes using fetchMGs(Mende *et al.*, 2013; Sunagawa *et al.*, 2013). These genes were then used to build a phylogenetic tree using the standard FastTree workflow in ete3(Huerta-Cepas *et al.*, 2016b, Price *et al.*, 2010). Initial designations were provided via well-known genomes and previous designations, and subsequent classifications were
propagated to genomes within the same monophyletic clade of the tree. Genes in the ALOHA gene catalog that were taxonomically annotated to one of the clade-designated genomes were assigned to its respective clade.

mOTUs

mOTUs are species-level sequence clusters of a set of near universal, single-copy, protein-coding genes (Mende et al., 2013; Sunagawa et al., 2013). This gene set encompasses multiple orthologous groups, each of which have a different rate of evolution and hence are clustered at the species level at a different sequence identity cutoffs. We established a customized version of the mOTUs using the gene catalog accompanying this publication in combination with genes from the Tara Expedition (Sunagawa et al., 2015) and our custom database of select marine genomes. The fetchMGs tool (Sunagawa et al., 2013) was used to extract the universal, single-copy genes from all datasets. The nucleotide identity between all pairs of orthologous sequences were computed using vsearch (version v1.9.3) and only alignments with more than 20 aligned bases were retained (Rognes et al., 2016). The resulting distance matrix was used to carry out average linkage hierarchical clustering and the mOTU species level clusters were extracted from this using optimized cutoffs (Rognes et al., 2016; Sunagawa et al., 2013). Linking of different orthologous groups was not possible in this dataset, likely due to the high degree of co-abundance between different phylogenetic groups, and we therefore focused our analyses on individual mOTUs belonging to COG0012 (a ribosome associated GTPase). Phylogenetic tree of all COG0012 mOTUs are provided with annotations as Figure S3.15.

mOTUs were designated to clades and other taxonomic levels (such as phyla and genera) using LAST v. 756 (Kielbasa et al., 2011) alignments to the custom version of RefSeq release 75 (Tatusova et al., 2014) described above. The best alignment for each gene sequence of a mOTU was used for taxonomic assignments.

mOTU phylogenetic tree
A phylogenetic tree of all COG0012 mOTUs found at Station ALOHA as well as selected reference genomes was created using ete3 (Huerta-Cepas et al., 2016b). For this purpose, we selected the longest sequence of every COG0012 mOTU cluster as its representative and added COG0012 sequences from 439 of 480 reference genomes (Also see section “Clade designations”). We then used ete3 with the workflow “clustalo_default-trimal01-protest_default-fasttree_full” to build the phylogenetic tree. In more detail: A multiple sequence alignment was generated using Clustal Omega (Sievers and Higgins, 2014), the alignment was trimmed using trimAl (Capella-Gutierrez et al., 2009), optimal parameters for tree generation were calculated using PhyML 3.0 (Guindon et al., 2010) and the phylogenetic tree was generated using FastTree 2. The tree was displayed with clade annotations on iTOL (Letunic and Bork, 2016).

mOTU richness

COG0012 mOTU richness was calculated from read mapping count data (for this purpose only uniquely mapping inserts were used). For each sample, read counts were downsampled to the lowest total read count of all samples using the function rarefy of the R package Vegan (Dixon, 2003). Species accumulation curves for each depth (excluding 125m samples collected above the DCM) were calculated using the specaccum function (Fig. S3.4) and mOTU richness was calculated using the function specnumber, both of the R package Vegan.

miTags and SSU rRNA gene GC content

We assembled a SSU rRNA OTU database, based on the miTag approach described by Logares et al. (Logares et al., 2014), by using usearch v. 8.1 (Edgar, 2010), to cluster all nearly full-length sequences in the SILVA SSU Ref NR99 database (release 123) (Quast et al., 2013; Yilmaz et al., 2013) and then compiled a set of genes that shared less than 97% sequence similarity to one another. SSU rRNA gene fragments were then extracted from our quality-filtered unassembled Illumina datasets using riboFrame (Ramazzotti et al., 2015). Extracted fragments were aligned to our custom SSU rRNA OTU database using bowtie2 with the
parameters "--local", "--very-sensitive" and "-k 100" (Langmead and Salzberg, 2012). Fragments were assigned to their top database OTU when alignment lengths were greater than 70 bps with a sequence identity of at least 97%. Reads aligning to multiple database sequences equally well were assigned to OTUs in equal proportions to unambiguous matches to each OTU in the same sample. Read counts in each sample were randomly downsampled to the sample with the lowest total read counts (6,990 SSU rRNA gene fragments). OTU relative abundances for the subsampled dataset were then used to generate an NMDS plot to visualize Bray-Curtis distances with the metaMDS command in the R package Vegan (Dixon, 2003). GC content of all extracted SSU rRNA gene fragments were used to calculate sample mean SSU rRNA GC for Figure S3.7.

**Mapping/abundance estimation**

Quality-trimmed sequencing reads were aligned to the ALOHA gene catalog using BWA (standard parameters) [Li and Durbin, 2009]. Results were filtered using a 95% ID cutoff and a minimum alignment length of 45 bps using msamtools (Arumugam et al., 2010). For alignments that did not encompass a complete read, a more stringent minimum alignment length of 60 bp was applied. Alignment quality was assessed using BWA alignment scores. If both reads of an insert could be aligned to the same reference a summed alignment score for the insert was calculated. The highest scoring alignment for each insert was kept for abundance counting. Inserts with multiple highest scoring alignments were flagged as multiple mappers. To estimate the abundance of each gene, we first counted all unique alignments to each of the genes (alignments not flagged as multiple mappers). In a second step all multiple mappers were distributed among the different genes according to the abundance profiles of the unique alignments. Gene coverage was calculated by calculating the total number of bases mapping to a gene and then dividing this number by the length of the gene. In order to calculate an average-per-genome-copy number all coverages were divided by the average coverage of all 10 universal, single copy genes (mOTUs) (Sunagawa et al., 2013) found in the same sample.
Hierarchical clustering cladogram

mOTU abundances were calculated as coverage as described and then normalized so the total sum equals to 1. The abundances were then used to calculate Bray-Curtis distances between all pairs of samples using the R package Vegan (Dixon, 2003). The distances were used to compute a complete linkage hierarchical clustering.

Weighted Gene Co-Abundance Network Analysis

Abundance profiles for the COG0012 mOTUs were used to create a weighted gene correlation network using methods similar to those previously described (Aylward et al., 2015; Bryant et al., 2016; Langfelder and Horvath, 2008). Total reads mapped to mOTUs was used as the abundance criterion, and counts were normalized using the variance stabilizing transformation implemented in DESeq2 (Love et al., 2014). A soft-threshold of 3 was chosen based on the scale-free network criterion (Zhang and Horvath, 2005), and modules of co-abundant mOTUs were constructed using the blockwiseModules command (parameters: minModuleSize = 2, mergeCutHeight = 0.25). Module eigengenes, or first principle components, were constructed using the moduleEigengenes command. Networks were visualized using igraph.

Genome Size estimation

We utilized the 10 mOTU genes (near single copy, near universal marker genes) (Sunagawa et al., 2013) to estimate genome sizes. For this purpose, we normalized the total gene abundance in each sample so that the average abundance of the 10 mOTU genes is 1. Hence, the total gene abundance represents an estimate for the average number of genes per genome. The results are displayed in Fig. S3.8.

GC, codon usage, N-ARSC and C-ARSC calculations

For each assembled gene, GC was calculated using sequence utilities within biopython (Cock et al., 2009). The effective codon number, which ranges from 20, where only one codon from each synonymous codon set is used, to 61, where all
synonymous codons for each amino acid are used at even frequencies, was then calculated for each gene (Wright, 1990). To observe preferences for codon GC content we developed a degenerate codon, ranking scheme from 0-1, with 0 and 1 indicating the codon(s) with highest or lowest G+C content, respectively, relative to all codons encoding each amino acid. Codon rankings for each amino acid with degeneracies in the codon table were averaged across each assembled gene. N-ARSC and C-ARSC for each gene was calculated by translating DNA sequences into amino acids (NCBI’s codon table 11 as used by Prodigal), tallying the number of nitrogen and carbon atoms in the side chains of encoded amino acids residues and dividing N and C counts by the amino acid length of the gene. GCs, effective codon number, codon GC ranks, N-ARSC and C-ARSC were averaged for all genes assembled within a sample.

**GC and encoded protein elemental composition (N-ARSC, C-ARSC) in the Tara dataset**

Tara Ocean’s reference gene catalog, gene abundance data and sample environmental data were downloaded from the Tara Ocean’s Project Companion Website (Sunagawa et al., 2015) (http://ocean-microbiome.embl.de/companion.html, accessed 10/15/16). The Tara Ocean’s reference gene catalog contains both full and partial genes. We selected full-length genes from the catalog by screening each sequence for start and stop codons and calculated the GC, N-ARSC and C-ARSC of full-length sequences. Next we calculated the weighted average GC, N-ARSC and C-ARSC values for each sample using the sample abundances (coverage) of the full-length genes reported by Tara. To compare these values between different TARA sampling depths, we selected 13 open ocean samples for which metagenomic data for all three depths (surface, DCM and mesopelagic) was available. (Stations: 38, 64, 72, 76, 78, 98, 100, 102, 112, 132, 133, 138 and 142). The results are shown in Fig. S3.14.

**Functional analysis of nitrogen metabolism**
All functions annotated to the KEGG pathway for "Nitrogen Metabolism" (map00910) (Kanehisa et al., 2016) were extracted from the KEGG ortholog abundance matrix. From these, the average abundance per depth as well as the correlation with Nitrate+Nitrite concentrations was calculated. KEGG orthologs that were either highly abundant (>0.1 average gene copies per genome) or highly correlated or anticorrelated with Nitrate+Nitrite concentrations (Spearman correlation; Bonferroni corrected p-value <0.01) were flagged as central genes in nitrogen metabolism at Station ALOHA. These key KEGG orthologs in Nitrate+Nitrite poor surface and the Nitrate+Nitrite rich deeper water layers were then mapped onto the KEGG pathway for "Nitrogen Metabolism".

**Supplementary Material References**


Fig. S3.1. Summary of workflow from sampling final data products.
Fig. S3.2. Dendrogram of gene abundances across the time series. Complete linkage clustering samples used in this study based on Bray-Curtis distances calculated from gene abundances of the whole gene catalog of 8.9 million genes assembled from station ALOHA. Branch colors indicate depth of sample origin as listed in Fig. 3.1.
Fig. S3.3. Contour plots depicting temporal and vertical distributions of chlorophyll a. Chlorophyll a concentration derived from measurements made over 0m to 200m depth at Station ALOHA from August 2010 to December 2011. Note that whether or not peak chlorophyll concentrations are above or below the 125 m sample collection locations is not well represented in this figure, as chlorophyll and metagenomic samples were matched by corresponding densities, not pressure.
Fig. S3.4. Species accumulation curves generated from mOTUs for samples from each depth. mOTU abundance counts were down-sampled to the lowest sequencing depth before calculating species accumulation curves. Colors indicate depth of sample origin as listed in Fig. 3.1.
Fig. S3.5. Codon diversity by depth. (A) Averaged effective codon number (Nj) of the genes assembled from each sample. Nj values for individual genes range from 20, where only one codon from each amino acid's synonymous codon sets is used; to 61, where all synonymous codons are used at equal frequencies. (B) Degenerate codon usage ranked by GC content. We ranked degenerate codons in each gene from 0-1, with 0 and 1 indicating the codon had the highest or lowest GC content of all codons encoding same amino acid, respectively. Ranks were first averaged across each gene then averaged across all genes assembled in a given sample. Yellow and blue indicate a sample was collected above or below the DCM, respectively. White triangles mark average values of all samples from the same depth.
Fig. S3.6. Average GC content of individual mOTU gene sets. Numbers in subplots indicate that GC values above and below the DCM significantly differed (Wilcoxon rank sum test with Benjamini-Hochberg FDR for multiple comparisons). Colors as indicated in Fig. S3.5.
Fig. S3.7. Average GC content of regions of Illumina reads that mapped to SSU rRNA genes. Colors as indicated in fig. S3.5. GC values above and below the DCM significantly differed (Mann-Whitney test p < 0.001).
**Fig. S3.8. Average genome size estimations.** Estimates are expressed in average number of genes per genome. Colors indicate depth of sample origin as listed in Fig. 1.
Fig. S3.9. **Average length of the non-coding regions located between full-length genes identified on assembled contigs.** Contigs assembled from samples at each depth were combined and only contigs that were longer than 8,000 base pairs and did not contain ribosomal RNA genes were included. Colors indicate depth of sample origin as listed in Fig. 3.1.
Fig. S3.10. Abundance of genes indicative of larger sized genomes. Analyses include genes involved in flagellar assembly as well as mobile genetic elements across different depth based on KEGG annotations. Abundance estimated as genes per genome using genome size normalization. Colors indicate depth of sample origin as listed in Fig. 1.
Fig. S3.11. GC of individual clades. Average GC content of assembled genes that mapped to select taxa at greater than 80% amino acid identity. Samples were only included in plots if at least 100 unique gene catalogue sequences mapping to a given taxa were identified in that sample. Numbers in subplots indicate that GC values above and below the DCM significantly differed (Wilcoxon rank sum test with Benjamini-Hochberg FDR for multiple comparisons). Yellow indicates the sample was collected above the DCM. Blue indicates the sample was collected below the DCM.
Fig. S3.12. N-ARSC of individual clades. Average N-ARSC of assembled genes that mapped to select taxa at greater than 80% amino acid identity. Samples were only included in plots if at least 100 unique gene catalogue sequences mapping to a given taxa were identified in that sample. Numbers in subplots indicate that N-ARSC values above and below the DCM significantly differed (Wilcoxon rank sum test with Benjamini–Hochberg FDR for multiple comparisons). Colors as indicated in Fig. S12.
Fig. S3.13. C-ARSC of individual clades. Average C-ARSC of assembled genes that mapped to select taxa at greater than 80% amino acid identity. Samples were only included in plots if at least 100 unique gene catalogue sequences mapping to a given taxa were identified in that sample. Numbers in subplots indicate that C-ARSC values above and below the DCM significantly differed (Wilcoxon rank sum test with Benjamini-Hochberg FDR for multiple comparisons). Colors as indicated in Fig. S12.
**Fig. S3.14. Genomic trends at oligotrophic TARA sites.** Boxplots displaying the weighted average GC (A), N-ARSC (B) and C-ARSC (C) of oligotrophic open ocean microbial samples collected by the TARA Oceans Project. Samples from surface (yellow), DCM (red) and mesopelagic (blue) layers are shown.
Fig. S3.15. (previous page) mOTU tree. Phylogenetic tree generated from the near-universal single copy gene COG0012 (Ribosome-binding ATPase YchF) extracted from sequences assembled from Station ALOHA metagenomes and selected reference genomes. Taxonomic annotations are provided as color coded circles. The tree can interactively explored at: http://itol.embl.de/tree/128171209201372251480989370.
Fig. S3.16. Spearman Correlations between average GC, N-ARSC, C-ARSC and measured environmental parameters across 125 m samples. Abbreviations as follows: phos: phosphate, sil: silicate, dic: dissolved inorganic carbon, temp: temperature, pc: particulate carbon, pn: particulate nitrogen, pp: particulate phosphorus, csal: salinity, oxy: oxygen, alk: alkalinity. Colors and diameter of circle indicate correlation strength. Blue shades = positive correlations, Red shades = negative correlations. GC and N-ARSC's strongest positive correlations were to NO3+NO2 and dissolved inorganic carbon (spearman rho > 0.55) while C-ARSC's strongest positive correlations were to chlorophyll a and pH (spearman rho > 0.75).
Chapter 4. Diversity and activity of communities inhabiting plastic debris in the North Pacific Gyre

This work has been published with the following co-authors in the manuscript:

Bryant JA\textsuperscript{1,2}, Clemente TM\textsuperscript{2,3}, Viviani DA\textsuperscript{2,3}, Fong AA\textsuperscript{2,3,5}, Thomas KA\textsuperscript{3,6}, Kemp P\textsuperscript{2,3}, Karl DM\textsuperscript{2,3}, White AE\textsuperscript{2,4} and DeLong EF\textsuperscript{2,3} (2016) Diversity and Activity of Communities Inhabiting Plastic Debris in the North Pacific Gyre. \textit{mSystems} 3: e00024-16.

\textsuperscript{1}Department of Civil and Environmental Engineering, 15 Vassar St., Massachusetts Institute of Technology, Cambridge, MA 02139
\textsuperscript{2}Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawaii, Honolulu, HI, USA
\textsuperscript{3}Department of Oceanography, 1000 Pope Rd., University of Hawaii, Honolulu, HI 96822, USA
\textsuperscript{4}College of Earth, Ocean \& Atmospheric Sciences, 104 CEOAS Administration Building, Oregon State University, Corvallis, OR 97331, USA
\textsuperscript{5}Biosciences Division, Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research, Am Handelshafen 12, D-27570 Bremerhaven, Germany
\textsuperscript{6}Department of Earth and Environmental Studies, Department of South Asia Studies, University of Pennsylvania, 820 Williams Hall, 255 S 36\textsuperscript{th} St., Philadelphia, PA 19146

Abstract

Marine plastic debris has become a significant concern in ocean ecosystems worldwide. Little is known however, about its influence on microbial community structure and function. In 2008, we surveyed microbial communities and metabolic activities in seawater and on plastic on an oceanographic expedition through the ‘Great Pacific garbage patch’. The concentration of plastic particles in surface seawater within different size classes (2-5 mm and > 5 mm) ranged from 0.35-3.7 particles m\textsuperscript{-3} across sampling stations. These densities and the particle size distribution were consistent with previous values reported in the North Pacific Ocean. Net community oxygen production (NCP = gross primary production – community respiration) on plastic debris was positive and so net autotrophic, whereas NCP in bulk seawater was close to zero. Scanning electron microscopy and metagenomic sequencing of plastic-attached communities revealed the dominance of a few metazoan taxa, and a diverse assemblage of photoautotrophic and
heterotrophic protists and bacteria. Bryozoa, *Cyanobacteria, Alphaproteobacteria* and *Bacteroidetes* dominated all plastic particles, regardless of particle size. Bacteria inhabiting plastic were taxonomically distinct from surrounding picoplankton, and appeared well adapted to a surface-associated lifestyle. Genes with significantly higher abundances among plastic-attached bacteria included *che* genes, secretion system genes and *nifH* genes, suggesting enrichment for chemotaxis, frequent cell-to-cell interactions and nitrogen fixation. In aggregate, our findings suggest that plastic debris forms a habitat for complex microbial assemblages that have lifestyles, metabolic pathways and biogeochemical activities that are distinct from free-living planktonic microbial communities.
Introduction

In the last decade, there has been a growing concern about the ecological impact of plastics in the marine environment. From 1950 to 2012, the production rates of plastic have increased by an average of 8.7 percent per year, with annual production rates nearing 300 million tons of plastic in 2013 (PlasticsEurope, 2014, 2010). A fraction of this material accumulates in the marine environment. Current estimates of the mass of plastic in the global ocean range from 7,000 - 300,000 tons (Cózar et al., 2014; Eriksen et al., 2014). This debris is found in all ocean basins, albeit not uniformly distributed. In 1988, scientists correctly predicted that buoyant plastic debris entering the ocean would become concentrated in regions termed “gyres” where large scale subtropical currents converge (Day, R.H., Shaw, D.G., Ignell, 1990). This prediction has since been confirmed by multiple sampling efforts spanning the Pacific and Atlantic ocean gyres (Cózar et al., 2014; Moore et al., 2001; Morét-Ferguson et al., 2010; Law et al., 2010, 2014).

While these gyres do not collect cohesive patches or floating islands of refuse, they are certainly zones where plastic debris is observed in elevated concentrations. The most well publicized ‘patch’, the so called ‘Great Pacific garbage patch’, is an accumulation zone roughly centered at 31°N, 139°W (Maximenko et al., 2012) where large-scale anticyclonic (clockwise) ocean circulation acts to trap and retain floating debris (Day and Shaw, 1987; Moore et al., 2001). Despite the increasing research efforts to understanding the spatial distribution and temporal variance of marine plastic debris, the ecological implications of this refuse field are still largely unknown, particularly in regards to the potential consequences for lower tropic levels (e.g. phytoplankton and marine bacteria).

Plastic debris is known to impact marine organisms including turtles, birds, mammals, fish and invertebrates through entanglement and ingestion (Laist, 1997; Wilcox et al., 2013; Derraik, 2002). There is also concern that some types of plastic debris are a source of toxic chemicals and/or adsorb persistent organic pollutants (POPs), including polychlorinated biphenyls (PCBs) that could be biomagnified
throughout the food chain (Mato et al., 2001; Endo et al., 2005; Teuten et al., 2009; Hirai et al., 2011). Additionally, a number of studies have clearly demonstrated that diverse biofouling organisms, such as bryozoans, settle on marine plastic debris (Winston, 1982; Goldstein et al., 2014; Reisser et al., 2014). In this regard plastic can serve as a vector for the introduction of non-native species into new environments (Barnes, 2002; Masó et al., 2003). Small plastic particles, including those called microplastics (generally < 5mm in size but see reference (Andrady, 2011)), may be particularly harmful, given that they are more abundant and that their reduced size makes them ingestible to small grazers that form the lower levels of the marine trophic system (Andrady, 2011).

Despite known impacts of plastic on higher organisms much less is known about the interactions between marine microbiota and plastic (Oberbeckmann et al., 2015). Colonization of plastic particles by microbes was first reported in 1972 (Carpenter et al., 1972). Subsequent studies have shown that microbes rapidly colonize debris and that in the Atlantic Ocean communities on plastic are taxonomically distinct from those in the surrounding water column (Dang et al., 2008; Lobelle and Cunliffe, 2011; Zettler et al., 2013; Oberbeckmann et al., 2014; Harrison et al., 2014). Little is known however about the nature of plastic-microorganism interactions, especially in the context of the entire biofouling community. More significantly, the potential for functional differences between microbes found on plastics and those in the surrounding water column has yet to be explored.

To address these uncertainties and to learn more about the nature of microbes that colonize plastics, we mounted the SUPER HI-CAT expedition to observe and sample plastic debris along a transect through the North Pacific Subtropical Gyre (NPSG) in 2008. We hypothesized that microplastics in the Pacific plastic patch harbor communities that: 1) are metabolically active, with productivity and respiration rates that differ in magnitude from equivalent volumes in the surrounding water column, 2) are taxonomically distinct from free-living picoplankton but similar to plastic-attached communities sampled in the Atlantic.
have different protein coding genes compared to those in the surrounding free-living picoplankton.

**Materials and Methods**

The SUPER HI-CAT (Survey of Underwater Plastic and Ecosystem Response, Hawaii to California Transect) expedition took place aboard the R/V Kilo Moana and transited from Oahu, Hawaii to California between August 25th and September 5th 2008. Hydrographic and biogeochemical data were collected along the expedition route to characterize the upper 150 m of the water column at discrete depths at each station. Water samples for measurements were collected via 10-liter polyvinyl chloride bottles affixed to a conductivity-temperature-density (CTD) rosette sampler. In order to quantify neustonic plastic debris, a manta trawl (Brown and Cheng, 1981) provided by the Algalita Marine Research Foundation with a rectangular opening of 0.9 x 0.15 m, and a 3.5 m long, 333μm mesh net and flow meter was towed off the stern for ~ 90 minutes at a speed of 1-2 knots. Upon recovery of the manta trawl, samples were separated into three different size classes using mesh-lined screens: >5 mm, 2-5 mm and 0.2-2 mm. With the aid of a dissecting scope and forceps, we carefully separated identifiable plastic fragments from any natural particles captured by the trawl. Previous studies have used similar approaches and shown visual inspection with the aid of a dissecting scope is sufficient to discriminate between plastic and natural particles down to at least 1 mm in diameter (Cózar et al., 2014; Goldstein et al., 2013; Song et al., 2015). In addition, the base polymer of a subset of plastic particles from each size class was identified using FTIR by the analytical chemistry consulting company Analytical Answers (Woburn, MA, USA). The one dimensional area of the largest surface and length of individual plastic particles (n=554) was determined using ImageJ software (http://rsb.info.nih.gov/ij/) and the total surface area of each plastic particle was then estimated based on the approximate shape of the particle.
Biotic Measurements

Plastic particle and water column chlorophyll \( a \) measurements were carried out using a Turner Designs model 10-AU fluorometer and the standard protocol used by the Hawaii Ocean Time-series program (http://ahana.soest.hawaii.edu). Plastic particle chlorophyll \( a \) values were then normalized to the surface area of individual plastic particles in order to approximate the relationship between chlorophyll concentrations and plastic particle size. Rates of community metabolism were estimated by utilizing light–dark bottle oxygen production and consumption measurements (Carritt and Carpendter, 1966; Williams et al., 2004). These provided estimates of net community production (NCP) - the balance of oxygen produced and consumed in the light bottle incubations relative to a time zero value; respiration (R) - the total oxygen consumption in the dark bottle incubations relative to a time zero; and gross primary production (GPP) - calculated as NCP + R, the total production of oxygen. For water measurements, seawater was collected from near surface waters (~7 m), placed into a triple rinsed 20 L polycarbonate carboy. Subsamples were incubated as described in Viviani et al. (Viviani et al., 2011). Briefly, twenty-four 125 ml borosilicate iodine bottles were filled with seawater after overflowing 3 full volumes to fully flush out air bubbles. Eight bottles were immediately fixed with Winkler oxygen reagents, 8 bottles were placed in an opaque plastic container for incubation in the dark, and 8 bottles were incubated in the light. Bottles were incubated for 24 hours in surface seawater-cooled incubators, shaded to ~30% surface irradiance.

To assess the community metabolism of organisms associated with plastic particles, 10 to 14 plastic pieces from a given size classes (either >5 mm, 2-5 mm, or 0.2-2 mm) were chosen and placed individually into borosilicate iodine bottles filled with seawater. Plastic amended bottles were then divided and incubated in either light or dark conditions as described above at the same time, using the same seawater as water column metabolic rate determinations. An effort was made to ensure that plastic pieces used were similar in terms of color, approximate size (within each size-class), and presence or absence of visible biofilm. Measured
oxygen concentrations of bottles containing plastic particles were adjusted to take into account the approximate volume of the water displaced by the plastic pieces. To calculate GPP, NCP and R for individual plastic particles, background seawater community rates measured from the un-amended bottles were subtracted from the rates measured in plastic amended bottles.

**SEM**

SEM images were taken in 2015 using plastic particles that were fixed in formalin immediately after collection during the SUPER HI-CAT expedition. Formalin fixed samples were postfixed with 1% OsO4 in 0.1M sodium cacodylate, dehydrated through an ethanol series, and dried in a Tousimis Samdri-795 critical point dryer. Particles were mounted on aluminum stubs and sputter coated with palladium in a Hummer 6.2 sputter coater and viewed with a Hitachi S-4800 Field Emission Scanning Electron Microscope at an accelerating voltage of 5 kV.

**SUPER HI-CAT Library Construction, Sequencing and Annotation**

Immediately after size sorting, individual plastic particles collected for DNA analyses were placed in sterile 2.0 mL microcentrifuge tubes, flash frozen in liquid nitrogen and stored at -80°C until DNA extraction. To begin extractions, samples were defrosted on ice. Mechanical disruption of cells then took place using a bead beating approach where 0.1 g of sterile zirconia beads (Biospec Products, USA) and 200 μL of 1X TE buffer (pH 8.5) were added to each tube and tubes were reciprocated (Fast Prep machine, Bio 101, Carlsbad, CA) at setting 6.0 for a total of 2 minutes (two 45 second run times and one 30 second run time). Afterwards, 350 μL of lysis buffer #1 (final concentration 50 mM Tris, 20 mM EDTA, 1.2% Triton X-100, 10 gL-1 lysozyme, 200 mg L-1 RNase A) was added and samples were shaken at 250 rpm at 37°C for 2 hours. Following this, 200 μL of lysis buffer #2 (final concentration 1% SDS, 1% potassium xanthogenate, 50 mM Tris, 20 mM EDTA, 0.65 gL-1 proteinase K) was added and samples were shaken at 125 rpm for 18 hours at
56°C. Xanthogenate disrupts cyanobacteria cell walls and sequesters metal ions (Tillett and Neilan, 2000). Following this, 600 μL of Buffer AL from the Qiagen DNeasy Blood and Tissue Mini Kit was added, and samples were processed following the kit manufacturer’s instructions for DNA purification of bacterial cells. Two wash steps were performed with Buffer AWI.

DNA concentrations were quantified using a Picogreen assay (Invitrogen, Waltham, MA). Metagenomic libraries were constructed using the Illumina TruSeq library preparation protocol including a 2% PhiX spike-in (Illumina, Inc, San Diego, CA). Each library was first sequenced using the Illumina MiSeq system, to obtain a preliminary assessment of the plastic-associated communities, then sequenced with the Illumina NextSeq500 system, to achieve deeper sequencing depth per library.

For each sequenced library, adaptors were removed using Trimmomatic (Bolger et al., 2014) v. 0.27 (parameters: ILLUMINACLIP::2:40:15) and paired-end reads were then joined using PANDAseq (Masella et al., 2012) v. 2.4 (parameters: -F 6 -t 0.32). The ends of joined reads and reads unable to be paired with quality scores < 5 were clipped and sequences shorter than 40 bases or made up of more than 90% of a single base were discarded. Paired-end reads that did not overlap were joined with ‘NNNNNN’ inserted between them so non-overlapping paired-ends would not be double counted in statistical analyses. Low complexity regions of reads were masked with TanTan (Frith, 2011). Phage PhiX sequences were identified for removal by mapping reads to the PhiX genome using bowtie2 (Langmead and Salzberg, 2012) v. 2.1.0 (parameters: ‘-local ’). After trimming and quality filtering, libraries sequenced with the MiSeq and NextSeq systems contained between 0.5 - 2.4 M and 14.9 – 44.6 M reads, respectively. SortMeRNA (Kopylova et al., 2012) v. 1.7 with databases Rfam (Burge et al., 2013) v. 11.0 and Silva (Quast et al., 2013) release 111 were then used to separate reads into SSU rRNA gene and non-SSU rRNA gene bins and each bin for MiSeq and NextSeq library from the same sample were combined. The total number of reads assigned to SSU rRNA genes and non-rRNA genes are listed in table S2.

Reads identified as containing SSU rRNA genes (herein referred to as SSU rDNA reads) were then queried against the Silva SSU Ref database (Quast et al.,
2013) release 119 using Last version 418 (parameters: "-n 200 -u 2 -Q 1 -s 2 -m 2500"). Bacterial database sequences with a pintail value below 50, indicating a high probability the sequence is chimeric, and bryozoan database sequences that have been previously identified as chimeric or misannotated, were removed from the Silva database (Waeschencbach et al., 2012). In addition, some sample reads mapped to contaminant non-SSU regions at the ends of several database sequences. Reads mapping to these contaminated regions were removed.

For the remaining reads that mapped to the Silva database, all hits with a minimum alignment length of 100 bps, a minimum bit score of 50 and a bit score within 1% of the bit score of the best hit (including the best hit), were retained. Each read was then assigned to the lowest common ancestor (LCA) of those retained hits. For example, if a SSU rDNA read had two high scoring hits, each from the same family, but different genus, that read would be assigned to the common family, and not given a genus or species assignment. The 1% cut off was chosen for LCA assignments to allow for high-resolution taxonomic assignments, while also considering hits that only differed from the query sequence by a few base pairs less than the top hit. SSU rDNA reads mapping to SIVLA chloroplast SSU sequences were also annotated with the PhytoREF database (Decelle et al., 2015) v. 1.1, using the same approach.

To visualize distances between eukaryotic and bacterial communities in each plastic sample, class or family-level SSU rDNA read counts, for eukaryotes or bacteria, respectively, were used to generate non-metric multi-dimensional scaling (NMDS) plots using the metaMDS function with Bray-Curtis distances in the Vegan r package (VEGAN). Reads not able to be assigned to the targeted taxonomic level, were removed from this analysis. Subsequently, counts per clade were normalized by calculating their proportions relative to the total number of SSU rDNA read counts per sample and rounding proportions to the nearest thousandth (eukaryotes) or hundredth (bacteria) decimal place. Values were then square root transformed. Rounding the proportions accounted for variability in total read counts between samples, similar to randomly sub-sampling larger libraries down to the sequencing depth of the smallest library (rarefying; n = 3,560 (eukaryotes) and
n = 600 (bacteria)), but without adding artificial uncertainty (McMurdie and Holmes, 2014).

**Comparison to Plastic Debris from the North Atlantic Subtropical Gyre**

We downloaded bacterial amplicon sequences from the Zettler *et al.* study (Zettler *et al.*, 2013) at the NCBI Sequence Read Archive (SRP026054). SFF files were processed with the following QIIME v. 1.8.0 scripts (Caporaso *et al.*, 2010). We used `process_sff.py` to convert SFF files into FASTA and QUAL files, `split_library.py` (parameters: “-w 50 -r -l 100 -z truncate_only”) to demultiplex reads, `denoise_wrapper.py` and `inflate_denoiser_output.py` to denoise the flowgrams and `identify_chimeric_seqs.py` (parameters: “-m usearch61 -r gg_97_otus4feb2011.fasta”) with `filter_fasta.py` to identify and remove chimeric sequences. Reads were then annotated with the SILA database as described above. To control for varying read lengths and varying taxonomic information across different regions of SSU rRNA genes, we only included reads in the analysis that were able to be assigned a family level clade using the LCA approach.

**Comparing Taxonomic and Bacterial Functional Gene Abundances in Plastic Associated Communities and Surrounding Picoplankton Communities**

Samples used to generate picoplankton (0.22 – 1.6 μm seawater size fraction) 454 metagenomes (NCBI SRA #s SRX556050, SRX556052-SRX556067) were collected at 25 m depth at the Hawaii Ocean Time Series station ALOHA between December 2007 and September 2009 (Bryant *et al.*, 2015). Library preparation and processing have been described previously (Bryant *et al.*, 2015). It has previously been demonstrated the 454 and Illumina platforms sample the same fraction of diversity and produce similar relative abundances of genes and genomes (Luo *et al.*, 2012), although 454 does produce lower read depths per sequencing run (addressed below).

SSU rRNA genes in picoplankton libraries were identified and annotated as described above for the SUPER Hi-CAT datasets. Since the picoplankton samples were sequenced with the older technology, there were fewer SSU rDNA read counts
per samples. We used Mann-Whitney U tests with the Benjamini-Hochberg procedure for false positive rate (FDR) correction (FDR < 0.005), to test for differential abundances of microbial families between the two communities, as recommended for comparing categories of samples with uneven library sizes (Weiss et al., 2015). To account for differing read depths between samples, we calculated the proportion of SSU rDNA read counts to each prokaryotic family in each sample, relative to the total SSU rDNA read counts to all prokaryotic families per sample and rounded these proportions to the hundredth decimal place (approximating rarefying to ~ 200 reads).

In order to target bacterial protein-coding genes, low complexity regions of reads from the SUPER HI-CAT and picoplankton datasets were masked with TanTan (Frith, 2011) and then compared to the NCBI RefSeq database 69 using Last (parameters: “-b 1 -x 15 -y 7 -z 25 -F 15 -u 2 -m 10 -Q 0 ”) including the default 1e-06 E-value cutoff. Reads were considered to originate from bacterial cells if all the best scoring hits, with an alignment length of at least 150 amino acids, were to bacterial genomes. These reads were queried against the KEGG database ((Kanehisa and Goto, 2000) accessed April 4, 2014) with Last (parameters same as above). Reads were assigned to the KEGG Orthology annotation of their top LAST hit. This produced the same results as adding an additional bit score 50 requirement.

To test for KEGG orthology groups (KO) with a greater than 2 log2 fold difference in abundance (log2 fold change > 2, FDR < 0.005) between the picoplankton and SUPER HI-CAT bacterial communities, we used DESeq2 (Love et al., 2014). In brief, the DESeq2 algorithm uses negative binomial generalized linear models to test for differential abundances in count data, and estimates size factors to control for variation in sequencing depth between libraries. It applies the Benjamini-Hochberg procedure to account for multiple comparisons. To confirm that KOs identified as differentially abundant by DESeq2 are not false positives due to uneven library sizes between the picoplankton and SUPER HI-CAT datasets, we also transformed KO read counts to proportions relative to the total number of reads assigned to a KO, rounded proportions to the 5th decimal place (approximating rarefying to ~ 216,000 reads) and applied Mann-Whitney U tests
followed by the Benjamini-Hochberg procedure (FDR < 0.005).

**Nucleotide Sequence Accession Numbers.** Illumina TruSeq and NextSeq500 metagomic libraries generate for this study, were deposited in the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under BioProject number PRJNA318384 with sample accession numbers SRS1401924-SRS1401935 (see Table S4.2).

**Results**

**Concentration and Size Distribution of Plastic**

Plastic fragments were recovered from 14 manta trawls carried out between the Hawaiian Islands and California (Figure 4.1, Table S4.1). A subset of these particles (3 particles in each of the two larger size classes and 4 particles from the 0.2-2 mm size class) was analyzed by Fourier transform infrared spectroscopy (FTIR), and confirmed to be composed of either polyethylene or polypropylene polymers. The concentration of plastic encountered along this transect varied by an order of magnitude. For the two largest size classes sampled by the manta trawl (>2 mm), surface concentrations ranged from 0.35-3.7 fragments per m³ with the highest values (3.71 pieces per m³) recorded at approximately 35°N, which is roughly the position of the subtropical front (Figure 4.1). When integrated over the upper 0.15 m of the water column, neustonic plastic concentrations ranged from 51,000 to 556,500 fragments per km² of sea surface (sum of >2-5 and >5mm). For reference, analysis of existing plastic concentrations in other studies in the NPSG (data from 1972-2012) ranged from 18,160-557,700 pieces per km² (see summary in Law *et al.* 2014 (Law *et al.*, 2014) and Goldstein *et al.* 2013 (Goldstein *et al.*, 2013)). The distribution of plastic particle sizes was reasonably modeled with a power-law scaling exponent in the size bins under 3 mm, beyond which plastic concentrations begin to decrease (Figure 4.2).

**Biotic Activity On Microplastics**
Chlorophyll a (Chl a) measurements combined with oxygen production and respiration measurements demonstrated that metabolically active photosynthetic and heterotrophic organisms were attached to plastic debris (Figure 4.3 and 4.4). Chl a concentrations measured on the plastic debris ranged from approximately 0.03 to 0.42 mg m\(^{-2}\) while Chl a in the surrounding seawater ranged from approximately 0.04-0.1 mg m\(^{-3}\). Assuming the water column Chl a concentrations we measured at each station and the Chl a concentrations on the > 5 mm plastic particles, a spherical plastic particle with a diameter of 5 mm contains the equivalent amount of chlorophyll as approximately 30-700 ml of seawater.

The concentrations of Chl a scaled to surface area were also higher on larger pieces of plastic (Figure 4.3). Chlorophyll concentrations on the three size classes significantly differed (Kruskal-Wallis rank sum test; P < 0.05), and a post hoc Dunn test showed that the > 5mm and 2-5 mm size classes had significantly larger chlorophyll concentrations than the 0.2-2 mm size class (FDR < 0.05).

We estimated bulk community metabolic rates on plastic and in the surrounding seawater in terms of net oxygen production (Net Community Production (NCP)), total oxygen consumption (Community Respiration (R)) and total oxygen production (Gross Primary Production (NCP+R=GPP)). The >5 mm particle size fraction NCP and R rates were significantly higher than the seawater rates (Mann-Whitney U test; p < 0.01). In addition, both experiments using >2-5 mm size fraction pieces had greater R than seawater (One-way ANOVA; p<0.05) and at Station 12, the >2-5 mm size fraction pieces demonstrated greater rates of NCP than seawater (One-way ANOVA; p<0.05). Seawater amendments with the smallest size fraction particles resulted in production and respiration values that were similar to un-amended water samples (data not shown).

**Eukaryotic and Prokaryotic Organisms on Microplastics**

Inspection of > 5mm plastic particles collected from station 2, 14 and 15 using scanning electron microscopy (SEM) revealed that samples were heavily colonized by encrusting bryozoans (Figure 4.5). In particular, the frontal
membranes of the bryozoans were associated with multispecies microbial biofilms that included pennate diatoms as well as coccus, rod and spiral-shaped cells. Bacterial cells with prosthecae and long filaments were also observed on bryozoan surfaces. Similar cell morphologies were seen directly on the surface of the plastic particles, with some cells nested within pores in the plastic.

We extracted DNA from communities attached to 12 plastic particles collected across the oceanographic transect, and analyzed the DNA using metagenomic shotgun sequencing (referred to henceforth as metagenomic samples). Metagenomic sample numbers correspond to sampling station numbers as in Figure 4.1. The ‘a’ and ‘b’ indicate particles from the > 5 mm or 2-5 mm sample size classes, respectively. We used the paired-end reads within our metagenomic libraries, that mapped to small subunit ribosomal RNA genes (SSU rDNA) in the SILVA database, to identify the taxonomic origins of organisms in our samples. Between 40-99% of reads in each sample that mapped to SSU rRNA genes, mapped to eukaryotic SSU rRNAs, with the remainder mapping to bacteria (Figure S4.1). Some, but not all eukaryotic and bacterial communities on plastic particles from the same station clustered together in NMDS plots (Figure S4.2).

Consistent with the SEM images, between 30-90 % of the eukaryotic SSU rDNA reads from all twelve plastic particles mapped to bryozoan rRNA genes (Figure 4.6). In addition, samples 2a, 2b and 15b also harbored a high abundance of polycystine radiolarians and a large percentage of reads from 11a and 11b mapped to Hydrozoa, Maxillopoda and Aphragmophora database sequences. Sample Sb also contained a high abundance of both Dinophyceae and Anthozoa.

Diatom clades did not make up more than 1% of the eukaryotic SSU rRNA genes in any of our metagenomic libraries, despite being evident in SEM images (Figure S4.3) and being frequently abundant on plastic debris in other studies (Carpenter and Smith, 1972; Thiel and Gutow, 2005; Zettler et al., 2013; Carson et al., 2013; Reisser et al., 2014). Their low representation in the metagenomic samples may be due to their low biomass (as opposed to number of individuals) compared to other eukaryotes present. Between 10-50% of reads mapping to chloroplast rRNA genes did map to diatom clades (Figure S4.4). Other chloroplast
sequences mapped to algal classes including *Stylonematophyceae, Filosa-Chlorarachnea* and *Pelagophyceae* (Figure S4.4).

Bacterial SSU rRNA genes revealed that *Cyanobacteria* and *Alphaproteobacteria* were consistently among the most abundant prokaryotic groups on plastic particles. *Flavobacteriia, Cytophagia, Sphingobacteriia, Gammaproteobacteraia* and *Deltaproteobacteria* were also present across all samples (Figure S4.4). Consistent with the filaments observed in the SEM images, higher resolution taxonomic assignments showed that the most abundant *Cyanobacteria* across the 12 samples were most closely related to the filamentous genera *Phormidium, Rivularia, and Leptolyngbya*. *Rhodobacteraceae* and *Hyphomonadaceae*, the latter having prosthecae, appendages like those in Figure 4.5F, were among the most abundant *Alphaproteobacteria* groups (Figure 4.7). The *Bacteroidetes* genera *Tunicatimonas* and *Tenacibaculum* each comprised approximately 10 percent of one sample while *Muricauda* and *Lewinella* were identified at lower abundances across all the samples.

A recent study conducted by Zettler *et al.* (2013) using amplicon sequencing to characterize the bacterial communities on three pieces of polyethylene and three pieces of polypropylene collected in the oligotrophic North Atlantic Subtropical Gyre reported observing similar bacterial taxa (Zettler *et al.*, 2013). We re-annotated their amplicon data using our workflow for our (unamplified) metagenomic rRNA sequences in order to directly compare our SSU rRNA gene data with theirs. *Cyanobacteria* Subsection III Family I group, which includes *Phormidium* and *Leptolyngbya*, and *Rhodobacteraceae* were the most abundant microbial families within and across both studies (Figure 4.8). *Hyphomonadaceae, Flavobacteriaceae, Saprospiraceae* and *Flammeovigaceae* were also consistent members of the microbial plastic communities. Overall, the most abundant groups in the previous study and those reported here were strikingly similar. The one major exception was *Vibrionaceae*, which had a very high abundance on one sample from the Atlantic, but was otherwise not common in samples from either study.
Taxonomic and Functional Gene Comparison to Surrounding Seawater

The bacterial taxa making up the plastic-associated microbial communities were distinct from the previously well-characterized free-living clades found in the surrounding seawater in the NPSG, that are consistently dominated by the oligotrophic Prochlorococcus (Cyanobacteria Subsection I, Family I) and SAR11 clades (DeLong et al., 2006; Giovannoni and Vergin, 2012; Bryant et al., 2015). To illustrate this, we compared the microbial families found in our plastic samples to 17 metagenomic libraries from free-living, picoplankton surface water microbial communities collected in the NPSG over a two-year period bracketing the dates that our plastic samples were collected. Only two clades, Rhodobacteraceae and Flavobacteraceae, made up at least 1% of the reads mapping to SSU rRNA genes in more than half the samples from either community. However, both were significantly more abundant in the plastic-associated communities (Figure S4.6, FDR < 0.005). In addition, bacterial family richness was higher in plastic samples compared to planktonic samples (P<0.001, Figure S4.7).

To better understand how these taxonomic differences correspond to differences in the functional gene repertoires, we compared the abundances of KEGG orthologs (KOs) found in our plastic-attached versus free-living communities. Of the 5,912 KOs tested, 18% (1064) were at least 4 times more abundant in the plastic-attached metagenomes while only 2% (129) were more abundant in the water column (FDR < 0.005, see Supplementary data set in (Bryant et al., 2016)). Of the KOs with significantly different abundances, as determined by the DESeq2 algorithm, only 13 were not detected as significantly different by Mann-Whitney U tests (FDR > 0.005). This in part simply reflects the larger genomes associated with abundant taxa on the plastics, but it also indicates the enrichment of taxa with specific metabolic pathways and genes (secretion systems, nitrogen fixation, motility, etc.) in plastic versus free-living communities (Figure 4.9; see Discussion below).
Discussion

Plastic particle densities across our transect were of the same order of magnitude (52,233 to 556,152 particles per km²) as those previously reported in the North Pacific (Eriksen et al., 2014; Law et al., 2014; Goldstein et al., 2013). In addition our particle size distribution was strikingly consistent with previous observations. Cozar et al. (Cózar et al., 2014) predicted that the fractal nature of plastic fragmentation should result in the smallest plastic size classes having the largest number of particles and that a steady state plastic abundance-size distribution should follow a power-law with a scaling exponent of 3. However, consistent with other studies (Eriksen et al., 2014), Cozar et al. only observed a power-law relationship for size classes greater than approximately 5 mm followed by a steep decline in plastic concentrations in the smaller size classes. We observed the same decline in the smallest size classes. Although we were careful to separate plastic particles from organic matter during sampling, we cannot rule out the possibility that we under-estimated small plastic particles with diameters between 1-3 mm. A recent study also suggested that smaller particles may be distributed across deeper ocean depths compared to larger particles and therefore be captured at lower rates by surface trawls (Reisser et al., 2015). Alternatively, as suggested in previous studies, small plastic particles may be selectively lost from the upper ocean due to unknown processes (Cózar et al., 2014; Eriksen et al., 2014).

The high density of Chl a we observed on plastics, combined with the higher oxygen production measurements compared to the surrounding water column, suggest that larger microplastic particles are creating net autotrophic “hotspots” in the oligotrophic ocean. Determining the exact source of the increased oxygen production and respiration rates is complicated by potentially enhanced activity of plankttonic organisms surrounding the plastic particles during incubations. The biofilms on microplastics are the most likely source, however, given the short incubation times as well as the high density of Chl a and diverse array of eukaryotic and prokaryotic organisms observed on microplastics.
Our findings are consistent with earlier work demonstrating that plastics, in particular microplastics, harbor a distinct biota and represent a new habitat for rafting organisms in the NPSG, especially within accumulation zones (Cózar et al., 2014; Eriksen et al., 2014; Goldstein et al., 2014). The eukaryotic groups we observed in our metagenomic libraries (including Anthozoa, Hydrozoa, Maxillopoda and Aphragmophora) have all been reported in association with rafting communities on either natural (e.g. macroalgae, wood, pumice) or artificial substrates (Barnes, 2002; Thiel and Gutow, 2005; Gregory, 2009; Goldstein et al., 2014). Encrusting bryozoans in particular have been reported as abundant organisms in previous marine debris surveys, including in the NPSG (Winston, 1982; Barnes, 2002; Gregory, 2009; Goldstein et al., 2014; Reisser et al., 2014).

It was intriguing that radiolarians were observed in such high abundances in three of our metagenomic samples, since they have only been visually observed in low abundances in one previous debris study (Carson et al., 2013), but were also observed on plastic debris using molecular approaches by Zettler et al. (2013) (Zettler et al., 2013). This may represent radiolarian “by-catch” in the plankton net tows, as opposed to true association with microplastic particles. Also of interest was the co-occurrence of both Dinophyceae and Anthozoa in sample 5b, suggesting that coral and their photosynthetic dinoflagellate symbionts may sometimes occupy this niche. The closest database hits of reads mapping to Dinophyceae were to *Symbiodinium* species. However the low sequence similarity to many eukaryotic database sequences preclude definitive classification at a higher taxonomic level at this time.

Consistent with previous marine plastic debris work in different systems, the bacterial taxa we observed on plastic particles are strikingly different from the clades known to reside in the surrounding water column (Zettler et al., 2013; Oberbeckmann et al., 2014). In the NPSG, *Prochlorococcus* spp. are the most abundant planktonic cyanobacteria, consistent with Figure S4.6 (Campbell and Vaulot, 1994). In addition, previous studies in tropical and subtropical waters have shown that *Trichodesmium* and *Crocosphaera* represent the major diazotrophic cyanobacteria (Zehr and Kudela, 2011; Thompson and Zehr, 2013). We did not
observe these two taxa in this study, at least in part because they would have been excluded by the sampling prefilters used to generate the planktonic data. Regardless, these and other known open ocean marine cyanobacteria (Thompson and Zehr, 2013) differ from the abundant cyanobacteria genera, *Phormidium*, *Leptolyngbya*, *Prochlorothrix* and *Rivularia* we observed in the plastic microbiota. *Phormidium* and *Rivularia* have also been observed on marine plastic debris in subtropical Atlantic and Northern European waters, as well as mats and benthic environments (Zettler et al., 2013; Oberbeckmann et al., 2014; Paerl et al., 2000). *Prochlorothrix* to our knowledge has previously only been identified in fresh or brackish water (Pinevich et al., 2012). Similarly, *Tunicatimonas*, *Tenacibaculum*, *Hyphomonadaceae*, *Chitinophagaceae*, *Muricauda* and *Lewinella* are not commonly observed as planktonic heterotrophs in the NPSG (DeLong et al., 2006; Giovannoni and Vergin, 2012; Bryant et al., 2015). At the family level *Rhodobacteraceae* and *Flavobacteraceae* were the only abundant clades found in our plastic metagenomic samples that are also reported in open ocean picoplankton communities including the NPSG (Figure S4.6) (Buchan et al., 2005; Gómez-Pereira et al., 2010; Bryant et al., 2015).

The clades we observed associated with microplastics all appear well adapted to take advantage of niches created by surfaces. The abundant bacterial families we found on our microplastic samples, which were separated by as much as 1700 km in the NPSG, were consistent with the clades observed on plastic debris in the Atlantic Ocean. These observations suggest that a predictable, core group of clades occupy the niche created by small plastic debris in oligotrophic surface waters worldwide.

With respect to particular taxonomic groups, members of the *Rhodobacteraceae* clade are known to alternate between diverse lifestyles (e.g. planktonic and attached) and are also capable of rapid response to varying resources (Moran et al., 2004; Polz et al., 2006). It is likely that these characteristics explain their observed frequency as early colonizers on artificial surfaces including glass, polyvinyl chloride and Plexiglas surfaces and explain why we observed them in high abundances across our samples (Dang and Lovell, 2000; Dang et al., 2008;
Likewise, marine Bacteroidetes groups, Flavobacteriaceae and Saprospiraceae have a known preference for growth on particles, surfaces and algae (DeLong et al., 1993; Fernández-Gómez et al., 2013) and Hyphomonadaceae are considered oligotrophs that readily form biofilms on surfaces (Abraham and Rohde, 2014). Together this suggests the microbial communities we observed seem more indicative of general proclivity for surface attachment, as opposed to any specific selection of microbiota by the chemical composition of the plastic substratum itself.

We were unable to determine which bacteria might have been attached to the eukaryotic organisms (like the Bryozoa that can cover much of the substrate surface area) versus directly attached to the plastic substrate. Many of the bacterial groups we observed have been previously documented living in association with eukaryotic organisms including marine invertebrates, corals and sponges (Bourne et al., 2013; Hentschel et al., 2012; Mouchka et al., 2010). The genus Tunicatimonas, which we observed was most abundant on the plastic particle with the highest number of hydrozoan reads, was first isolated from a sea anemone (Yoon et al., 2012). The Flavobacteria Tenacibaculum (also observed by Zettler et al. 2013(Zettler et al., 2013)) and Oberbeckmann et al. 2014(Oberbeckmann et al., 2014) have been isolated off a variety of marine organisms including bryozoans (Heindl et al., 2008) and a few members of this genus are known fish pathogens (Piñeiro-Vidal et al., 2008; Avendaño-Herrera et al., 2006). Lepolyngbya and Phormidium strains have been identified as members of coral black band disease consortium (Myers et al., 2007).

The large number of KEGG orthologs (KOs) that were significantly more abundant in the plastic-associated versus free-living metagenomes provided further evidence that microplastics create a niche that is distinct from niches utilized by the surrounding picoplankton. While it is possible that some differences are a result of differences in the sequencing technologies used to generate data from these two communities, previous studies suggest this is unlikely. For example, a previous study comparing 454 and Illumina platforms in an aquatic system showed both platforms sample the same fraction of diversity and produce similar relative abundances of genes and genomes (Luo et al., 2012). Future gene or protein
expression studies will provide additional information on the potential importance of these genes functions in plastic-associated habitats.

Not surprisingly, our results suggest that the microbial communities that develop on the plastics are enriched for traits necessary for a surface attached lifestyle. Consistent with the stalked cells observed in the SEM image and the presence of *Hyphomonadaceae* across our samples, many of the Caulobacter-like cell-cycle genes that are involved in transitions between a flagellated motile lifestyle, to a sessile cell with a prosthecum, were significantly more abundant in the plastic metagenomes (Abraham and Rohde, 2014) (Figure 4.9). Similarly, methyl-accepting chemotaxis protein and the majority of the two-component system, CheA family (chemotaxis-like) genes, *potABCD* spermidine/putrescine transporter components and KOs within the tight adherence export apparatus system were all more abundant in the plastic-associated metagenomes. These genes have been implicated in chemotaxis, signaling of swarming activity, surface motility, colonization and biofilm formation (Kurihara et al., 2009; Kurihara and Suzuki, 2015; Porter et al., 2011; He and Bauer, 2014; Tomich et al., 2007), (Figure 4.9).

KOs belonging to secretion system pathways, including numerous secretion system IV (T4SS) genes and the majority of type VI secretion system (T6SS) components were also more abundant in the plastic metagenomes. The most common function of T4SSs is to conjugate plasmid DNA and hence T4SS plays an important role in gene flow between cells (Costa et al., 2015). The T6SS transports effector proteins directly into neighboring eukaryotic or prokaryotic cells, thereby playing a key role in competition or pathogenesis (Russell et al., 2014). T4SSs have also been shown to mediate the transfer of toxins and other effector proteins in several pathogens.

Concerns have been raised that plastic debris could transport pathogens or other unfavorable organisms including dinoflagellates that cause harmful algal blooms (Zettler et al., 2013; Masó et al., 2003). The presence of bacterial clades with some pathogenic members has been interpreted by some as evidence that plastic debris may act as a disease vector (Zettler et al., 2013). We cannot however draw any definitive conclusions in this regard. Potentially pathogenic species for example,
frequently contain strains that are benign. Additionally, secretion systems are also used in many other processes including non-pathogenic, non-toxic interbacterial interactions. Finally, little is known about the natural distribution and dispersal mechanisms of many pathogenic and non-pathogenic marine microbes and traits, so it is difficult to postulate how plastic debris impacts these natural processes (Russell et al., 2014; Caporaso et al., 2012).

The significantly higher abundance of phycobilisome antenna protein genes in the plastic metagenomes compared to the increase of some chlorophyll a/b binding light-harvesting protein (LBP) genes in the surrounding water column shows that the dominant cyanobacteria in the two habitats use different light harvesting machinery (Figure 4.9). The majority of cyanobacteria are believed to absorb photons for photosynthesis using phycobilisome complexes while *Prochlorococcus*, the dominant cyanobacteria in the surrounding water column, utilizes chlorophyll-binding complexes. It has been postulated that *Prochlorococcus* evolved the alternative light harvesting mechanism to cope with limited nutrients including iron and nitrogen in oligotrophic gyres (Ting et al., 2002). Assuming this, the prevalence of phycobilisome utilizing cyanobacteria on plastics, in combination with the elevated rates of oxygen production and respiration on plastics relative to background seawater, suggest that the nutrient limitation in the NPSG is less severe in plastic particle communities.

Consistent with this, the increased abundance of nitrogenase genes *nifH*, *nifD* and *nifK* in the plastic-associated metagenomes suggests that nitrogen fixation could be reducing nitrogen limitation on the plastics. Additionally, key enzymes involved in phosphonate utilization were more abundant in plastic communities. Phosphonates are increasingly being recognized as an important source of phosphorus (Dyhrman et al., 2006; Villarreal-Chiu et al., 2012; Martinez et al., 2010; Karl, 2014) and it has been suggested that some microbes use phosphonates when nitrogen fixation relieves nitrogen limitations (Karl 2008). This is consistent with the possibility that oligotrophic conditions are reduced on plastic particles, at least in relation to nitrogen. Future studies focusing on biomass accumulation and
nutrient fluxes on microplastics will clarify the extent to which microplastics are creating a eutrophic niche in oligotrophic waters.

A large number of membrane transporters were significantly more abundant on plastics. Notably, three genes encoding TonB-dependent iron complex outer membrane receptors that import chelated iron and two genes forming an inner membrane iron complex transport system involved in siderophore import were more abundant in plastic metagenomes (Krewulak and Vogel, 2008; Noinaj et al., 2010). Siderophore uptake is important to community dynamics on large marine particles and siderophore biosynthesis is important for biofilm maturation in some taxa (Cordero et al., 2012; Saha et al., 2013).

Whether or not the microorganisms residing on plastic debris are degrading plastics and thereby significantly contributing to the loss of plastic from marine surface waters and in particular to the loss of the small size classes in the particle size distribution (Figure 4.2), is an ongoing question (Osborn and Stojkovic, 2014). It has also been hypothesized that microbial communities associated with plastic debris could be degrading organic pollutants adsorbed on plastic debris, as biofilms are valuable for remediation (Osborn and Stojkovic, 2014; Edwards and Kjellerup, 2013). Similar to observations by Zettler et al. (Zettler et al., 2013), we observed SSU rRNA genes related to bacterial clades with hydrocarbon-degrading members or members that have sometimes been associated with oil contaminated environments including Phormidium, Muricauda, Hyphomonedaeceae and Rhodobacteraceae (Munn, 2004; Hwang et al., 2009; Abraham and Rohde, 2014; Lamendella et al., 2014). It has also been suggested that some Rhodobacteraceae strains isolated from coastal environments are capable of lignin degradation, an activity that is associated with plastic degradation (Buchan et al., 2001; Sivan, 2011).

Several xenobiotic biodegradation genes were more abundant on plastic particles including homogentisate 1,2-dioxygenase, N-ethylmaleimide reductase, a cytochrome P450 and 2,4-dichlorophenol 6-monooxygenase (Figure 4.9). In particular, homogentisate 1,2-dioxygenase is a ring-cleaving enzyme that has been implicated in degrading polycyclic aromatic hydrocarbons (PAHs) as well as styrene (Cao et al., 2015). 2,4-Dichlorophenol 6-monooxygenase is a hydroxylase involved in
the degradation of chlorinated aromatic pollutants (Beadle and Smith, 1982; Ledger et al., 2006). The genes encoding the two subunits of protocatechuate 3,4-dioxygenase, an aromatic ring-cleaning enzyme implicated in lignin degradation, were also observed in plastic metagenome samples and the alpha subunit was significantly more abundant in this niche (Buchan et al., 2001). These data only allow for speculation as to whether the microorganisms residing on plastic debris are degrading plastic, co-metabolizing adsorbed pollutants or are instead primarily relying on carbon and other nutrients accumulated by filter feeding bryozoans, other marine eukaryotes and autotrophic activity.

In the present study, we applied an integrated approach, by focusing on microbial taxonomic and functional composition in the context of the metabolic activity and composition of the entire community. We observed that microplastics create a habitat for metabolically active and net autotrophic communities that may harbor a predictable, core group of microbial clades that are functionally distinct from the surrounding picoplankton community in the water column. Future studies aimed at specifically elucidating how natural microbial assemblages interact with plastic might include excluding multicellular eukaryotes. Alternatively, approaches that differentiate microbes growing directly on plastic surfaces from those co-associated with colonizing eukaryotic organisms, would, help clarify intra-community biotic interactions occurring on microplastics. Further insights will be gained by comparing microbes on plastic to those on natural surfaces in the open ocean such as driftwood, floating algae, plankton, migratory fish and other wildlife. Such future work will be useful to further determine how plastic debris may uniquely impact open ocean communities, processes and dispersal.
Figure 4.1. The locations of sampling stations along our transect. The area of each circle corresponds to the concentration of plastic particles with a diameter greater than 2 mm. Station numbers are written next to the stations where samples used for molecular analyses were collected. Composite satellite SeaWiFS measurements of sea surface chlorophyll a (up to approximately 25 m depth) from August-September 2008 are shown for context. For reference, the center of the NPSG accumulation zone is 31°N, 139°W (Maximenko et al., 2012).
Figure 4.2. The particle size distribution (PSD) of 554 microplastic particles collected in the NPSG in August, 2008. Bins are spaced 0.1 log units apart, and the x-axis represents the upper edge of these logarithmic bins. For reference, bin diameter in mm is also shown. The relationship between particle diameter and particle abundance normalized to bin-width ($A_n$) is characterized by a power law with an exponent ($\xi$) equal to 3. This fit is shown in red. The PSD of microplastic collected in the NPSG adheres to this fit at diameters $> 3$ mm.
Figure 4.3. Chlorophyll \(a\) concentrations on the three size classes of plastic debris and in the surrounding surface water at each station.
Figure 4.4. A) Gross primary production (GPP), net community production (NCP) and B) respiration (R) rates of the communities attached to plastic particles and in the surrounding surface seawater, measured using oxygen fluxes. GPP was calculated as the sum of NCP and R. Bars represent standard deviation.
Figure 4.5. Scanning electron microscopy images of microplastic particle surfaces. Scale bar is located at the bottom right of each image with the value designating the length of the entire scale bar. A) A bryozoan colony on the suface of a plastic particle. B) An individual bryozoan zooid with diatom-shaped organisms attached to its operculum. C) Region of a bryozoan zooid frontal membrane densely covered with cells of various phenotypes. D-G) Cells on the surface of plastic particles.
Figure 4.6. Bar chart displaying the abundance of eukaryotic classes within reads mapping to SSU rRNA genes. Reads were assigned to the lowest common ancestor of top hits to the SILVA database. Clade abundances in each sample are relative to the total number of reads per sample mapping to a eukaryotic SSU rRNA gene. Clades with abundances greater than 1% in at least one sample are shown. The average percent identities of sample reads to their top hit within each taxonomic group are displayed in parentheses.
Figure 4.7. Bar chart showing the relative abundance of prokaryotic groups based on reads mapping to prokaryotic SSU rRNA genes. Reads were assigned to the lowest common ancestor of top hits to the SILVA database. Where possible reads were assigned prokaryotic genera. Broader taxonomic groups are comprised of reads that could not be assigned to a genus or had abundances below 3% in all libraries. Clade counts in each sample were normalized to the total number of SSU rDNA reads mapped to bacterial taxa in each sample. The average percent identities of sample reads to their top hit within each taxonomic group are displayed in parentheses.
Figure 4.8. Heatmap displaying the abundance of bacterial families identified on plastics particles collected from the North Atlantic Subtropical Gyre (NASG, samples PP1-PE3; abbreviations: PP= polypropylene, PE=polyethylene; Zettler et al. 2013) and the North Pacific Subtropical Gyre (NPSG, samples 2a-15b). The abundance of each bacterial family in a sample is relative to the total number of SSU rDNA reads in that sample assigned a prokaryotic family. Families with abundances greater than 1% in at least one sample are shown. The average percent identities of sample reads to their top hit (NASG/NPSG) are displayed in parentheses.
Figure 4.9. Select KEGG genes that were significantly more abundant (log2 fold change > 2, FDR adjusted P < 0.005) in plastic-associated metagenomic libraries compared to the picoplankton community in the surrounding water column. Some gene-encoded products with related functions have been condensed and the log2 fold change averaged (e.g. phnGHIJLMN). A list of all KEGG genes identified in this study and a list of KEGG genes and descriptions included in this figure are available in the Supplementary data set in (Bryant et al., 2016). Abbreviations used for gene descriptions in this figure are as follows: 2-CS: two component system, resp reg: response regulator, OM: outer membrane, SS: secretion system, sys: system, trans: transport.
References


47. Heindl H, Wiese J, Imhoff JF. (2008). *Tenacibaculum adriaticum* sp. nov., from a


65. Luo C, Tseverity D, Kyripides N, Read T, Konstantinidis KT. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same


**Acknowledgments**

This work was supported by grants from the Gordon and Betty Moore Foundation (E.F.D., 3777), and (D.M.K., 3794), the US Environmental Protection Agency STAR Fellowship (J.A.B.), the National Science Foundation (D.M.K and E.F.D., C-MORE; EF0424599) and the Simons Collaboration on Ocean Processes and Ecology (E.F.D. and D.M.K., SCOPE; 329108) and the Alfred P. Sloan Foundation Research Fellowship (A.W.). This work is a contribution of the Center for Microbial Oceanography: Research and Education, and the Simons Collaboration on Ocean Oceanography: Research and Education, and the Simons Collaboration on Ocean
Processes and Ecology. We are indebted to the captain and crew of the research vessel R/V *Kilo Moana* for logistical support, to Tsultrim Palden and Anna Romano for preparing samples for pyrosequencing, to all members of the DeLong lab for valuable comments and to the Algalita foundation for use of their manta trawl.

**Supplementary Materials for Chapter 4**
Table S4.1. Sampling station information and summary of plastic concentrations in the >2-5 mm and > 5 mm size classes at each station in the North Pacific Subtropical Gyre.

<table>
<thead>
<tr>
<th>Station</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Collection Date</th>
<th>2-5 mm, # m⁻³</th>
<th>&gt;5 mm, # m⁻³</th>
<th>Total Areal Abundance, # km⁻¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>STN-2</td>
<td>27° 43.795 N</td>
<td>154° 53.897 W</td>
<td>27-Aug-08</td>
<td>0.22</td>
<td>0.12</td>
<td>51,000</td>
</tr>
<tr>
<td>STN-3</td>
<td>28° 35.042 N</td>
<td>154° 34.317 W</td>
<td>27-Aug-08</td>
<td>0.28</td>
<td>0.07</td>
<td>52,500</td>
</tr>
<tr>
<td>STN-4</td>
<td>31° 12.787 N</td>
<td>153° 29.632 W</td>
<td>28-Aug-08</td>
<td>1.53</td>
<td>0.25</td>
<td>267,000</td>
</tr>
<tr>
<td>STN-5</td>
<td>31° 59.450 N</td>
<td>153° 03.452 W</td>
<td>28-Aug-08</td>
<td>0.85</td>
<td>0.11</td>
<td>144,000</td>
</tr>
<tr>
<td>STN-6</td>
<td>34° 05.050 N</td>
<td>151° 37.916 W</td>
<td>29-Aug-08</td>
<td>0.78</td>
<td>0.11</td>
<td>133,500</td>
</tr>
<tr>
<td>STN-7</td>
<td>34° 28.125 N</td>
<td>151° 19.283 W</td>
<td>29-Aug-08</td>
<td>0.68</td>
<td>0.11</td>
<td>118,500</td>
</tr>
<tr>
<td>STN-8</td>
<td>34° 38.423 N</td>
<td>150° 45.928 W</td>
<td>29-Aug-08</td>
<td>2.3</td>
<td>0.19</td>
<td>373,500</td>
</tr>
<tr>
<td>STN-9</td>
<td>34° 55.675 N</td>
<td>148° 01.812 W</td>
<td>30-Aug-08</td>
<td>1.2</td>
<td>0.31</td>
<td>226,500</td>
</tr>
<tr>
<td>STN-10</td>
<td>34° 59.026 N</td>
<td>147° 24.349 W</td>
<td>30-Aug-08</td>
<td>3.03</td>
<td>0.68</td>
<td>556,500</td>
</tr>
<tr>
<td>STN-11</td>
<td>35° 02.066 N</td>
<td>146° 48.626 W</td>
<td>30-Aug-08</td>
<td>1.19</td>
<td>0.24</td>
<td>214,500</td>
</tr>
<tr>
<td>STN-12</td>
<td>35° 14.797 N</td>
<td>143° 47.303 W</td>
<td>31-Aug-08</td>
<td>1.4</td>
<td>0.25</td>
<td>247,500</td>
</tr>
<tr>
<td>STN-13</td>
<td>35° 18.046 N</td>
<td>142° 47.506 W</td>
<td>31-Aug-08</td>
<td>1.63</td>
<td>0.13</td>
<td>264,000</td>
</tr>
<tr>
<td>STN-14</td>
<td>35° 25.376 N</td>
<td>139° 15.016 W</td>
<td>1-Sep-08</td>
<td>1.41</td>
<td>0.19</td>
<td>240,000</td>
</tr>
<tr>
<td>STN-15</td>
<td>35° 26.331 N</td>
<td>138° 15.861 W</td>
<td>1-Sep-08</td>
<td>1.02</td>
<td>0.12</td>
<td>171,000</td>
</tr>
</tbody>
</table>
Table S4.2. Metagenomic library information.

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Station #</th>
<th>Size Class (mm)</th>
<th>NCBI Sample Accession Numbers</th>
<th>Number of SSU rRNA fragments*</th>
<th>Number of non-rRNA reads*</th>
<th>Number of reads mapping to bacterial coding sequences in RefSeq*</th>
</tr>
</thead>
<tbody>
<tr>
<td>2a</td>
<td>STN-2</td>
<td>&gt;5</td>
<td>SRS1401934</td>
<td>38,691</td>
<td>22,897,674</td>
<td>2,521,340</td>
</tr>
<tr>
<td>2b</td>
<td>STN-2</td>
<td>&gt;2-5</td>
<td>SRS1401935</td>
<td>33,112</td>
<td>27,103,800</td>
<td>3,843,816</td>
</tr>
<tr>
<td>5a</td>
<td>STN-5</td>
<td>&gt;5</td>
<td>SRS1401933</td>
<td>62,001</td>
<td>46,784,460</td>
<td>2,779,836</td>
</tr>
<tr>
<td>5b</td>
<td>STN-5</td>
<td>&gt;2-5</td>
<td>SRS1401932</td>
<td>7,941</td>
<td>31,300,709</td>
<td>687,320</td>
</tr>
<tr>
<td>9a</td>
<td>STN-9</td>
<td>&gt;5</td>
<td>SRS1401931</td>
<td>62,941</td>
<td>45,214,833</td>
<td>1,821,896</td>
</tr>
<tr>
<td>9b</td>
<td>STN-9</td>
<td>&gt;2-5</td>
<td>SRS1401930</td>
<td>30,629</td>
<td>38,649,677</td>
<td>2,216,048</td>
</tr>
<tr>
<td>11a</td>
<td>STN-11</td>
<td>&gt;5</td>
<td>SRS1401929</td>
<td>9,006</td>
<td>15,331,390</td>
<td>2,613,731</td>
</tr>
<tr>
<td>11b</td>
<td>STN-11</td>
<td>&gt;2-5</td>
<td>SRS1401927</td>
<td>15,660</td>
<td>20,055,718</td>
<td>1,559,515</td>
</tr>
<tr>
<td>14a</td>
<td>STN-14</td>
<td>&gt;5</td>
<td>SRS1401926</td>
<td>40,253</td>
<td>46,178,956</td>
<td>3,420,774</td>
</tr>
<tr>
<td>14b</td>
<td>STN-14</td>
<td>&gt;2-5</td>
<td>SRS1401928</td>
<td>50,737</td>
<td>28,251,664</td>
<td>381,636</td>
</tr>
<tr>
<td>15a</td>
<td>STN-15</td>
<td>&gt;5</td>
<td>SRS1401925</td>
<td>19,245</td>
<td>38,613,467</td>
<td>1,860,711</td>
</tr>
<tr>
<td>15b</td>
<td>STN-15</td>
<td>&gt;2-5</td>
<td>SRS1401924</td>
<td>20,959</td>
<td>20,485,893</td>
<td>2,638,598</td>
</tr>
</tbody>
</table>

*Read statistics after quality trimming and paired-end assembly of combined MiSeq and NextSeq500 DNA sequencing data.
Supplementary Figures

Figure S4.1. A bar chart displaying the relative abundance of Eukaryota, Archaea and Bacteria within reads from plastic particle metagenomic libraries mapping to SSU rRNA genes. Reads were assigned to the lowest common ancestor of top hits to the Silva database.
**Figure S4.2.** Non-metric multidimensional scaling plots visualizing Bray-Curtis distances between eukaryotic (A) and bacterial (B) communities. Symbol shape and color indicates the station from which the sample originated. The size of each symbol indicates the size fraction, with the larger symbol indicating the > 5mm class and the smaller symbol indicating the >2-5mm size class. The figure was generated using reads mapping to SSU rRNA genes that could be assigned to the class (A) or family (B) taxonomic level using the LCA approach.
Figure S4.3. Additional diatom SEM images. A) Diatoms nested within a crevice on the surface of a plastic particle. B) Diatoms attached to a bryozoan frontal membrane. C) Diatoms attached to bryozoan tentacles.
Figure S4.4. Bar chart displaying the relative abundance of photosynthetic eukaryotic classes within reads mapping to chloroplast SSU rRNA genes. The PhytoRef database was used for taxonomic assignments and abundances are relative to the total number of SSU rDNA reads in each sample mapping to sequences in the PhytoRef database. The average percent identities of sample reads to their top hit within each taxonomic group are displayed in parentheses in the legend. The percentage of total SSU rDNA reads mapping to chloroplasts in the Silva database is displayed in parentheses next to sample names on the bottom axis.
Figure S4.5: Bar chart displaying the relative abundance of bacterial classes within reads mapping to SSU rRNA genes. Reads were assigned to the lowest common ancestor of their top hits to the SILVA database. Clade abundances in each sample are relative to the total number of reads per sample mapping to bacterial SSU rRNA genes. Taxonomic groups with abundances greater than 3% in at least one sample are shown. The average percent identities of sample reads to their top hits within each taxonomic group are displayed in parentheses.
Cyano Subsection I, Family I (97.4/94.8)*
SAR11 clade (98/96)*
SAR86 clade (98.7/96.8)*
Rhodospirillaceae (97.9/94.9)*
SAR116 clade (98.2/97.9)*
OM1 clade (98.9/100)*
SAR408 clade (97.9/94.6)*
Archaea, Marine Group II (98.4/96.5)*
SAR324 clade (99.2/94.6)*
Rickettsiales, S25–593 (99.2/93)*
Gamma, Family Incertae Sedis (99.1/97.2)*
MB11C04 marine group (95/1/100)
NS9 marine group (98.3/93.2)
NS7 marine group (94.3/90.3)
Punicicoccaceae (99.3/96.2)
PAU/34 (97.9/95.7)
OM182 clade (98.6/95.8)
Oceanospirillaceae (98/92.7)
Archaea, Marine Group III (98.9/NaN)
Coxiellaceae (97.2/94.7)
Vibrionaceae (99.4/96.8)
Rickettsiaceae (96.2/92.1)
Pseudoalteromonadaceae (99.1/96.8)
Bdellovibrionaceae (97.9/95)
Sandaraciniaceae (NaN/93.5)
Cyano Subsection IV, Family I (NaN/94.6)
Truerigaceae (NaN/94.7)
Deferribacteres Incertae Sedis (NaN/92.6)
Hyphomicrobiaceae (NaN/95.8)
Cryomorphaceae (98.3/95.4)
Acidimicrobiaceae (98.7/93.6)
Simkaniaeae (96.4/94.2)
Anaerolineaceae (NaN/94.3)
GR–WP33–58 (98.3/93.7)
Parvularculaceae (NaN/96.4)
Cytophagaceae (NaN/95.1)
Sneathiellaceae (NaN/97.4)
Cyano Subsection I, Family II (NaN/95.6)
JTB255 marine benthic group (NaN/96.1)
Alteromonadaceae (97.8/96.1)
DEV007 (99.1/95.7)*
Rhodothermaceae (NaN/94.2)*
Planctomycetaceae (99.1/96.6)*
Phyllobacteriaceae (99.6/96.5)*
Erythrobacteriaceae (99.9/96.8)*
Physicicaeraceae (97.8/95.5)*
Chlophagaceae (NaN/94.3)*
Saprospiraceae (NaN/95.2)*
Hyphomonadaceae (99.9/97.2)*
Flammeovirgaceae (99.1/95.8)*
Flavobacteriaceae (97.5/96.5)*
Cyano Subsection IV, Family II (96.5/96.4)*
Rhodobacteraceae (97.9/96.5)*
Cyano Subsection II, Family I (NaN/96.4)*

**Frequency**

- **> 0.5**
- **0.25 - .50**
- **0.10 - .25**
- **0.01 - .10**
- **< 0.01**
**Figure S4.6** (previous page). Heatmap comparing abundances of prokaryotic families identified in the North Pacific Subtropical Gyre planktonic (samples H197-H215) and plastic-associated (samples 2a-15b) samples. The abundance values are relative to the total number of bacterial SSU rRNA reads in the sample assigned a prokaryotic family and have been rounded to the hundredth decimal place (unlike in figure 8 in the main text). Families with at least 1% abundance in one sample are shown. The average percent identities of sample reads to their top hit within each family (planktonic/plastic-associated) are displayed in parentheses. * Next to clade names indicates significantly different relative abundances between the two habitats (FDR < 0.005). Abbreviations: Cyano = Cyanobacteria, Gamma = Gammaproteobacteria

**Figure S4.7.** Bar chart displaying the average number of bacterial families identified in North Pacific Subtropical Gyre planktonic and plastic-associated samples. Family abundances in each sample were calculated as described in Figure S4.6 and families were required to have an abundance above 1% to be counted as present in a sample. Plastic-associated samples had significantly more families than NPSG samples. (Welch’s two-sample t-test, P < 0.001). Bars outline 95% confidence intervals.
Chapter 5. Successional dynamics of microbial biofilm communities on plastic surfaces.

Abstract

Plastic debris is an increasingly common feature of marine ecosystems. However, interactions between plastic debris and marine microorganisms, which have foundational roles in all biogeochemical cycles and marine food webs, are poorly characterized. Previous work has demonstrated that bacteria on plastic debris in the North Pacific subtropical gyre are part of a multi-domain community dominated by bryozoans and are taxonomically and functionally distinct from picoplankton communities in surrounding surface waters. In order to better understand how these complex communities develop, we observed colonization of polypropylene and polyethylene films incubated in an outdoor flowing seawater tank over a two-week period. Two days into the experiment, a single bacterial type within the Oleibacter clade rapidly colonized plastic surfaces and comprised up to 80% of the bacteria in a single sample. Seven days into the experiment the Oleibacter largely disappeared and was replaced with a more diverse community composed of diatoms and bacteria. A clear succession was observed, with time appearing to be the largest determinant of bacterial community composition, as samples from each time point were more similar to each other than to samples taken at other time points. Bacterial communities in early time points segregated by plastic polymer type as well, whereas communities in later time points appear to be more strongly shaped by attached photosynthetic eukaryote communities. In addition, bacterial communities on plastics were distinct from communities sampled on other surfaces found in the tank including macroalgae, sediment and large particulates filtered from the water. This work will help guide future marine biofilm succession studies as well studies aimed at understanding the impact of plastic and other persistent surfaces on microbial communities in marine systems.
Introduction

Scientists first reported plastic particles in open ocean surface waters in 1972 (Carpenter and Smith, 1972). Plastic debris has since become recognized as a persistent marine pollutant that can be transported long distances across ocean basins and has been observed in surface waters in every marine gyre, in addition to arctic seas and deep-sea sediments (Barnes and Milner, 2005; Law et al., 2010; Cózar et al., 2014; Woodall et al., 2014). It is estimated that there are currently trillions of plastic particles in surface oceans (van Sebille et al., 2015; Eriksen et al., 2014). In addition the quantity of plastic waste likely to enter the oceans is predicted to increase by an order of magnitude by 2025 (Jambeck et al., 2015). Therefore it is important that scientists understand the potential ecological impacts of this new and rapidly expanding feature of marine ecosystems.

The few plastic debris survey studies conducted to date have observed that microbial communities on marine plastic debris are largely different from the communities in the surrounding water column (Zettler et al., 2013; Amaral-Zettler et al., 2015; Oberbeckmann et al., 2014, 2016; Bryant et al., 2016). Given that microbial communities on natural particulate matter are also distinct from seawater communities, a more appropriate, but largely unexplored comparison would be to examine microbial biofilms on plastic versus natural particulate matter in the same water mass (DeLong et al., 1993; Fontanez et al., 2015; Oberbeckmann et al., 2016). Regardless, the persistent nature of plastic surfaces and the observation that plastic debris in open ocean environments harbors taxa not frequently observed in the open ocean, supports the present view that plastic debris presents a new ecological niche for marine microorganisms in the open ocean (Zettler et al., 2013; Reisser et al., 2014; Bryant et al., 2016).

To date most experiments aimed at characterizing bacteria on plastic debris in surface waters have either focused specifically on Vibrios or studied debris that has been in the environment long enough to form complex multi-trophic biofilms that have likely undergone numerous colonization events (Kirstein et al., 2016; Zettler et al., 2013; Oberbeckmann et al., 2014, 2016). By this time a significant proportion of the
microbes observed are presumably interacting and physically attached to other organisms rather than the plastic surface. Observing early colonization of plastic debris may improve our understanding of more direct bacterial plastic interactions that occur early in colonization and potentially help elucidate the mechanisms and time scales of changing complexity within biofilm communities.

Like all clean surfaces, upon entering the ocean plastic particles are subject to rapid biofouling by prokaryotic and eukaryotic organisms (Dang and Lovell, 2016). Biofilms can transform plastic particles in unexpected ways. For example, it was recently discovered that biofilms on marine plastic can release dimethyl sulfide (DMS), an olfactory cue that causes some clades of seabirds to mistakenly consume plastic debris while foraging (Savoca et al., 2016). Biofouling of clean surfaces is generally described as proceeding in four stages (Dobretsov; Wahl, 1989). Dissolved organic matter adsorbs to surfaces immediately after submersion into an aquatic environment. Next organisms settle on the surface in successional stages. Bacterial cells first attach to surfaces, followed soon afterwards by diatoms and other microscopic eukaryotes and then settlement and subsequent growth of algae and invertebrate larvae including barnacles and mussels. These stages can overlap, particular organisms with in stage can vary and complex colonizing dynamics are likely nested within each phase (Noël et al., 2009).

Only a handful of studies have investigated prokaryotic community diversity, composition and temporal variability during early colonization of abiotic surfaces (Dang and Lovell, 2000; Jones et al., 2007; Lee et al., 2008; Dang et al., 2008; Huggett et al., 2009; Chung et al., 2010). Mainly relying on low-resolution genetic fingerprinting techniques, these studies have observed that Alphaproteobacteria, Gammaproteobacteria and Bacteroidetes were early colonizers of abiotic surfaces including glass, stainless steel, plexiglass, polystyrene and various coatings. It has been argued that members of the Alphaproteobacteria clade are particularly important in promoting surface colonization and recruitment of secondary colonizers, as Alphaproteobacteria are often, although not always the most abundant clades during initial sampling (Lee et al., 2008).
Here we studied microbial colonization of transparent films made of either polypropylene or polyethylene polymers in an outdoor flowing seawater tank over a two-week period. These polymers have the highest demand compared to other plastic polymer types and are commonly used to make household items and food packaging (PlasticsEurope, 2014). We aimed to address the following questions: What are the pioneering taxa on plastic films? How does their composition change over time? Does the type of plastic impact community composition? What other niches in the flowing seawater tank harbor bacterial communities that are similar to the communities on plastic films?

**Methods**

The biofilm experiment took place in an 11,350 L, outdoor flowing seawater tank located at the Kewalo Field Station in Honolulu, HI, USA. At the time of the experiment, water circulating through seawater tanks was being drawn from the Kewalo Basin Harbor channel, adjacent to the lab. The tank contained a coral reef community that has been continuously maintained for over 10 years. *Montipora capitata*, a common coral species around the Hawaiian Islands, is the predominant coral species in the tank. Other organisms in the tank included two fish species *Thalassoma duperrey* and *Kuhlia sandwicensis*, various macroalgae, sea urchins and additional coral species.

We studied biofilm development on transparent 1.5 x 1.5 cm plastic film squares. Plastic squares were placed in six partially submerged square plastic baskets (mini storage trays, Sterilite Corporation, Townsend, MA, USA) held in place by a tether attached to a cable that ran above the tank. Basket openings were 16.2 cm long, 12.7 cm wide and buoyed above the water’s surface using strips of Styrofoam. Baskets were 5.4 cm deep. All baskets had a solid bottom and approximately 1 by 4 cm spaces on the sides through which water flowed freely. The sides of three of the baskets were covered in 105μm nylon mesh (Component Supply Company, Fort Meade, FL, USA) to manipulate the settling community.

1.5 x 1.5 cm squares were cut from 76.3μm (3 Mil) thick, transparent polyethylene film (PE, 100% virgin, high clarity polyethylene, Uline, Pleasant Prairie,
WI, USA) and 76.3μm thick, polypropylene film (PP, 95% virgin polypropylene, 5% FDA compliant undisclosed stabilizers, Uline). We differentiated polymer types by making one small cut on either opposite or adjacent sides of each square. Squares were placed in baskets on February 16, 2016. Six pieces of plastic (3 PE and 3 PP) were then sampled after two, seven and fourteen days for DNA extraction and SEM imaging. Samples used for DNA extractions were placed in 2mL cryotubes and flash frozen in liquid nitrogen immediately after being removed from the tank. Samples were then transported in a liquid nitrogen dewar to the laboratory at University of Hawaii Manoa and stored at -80°C until extraction. Samples used for SEM imaging were secured in racks, suspended in a fixative solution (4% formalin and sterile seawater) and stored at 4°C.

In addition, approximately 24 hours after the experiment began and at the final sampling time point, 2-liter water samples were collected in triplicate for DNA analysis by passing water from the tank through a 105μm prefILTER, then concentrating suspended material on a 0.22μm Supore filter (EMD Millipore Corporation, Billerica, MA, USA). At the final time point, triplicate 105μm prefilters that filtered 20L of water were also saved for DNA analysis in order to sample larger particulates. Samples of the following items found in the tank were also collected DNA analysis at the final time point: Montipora capitata fragments, sediments that had accumulated at the bottom of the tank, sediment that accumulated at the bottom of the baskets and various algae morphologies growing on the sides of the tank.

SEM Imaging
Formalin-fixed samples were post fixed with 1% OsO₄ in 0.1 M sodium cacodylate, dehydrated through an ethanol series, and dried in a Tousimis Samdri-795 critical-point dryer. Particles were mounted on aluminum stubs, sputter coated with palladium in a Hummer 6.2 sputter coater, and viewed with a Hitachi S-4800 Field Emission Scanning electron microscope at an accelerating voltage of 5 kV.
DNA Extractions

Cryotubes containing frozen samples were placed on ice and 400μl of sucrose lysis buffer (40 mM EDTA, 50mM Tris (pH 8.3) and 0.75 M sucrose) and 220 mg of zirconia/silica beads (equal mixture of 0.1 and 0.5 mm beads, Biospec Products, USA) were added to each tube. Cryotubes were then fixed to a vortex and reciprocated at maximum speed for ten minutes. Next 100μl of lysis buffer (as above) with an additional 5-mg/mL lysozyme was added and tubes were incubated for 30 minutes while slowly rotating. 50μl of lysis buffer containing 8mg/mL proteinase K was then added and tubes were incubated for an additional 2 hours at 55°C. Lysate was then transferred to deep 96-well plates and purified with a DNA Saliva Kit using a Chemagic Magnetic Separation Module instrument (Chemagen, Baesweiler, Germany) following the manufacturers protocol. DNA extracts were additionally purified with the MoBio PowerMag DNA Clean-Up Kit (MoBio, Carlsbad, CA, USA).

Barcoded SSU rRNA gene amplicon libraries with Illumina sequencing adaptors were generated by PCR amplifying the V4 region of the SSU rRNA as described by the Earth Microbiome Project protocol (http://earthmicrobiome.org, version 4-2016) with updated primers (Walters et al., 2015). 27 amplification cycles were carried out rather than the recommended 35 cycles. PCR products were visualized on an agarose gel, purified using AMPure beads (Beckman Coulter, Danvers, MA, US) and pooled to a final concentration of 6nM. Amplicons pools were then spiked with 15% PhiX DNA and paired-end sequenced using on an illumine MiSeq platform.

Sequence Processing and Statistical Analyses

Raw reads were quality trimmed using trimmomatic version 0.36 with parameters ILLUMINACLIP::2:30:10:8:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:100 (Bolger et al., 2014). Paired-end reads were then assembled with program Pear version 0.96 (Zhang et al., 2014). Assembled read were then demultiplexed using the script split_libraries_fastq.py within the QIIME package version 1.9 (Caporaso et al., 2010). Putative chimeras were identified and removed with vsearch version 1.11.1 using the green genes reference database gg_13_8_otus provided with QIIME (Rognes et al., 2016). Sequences were then binned into OTUs with 97%
sequence identity and given a taxonomic assignment using the QIIME script pick_open_reference_otus.py which first attempts to assign reads to OTUs in the gg_13_8_otus database, then bins unassigned reads into denovo OTUs.

T2 and T3 amplicon libraries were dominated by chloroplast sequences. Therefore the OTUs mapping to chloroplast (photosynthetic eukaryotes) and non-chloroplast and mitochondria SSU rRNA sequences were split and analyzed as separate datasets. Chloroplast 16s and 18S rRNA gene sequencing have been shown to produce consistent phytoplankton community dynamics (Needham et al., 2016). Over 99.9% of non-organelle SSU rRNA sequences mapped to bacteria and therefore will be referred to as the bacterial dataset. Chloroplast OTUs were additionally annotated with the RDP taxonomy using the least common ancestor algorithm (10 neighbors per query with 0.95 minimum sequence identity) implemented in the ARB-SILVA SINA tool and the SILVA database release 127 (Pruesse et al., 2007).

Rarefaction curves were drawn using the rarecurve command in the R package vegan (Oksanen et al., 2016). To compare richness across samples at a consistent sequencing depth, richness values were the calculated after subsampling communities to 4,225 reads. Next, Krustal-Wallis and Dunn tests were carried out to test for differences in bacterial richness between time points, using the kruskal.test and dunn.test commands in R’s base package (R Development Core Team, 2011).

For bacterial and chloroplast community composition analyses and related hypotheses testing, read counts per OTU were transformed to proportions (OTU count/total OTUs in sample) and square root-transformed. Bacterial OTUs with abundances below 0.1% before taking the square root were considered absent to account for variation detection limits due to differing sequencing depths. For chloroplast samples 1.0% was used as a cutoff.

Non-parametric multivariate analysis of variance (PERMANOVA) was conducted to investigate differences in community composition between time points, mesh, polymer type and corresponding interactions with the adonis command in the vegan package using Bray-Curtis dissimilarities (Oksanen et al., 2016; Anderson, 2001; Anderson and Walsh, 2013). Since replicate plastic pieces were nested within each basket, sample permutations used to calculate the significance of time and polymer type
were constrained to within each basket. The significance of the mesh treatment was calculated by permuting presence or absence of mesh across baskets with the nested.npmanova command in the BiodiversityR package (Anderson, 2001; Kindt and Coe, 2005). For all tests 9999 permutations were performed. Bray-Curtis distances between samples were visualized with 2 dimensional non-metric multidimensional scaling (NMDS) plots. To test whether spearman correlations between bacterial and chloroplast dissimilarities were greater or less than zero, exact p-values were calculated using cor.test in R’s base package.

Results

On February 16, 2016, six baskets containing 30 polyethylene and 30 polypropylene square pieces of plastic film were placed in an outdoor flowing seawater tank located at the Kewalo Field Station in Honolulu, HI, USA, to observe biofilm development (Figure 5.1). Triplicate pieces of both polymer types were removed from each basket 2, 7 and 14 days after the start of the experiment. DNA was then extracted from each piece and used for amplicon sequencing to characterize the microbial communities developing on the squares. Median amplicon read counts per square was 92,000 reads (Table S5.1). At each time point, samples were also collected for SEM imaging.

A succession from a bacteria dominated community in the 48-hour time point (T1) to a community dominated by photosynthetic eukaryotes in the 7 and 14-day time points (T2 and T3) was immediately evident in the SEM images and amplicon data (Figure 5.2, Figure 5.3). A single layer of predominantly rod-shaped prokaryotic cells covered the T1 samples. The T2 and T3 samples were covered with a layer of larger eukaryotic organisms, including diatoms that formed dense patches. SEM images showed microbial cells attached to eukaryotic organisms (Figure 5.1E). In addition, over 85% of all amplicon reads from T1 samples mapped to bacteria, whereas all the T2 and T3 samples were dominated by chloroplast sequences, of which at least 75% from each sample mapped to Bacillariophyta (diatoms).
Richness and composition of bacterial communities on plastic squares

Although bacterial relative abundances decreased as the experiment progressed, bacterial richness on the plastics significantly increased between T1 to T2 samples and from T2 to T3 samples (Figure 5.4, Kruskal-Wallis test: $P < 0.001$ followed by post hoc Dunn test: all pairwise comparisons between time points: $P < 0.01$). Bacterial richness on the plastics remained lower than richness in the macroalgae, sediment and water samples. Bacterial richness on average also appeared higher on polypropylene squares compared to polyethylene squares during the first time point, however differences were not significant after adjusting significance levels to accounting for comparing all combinations of time and polymer type ($P<0.11$). The majority of bacterial sequences in T1 samples mapped to a single *Oleibacter* OTU (Figure 5.5). The abundances of this *Oleibacter* OTU dropped below 3.4% across the T2 and T3 samples where communities were composed of numerous other Gammaproteobacteria OTUs as well as Alphaproteobacteria, Saprospirae, Deltaproteobacteria, Flavobacteriia, Verrucomicrobia, BME43 and Betaproteobacteria OTUs (Figure 5.5). Abundant bacterial families in the T2 and T3 samples included Rhodobacteraceae, Hyphomonadaceae, Altermonadaceae, Vibrionaceae and Saprospiraceae (Figure 5.6).

Compositional differences in bacterial communities between time points were evident in NMDS plots (Figure 5.7). T1, T2 and T3 plastic samples formed clear non-overlapping clusters. Non-parametric multivariate analysis of variance confirmed that bacterial community composition on plastics significantly differed across time points (Table 1, $p < 0.001$). Time explained the majority of the variation between plastic samples ($R^2=0.78$). The analysis also showed significant interactions between polymer type and time as well as whether the basket was wrapped in mesh and time (Table 1, $p < 0.001$). This combined with clustering patterns in NMDS plots suggests that bacterial community composition was most sensitive to the polymer type at the first time point, and presence of mesh at the final time point (Figure 5.7).
Comparison of Bacterial Communities on Plastics Versus Nearby Habitats

Bacterial communities on plastic particles clustered separately from communities observed on other substrata within the tank (Figure 5.8). Water samples clustered by size fraction and both sizes were distinct from other sampled substrata. Macroalgae, the two coral samples and the majority of the sediment samples also formed a distinct cluster. However, this macroalgae cluster was more similar to the bacterial communities on T2 and T3 plastic samples than to the T1 plastic samples.

To better understand the potential source of bacterial plastic colonizers, we compared the most abundant bacterial OTUs found on plastic samples (> 0.1% in at least one plastic sample) to the OTUs observed in other microbial habitats in the flowing seawater tank (Figure 5.9). The majority (62.2%) of abundant bacterial plastic OTUs were observed in all the habitats sampled and only 3.5% of the abundant OTUs were not observed in any other habitat.

Correspondence Between Bacteria and Photosynthetic Eukaryotes

We observed a positive relationship between bacterial and chloroplast dissimilarity in T2 and T3 samples (P < 0.001, Figure 5.10.). Similar to the bacterial results, the chloroplast samples clustered by time and presence or absence of mesh (Figure 5.11). Non-parametric multivariate analysis of variance confirmed that the effect of both time and the interaction between time and mesh on T2 and T3 chloroplast samples were significant (Table 2, P < 0.001).

Discussion

Our experiment captured two early stages of biofouling: colonization by bacteria, followed by colonization by photosynthetic eukaryotes. There is debate as to whether unicellular algae, particularly diatoms, settle on surfaces simultaneously or after bacteria have colonized and whether the presence of bacteria is required for algae settlement (Cooksey and Wigglesworth-Cooksey, 1995). However, recent work has been conducted showing that bacteria at least enhance biofilm formation by
diatoms (Bruckner et al., 2011; Yang et al., 2016). Although potential mechanisms are not addressed here, SEM images and amplicon data clearly showed that in our system, bacterial cells were present on the plastic films in large numbers before the photosynthetic eukaryotes became established on the plastic surfaces.

Several previous marine biofilm studies observed that early communities contained both Alphaproteobacteria and Gammaproteobacteria, but emphasized that Alphaproteobacteria (mainly Roseobacteria in the Rhodobacteraceae family) were likely the pioneering taxa on abiotic surfaces (reviewed by Dang and Llovell, 2016). We observed however that one Gammaproteobacteria OTU, mapping to the *Oleibacter* clade, made up the majority of the bacteria in the biofilms at 48 hours. Our first time point was collected earlier than most previous studies, suggesting we captured an earlier stage of succession. Consistent with our findings, Gammaproteobacteria was the first bacterial clade to colonize artificial surfaces in significant abundances in a study in Sacheon harbor, Korea where the first time point was collected three hours after the surfaces were placed in the harbor (Lee et al., 2008).

A recent study, following microbial succession on chitin particles observed a high diversity of microbes during the first 24 hours of colonization, followed by a dip in diversity from the 24 to 48 hour time points and then an increase in diversity that was maintained until the end of the 144 hour experiment (Datta et al., 2016). These observations, together with metagenomic data and functional assays of isolates, led the authors to propose three phases of microbial colonization that are driven by differing strengths and mechanisms of habitat filtering, as well as motility. They proposed that phase one is largely dictated by bacterial attachment processes, but the ability to attach to chitin particles is only a weak selective filter, resulting in high initial microbial diversities. During phase two, the ability to grow on chitin imposes a strong selective pressure that reduces diversity.

In our study, it was not clear if the high abundance of the *Oleibacter* OTU was because the ability to attach to polyethylene and polypropylene films was a stronger selective force than the ability to attach to chitin or alternatively, if the ability to grow on plastic surfaces imposed a selective sweep. The majority of cells visible in 48 hours SEM images had excreted an extracellular matrix to facilitate attachment,
demonstrating at a minimum that cells were forming biofilms, not simply transiently sticking to plastic surfaces. The ability to withstand high levels of light may also have imposed a strong selective pressure as once the bacteria were attached to the plastic they were held at the surface of the water rather than cycling through the mixed layer. However, little is known about whether *Oleibacter* produce pigments or can employ other adaptive strategies that allow them to thrive with high light exposure (Satomi and Fujii, 2014).

*Oleibacter* is a closely related to the genus *Thalassolituus*. Members of both these groups can degrade alkanes and have been found in crude oil contaminated seawater (Teramoto *et al*., 2011; Satomi and Fujii, 2014; Al-Bader *et al*., 2013). Additional work is required to determine if the *Oleibacter* we observed were utilizing either the plastic polymers or the conditioning film bound to the plastic surfaces as a nutrient source. There has been scant evidence that microbes break down or consume conventional plastic polymers in marine environments, especially at ecologically relevant time scales (Tosin *et al*., 2012; Nauendorf *et al*., 2016). Most evidence suggesting marine microbes decompose plastic is based on microscopic pockmarks or fissures observed in SEM images (Eich *et al*., 2015; Zettler *et al*., 2013). However these may also be artifacts from sample preparation.

Following our 48-hour time points, we observed increases in bacterial richness and highly consistent shifts in community composition. It is difficult to parse the relative impacts of deterministic (i.e. niche-based) verse stochastic (i.e. dispersal) processes in shaping the community changes across time. For example, the increases in diversity on the plastic squares may have been a result of continuous exposure to new organisms from the flowing seawater system. Alternatively compositional changes across squares could have been coordinated by unmeasured shifts in water chemistry between time points. Replicating this study at different times and locations as well in more controlled system will help clarify the role of immigration, environmental variability and successional processes in shaping the communities across time.

The increase in bacterial richness we observed is consistent however with the possibility that primary bacterial colonizers and settling eukaryotes created new niches for secondary colonizers. In addition, the relationship between the presence of mesh on
some baskets and composition of bacterial and photosynthetic eukaryote communities suggests that the mesh was impacting eukaryote communities which in turn was impacting bacterial communities in the later time points. A similar correlation between prokaryotic and eukaryotic communities was observed on PET bottles that had been attached to a buoy in the North Sea for over 5 weeks (Oberbeckmann et al., 2016). These trends align with successional phase III, proposed by Datta et al. (2016), where communities at later time points were dominated by taxa that lived off byproducts from substrate specialists that colonized earlier in the experiment.

Phytoplankton, in particular diatoms, are known to engage in complex interactions with bacteria (Amin et al., 2012). For example, phytoplankton are surrounded by a layer referred to as the phycosphere, where dissolved organic matter released from the cells is concentrated and attracts heterotrophic bacteria (Bell and Mitchell, 1972; Amin et al., 2012; Smriga et al., 2016). Some bacteria cultured from diatom microbiomes also have been shown to enhance or slow diatom growth or stimulate secretion of extracellular polymeric substances, depending on the identity of the interacting organisms (Bruckner et al., 2011; van Tol et al., 2017). Although research aimed at characterizing the bacterial communities living in conjunction with benthic diatoms is limited, there has been recent studies describing potential diatom-bacteria interactions during plankton blooms (Buchan et al., 2014; Sanli et al., 2015; Needham et al., 2016). Several of the bacterial clades reported to be associated with diatoms including Alteromonas, Ruegeria, Glaciecola, Flavobacteriia and Bacteroidetes, we also detected in later time points when photosynthetic eukaryotes were abundant on plastic samples (Amin et al., 2012).

Traditional biofouling succession models emphasize transitions from bacteria to microalgae to larger eukaryotic organisms and physical interactions with the substratum are believed to largely shape communities early in the fouling sequence and biological interactions largely shape communities later in the fouling sequence (Wahl, 1989). We took a more microorganism-centric approach by focusing on bacterial across all time points. We found that the bacterial composition on plastic surfaces changes dramatically across time, but that early bacterial colonizers on plastic appear to be more impacted by polymer type, whereas bacterial interactions with eukaryotic
settlers is more important in shaping bacterial community composition at later time points. This transition from the importance of interactions with the substrate to interactions between taxa is consistent with biofouling work that encompasses bacterial cells to complex invertebrates and the work by Datta et al that encompasses only heterotrophic prokaryotes (Wahl, 1989; Datta et al., 2016).

This suggests that succession on plastic debris is not fundamentally different than succession on other hard substrata. Bacterial communities at later time points on our study were also more similar to the bacterial communities on macroalgae raising the possibility that if our study continued the bacterial communities observed on plastic surfaces may eventually even become indistinguishable from those on biotic and other abiotic surfaces. Two studies observing microbial communities on plastic particles exposed to seawater for much larger time periods could not differentiate between bacterial composition on polyethylene versus polypropylene or between communities on polyethylene terephthalate versus natural marine particulate matter (Amaral-Zettler et al., 2015; Oberbeckmann et al., 2016). However, communities on foamed polystyrene have been shown to cluster separately and contain a higher abundance of bacterial cells than other polymers (Amaral-Zettler et al., 2015; Carson et al., 2013).

Concerns have been raised about the potential for plastic debris to transport pathogenic microorganisms. This concern largely arose from one study that observed a high abundance of Vibrio species in one sample (Zettler et al., 2013). However Vibrios are common in coastal waters and only a subset are pathogenic (Bruto et al., 2016; Hunt et al., 2008). We did find Vibrionaceae sequences in some of the T3 samples. Given that they were only abundant in the T3 samples, and a previous culture-based study suggested that Vibrios are secondary colonizers on polystyrene, it is likely that the Vibrios we observed were located on other organisms rather than the plastic surfaces (Foulon et al., 2016). Given that plastic surfaces are biochemically very different than biotic substrates, transmission of any microbial pathogen on plastic debris is likely dependent on whether a host can settle and persist on plastic debris after infection. Experimentation with sessile eukaryotes and known pathogens is necessary before for any real conclusions to be drawn.
Additional experiments that manipulate the taxa and environment plastics are exposed to will improve our understanding of the determinants of microbial communities on plastic over time. Higher resolution sampling will also be valuable. Here we observed a highly replicated succession of bacteria communities on plastic surfaces. The initial microbial communities were dominated by one OTU, but quickly became more diverse. This work highlights the importance of observing macro- and microorganisms while studying the microbial communities on plastic debris and puts colonization of plastic debris in the context of marine biofouling research.
<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean of Squares</th>
<th>F. Model</th>
<th>$R^2$</th>
<th>Pr(＞F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>2</td>
<td>21.5213</td>
<td>10.7606</td>
<td>260.136</td>
<td>0.78143</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Mesh</td>
<td>1</td>
<td>0.5941</td>
<td>0.5941</td>
<td>14.363</td>
<td>0.02157</td>
<td>0.11</td>
</tr>
<tr>
<td>Polymer type</td>
<td>1</td>
<td>0.0130</td>
<td>0.0130</td>
<td>0.313</td>
<td>0.00047</td>
<td>0.8259</td>
</tr>
<tr>
<td>Basket</td>
<td>4</td>
<td>0.4047</td>
<td>0.1012</td>
<td>2.446</td>
<td>0.01470</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Time: Mesh</td>
<td>2</td>
<td>0.7340</td>
<td>0.3670</td>
<td>8.873</td>
<td>0.02665</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Time: Polymer type</td>
<td>2</td>
<td>0.4317</td>
<td>0.2159</td>
<td>5.219</td>
<td>0.01568</td>
<td>0.0008***</td>
</tr>
<tr>
<td>Polymer type: Mesh</td>
<td>1</td>
<td>0.0329</td>
<td>0.0329</td>
<td>0.795</td>
<td>0.00119</td>
<td>0.4459</td>
</tr>
<tr>
<td>Time: Polymer type: Mesh</td>
<td>2</td>
<td>0.0863</td>
<td>0.0432</td>
<td>1.044</td>
<td>0.00313</td>
<td>0.3597</td>
</tr>
<tr>
<td>Residuals</td>
<td>90</td>
<td>3.7229</td>
<td>0.0414</td>
<td>0.13518</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>105</td>
<td>27.5410</td>
<td>1.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1. Results from non-parametric multivariate analysis of variance of bacterial samples on plastic squares. Probabilities associated with polymer type, time and their interaction terms were calculated by permuting polymer classification (polyethylene vs. polypropylene) or time point (T1 vs. T2 vs. T3) of samples within baskets. The significance of the mesh treatment was calculated by permuting treatment across baskets. Significance codes are as follows: < 0.001: '***'

<table>
<thead>
<tr>
<th></th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean of Squares</th>
<th>F. Model</th>
<th>$R^2$</th>
<th>Pr(＞F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>1</td>
<td>1.4223</td>
<td>1.4227</td>
<td>34.643</td>
<td>0.21349</td>
<td>0.0001***</td>
</tr>
<tr>
<td>Mesh</td>
<td>1</td>
<td>1.4780</td>
<td>1.47797</td>
<td>35.999</td>
<td>0.22185</td>
<td>0.099</td>
</tr>
<tr>
<td>Polymer type</td>
<td>1</td>
<td>0.0363</td>
<td>0.03629</td>
<td>0.884</td>
<td>0.00545</td>
<td>0.469</td>
</tr>
<tr>
<td>Basket</td>
<td>4</td>
<td>0.8900</td>
<td>0.22249</td>
<td>5.419</td>
<td>0.13359</td>
<td>0.0005***</td>
</tr>
<tr>
<td>Time: Mesh</td>
<td>1</td>
<td>0.3078</td>
<td>0.30782</td>
<td>7.498</td>
<td>0.04620</td>
<td>0.0004***</td>
</tr>
<tr>
<td>Time: Polymer type</td>
<td>1</td>
<td>0.0517</td>
<td>0.05168</td>
<td>1.259</td>
<td>0.00776</td>
<td>0.274</td>
</tr>
<tr>
<td>Polymer type: Mesh</td>
<td>1</td>
<td>0.0125</td>
<td>0.01252</td>
<td>0.305</td>
<td>0.00188</td>
<td>0.876</td>
</tr>
<tr>
<td>Time: Polymer type: Mesh</td>
<td>1</td>
<td>0.0413</td>
<td>0.04130</td>
<td>1.006</td>
<td>0.00620</td>
<td>0.384</td>
</tr>
<tr>
<td>Residuals</td>
<td>59</td>
<td>2.4223</td>
<td>0.04106</td>
<td>0.36359</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>6.6621</td>
<td>1.00000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.2. Results from non-parametric multivariate analysis of variance of chloroplast samples on plastic squares. Probabilities associated with polymer type, time and their interaction terms were calculated by permuting polymer classification (polyethylene vs. polypropylene) or time point (T2 vs. T3) of samples within baskets. The significance of the mesh treatment was calculated by permuting treatment across baskets. Significance codes are as follows: < 0.001: '***'
Figures

Figure 5.1. Images of experimental setup. (A) Outdoor flowing sweater tank with all 6 replicate baskets containing plastic film squares. Styrofoam floats were attached to the sides of each basket to buoy the top of the baskets above the waters surface (B). Baskets were also tethered to a cable suspended above the tank. (C) A view of the tank where the Montipora capitata reef is visible and Dr. Sarah-Jeanne Royer next to the tank for scale.
Figure 5.2 (previous page). Select SEM images from plastic collected 48 hours (T1, Figure 2A and 2B), 7 days (T2, Figure 2C and 2D) and 14 days (T3, Figure 2E and 2F) after plastic was placed in the flowing seawater tank. Inset in B zooms in individual cells and inset in E zooms in on cells attached to a larger organism. Scale bars for each image are on the bottom right with value reflecting the entire length of the scale bar.
Figure 5.3. The relative abundance of amplicon sequences mapping to chloroplasts in plastic samples. Each vertical bar represents one sample. Samples are grouped by replicate basket (B1-B6). Baskets B1, B4 and B6 were wrapped in 105μm mesh. Subdivisions in each vertical bar outline the abundance of individual chloroplast OTUs defined by 97% sequence identity. For each time point and basket, the first three replicate bars show communities on polyethylene and the second three bars show communities on polypropylene squares.
Figure 5.4. Rarefaction curves showing OTU richness (y-axis) per number of bacterial amplicon reads sampled (x-axis). Lines are colored by sample type as follows: red: T1 plastic squares, yellow: T2 plastic squares, purple: T3 plastic squares, light blue: 0.2-105μm seawater size fraction, dark blue: >105μm seawater size fraction, green: macroalgae growing in the tan, brown: sediments collected in the tank.
Figure 5.5. Abundant bacterial classes across plastic samples. Each vertical bar represents one sample. Samples are grouped by replicate basket (B1-B6). Baskets B1, B4 and B6 were wrapped in 105μm mesh. White, gray and dark gray horizontal bars along the x-axis delineate samples collected at T1, T2 and T3, respectively. Subdivisions in each vertical bar outline the abundance of individual bacterial OTUs normalized by the total number of bacterial reads per sample. For each sample, only OTUs with abundances greater than 0.1% that mapped to one of the highly abundance classes are displayed.
Figure 5.6. Abundant bacterial families across plastic samples. Same as figure 5.5, except OTUs are colored by their family level taxonomic assignments.
Figure 5.7. Nonmetric multidimensional scaling (NMDS) plot displaying Bray-Curtis distances between bacterial communities on plastic squares. A is colored to emphasize the division of between bacterial communities on polyethylene and polypropylene in T1 samples. B is colored to emphasize the division between pieces in meshed and unmeshed baskets in T3 samples. Stress = 0.064
Figure 5.8. NMDS plot displaying Bray-Curtis distances between all bacterial communities sampled in the tank. Sample shapes designate sampling time and color represents the sample substrates. 'Polyethylene' and 'polypropylene' are the two plastic film polymers studied. Stress = 0.082
Figure 5.9. Venn diagram comparing abundant plastic-associated bacterial OTUs with OTUs observed in other habitats and substrates in the flowing sea-water tank. To be included in the plastic bin, OTUs need to have > 0.1% abundance in at least one plastic sample. For other bins, all OTUs observed were included. Values in parentheses show percent relative to the total number of OTUs in the plastic bin.
Figure 5.10. Scatter plot showing a positive relationship between bacterial and chlorophyll community dissimilarity on plastic squares. Pairwise sample comparisons within T2 samples (A, spearman rho=0.59, p < 0.001), between T2 and T3 samples (B, spearman rho=0.73, p < 0.001) and within T3 samples (C, spearman rho=0.78, p < 0.001) are shown. Each point reflects pairwise Bray-Curtis dissimilarities between the chloroplast composition (x-axis) and bacterial composition (y-axis) on the same two pieces of plastic. Symbols and colors reflect sampling times of each pair being compared.
Figure 5.11. NMDS plot displaying Bray-Curtis distances between T2 and T3 chloroplast samples collected on plastic squares. Sample shapes designate sampling time and color represents sampling time and whether or not the sample came from a basket surrounded by 105μm mesh. Stress = 0.11

References


5. Anderson MJ, Walsh DCI. (2013). PERMANOVA, ANOSIM, and the Mantel test in
the face of heterogeneous dispersions: What null hypothesis are you testing? *Ecol Monogr* **83**: 557–574.


interactions lead to rapid micro-scale successions on model marine particles. *Nat Commun* 7: 11965.


Acknowledgements

We would like to thank Tina Carvalho for assistance with SEM imaging at the Pacific Biosciences Research Center’s Biological Electron Microscope Facility, Dr. Michael Hadfield and Kory Misaki for providing access and assisting with the flowing seawater tanks at the Kewalo Marine Station, Carla Gimpel and Sarah-Jeanne Royer for help with sample collection and Anna Romano and Paul Den Uyl II for efforts implementing amplicon sequencing in the DeLong lab. We would also like to thank members of the Cordero lab for helpful comments. This work was supported by grants from the Gordon and Betty Moore Foundation (E.F.D., 3777; D.M.K., 3794), the U.S.
Environmental Protection Agency STAR fellowship (J.A.B.), the National Science Foundation (E.F.D., C-MORE; EF0424599) and the Simons Collaboration on Ocean Processes and Ecology (E.F.D. and D.M.K., SCOPE; 329108).
Chapter 6. Summary and Future Directions

Marine microbes and the Earth's biogeochemistry are tightly intertwined. Understanding the biotic and abiotic factors that shape microbial communities and in turn how the activity of these communities modifies the environment around them is critical for understanding and maintaining the ecosystem services that support life on Earth (Azam and Worden, 2004). The processes that drive changes in microbial communities span broad temporal and spatial scales, resulting in communities that are dynamic but exhibit some predictable patterns (Fuhrman et al., 2015; Martiny et al., 2011). Both to improve our understanding of the drivers that shape microbial communities and to be alerted to unpredicted events, it is critical that marine microbial communities are closely monitored across both space and time.

A vision for the future of marine microbial ecology includes remote sensing systems that can monitor communities at high sampling densities at global scales, similar to the high frequency and density of global meteorological and oceanography data currently produced by programs such as the ARGO program and the National Oceanic and Atmospheric Administration's national data buoy network (Kirchman and Pedrós-Alió, 2007; Argo Data Management Team; NOAA National Buoy Data Buoy Center). Up to this point, marine microbe sampling technology has largely limited exploration of open ocean marine microbial communities to ship expeditions. However autonomous mobile platforms that sample communities in situ are under development and early trials have yielded valuable insights into microbial communities across short time scales (Scholin, 2010; Ottesen et al., 2014; Robidart et al., 2014). In addition, the ability to profile complex natural populations (largely via nucleic acid sequencing) has experienced technological and in particular bioinformatics and computational bottlenecks but rapid advancements are being made in these areas as well (White et al., 2016).
To date several multi-year studies have observed microbial community composition at major time-series sites including the Sargasso Sea, English Channel, San Pedro Channel and Mediterranean Sea (Treusch et al., 2009; Gilbert et al., 2012; Fuhrman et al., 2006; Galand et al., 2010) and the spatial distribution of microbes has been investigated through global-scale expeditions including the Sorcerer II Global Ocean Sampling Expedition and the Tara Ocean's Expedition, as well as numerous studies at smaller spatial scales (Rusch et al., 2007; Sunagawa et al., 2015). The diversity and structure of marine microbial communities vary across latitude, season and depth below the ocean’s surface (Fuhrman et al., 2015; Sunagawa et al., 2015; Raes et al., 2014; Ladau et al., 2013; DeLong et al., 2006). Environmental conditions such as temperature, light, water stratification and nutrient availability in addition to interactions between taxa across trophic levels all play important roles in shaping these trends (Lima-Mendez et al., 2015; Needham et al., 2016). The role of processes such as endemism and dispersal in shaping communities are also still under debate (Lennon and Jones, 2011; Caporaso et al., 2012).

Survey-based studies generate a wealth of hypotheses to explain environmental observations (Gilbert et al., 2011). However the expansive phylogenetic and functional diversity of microbial communities as well as the complex nature of environmental variation make it difficult to establish definitive underlying mechanisms. Therefore although model systems and controlled experimental settings inevitably differ from natural conditions, they offer the ability to isolate specific potential drivers and determine cause–and-effect relationships (Zhou, 2009). Therefore experiments that test environmental observations complement survey work and are key to better understanding underlying mechanisms.

The work in this thesis helps expand our understanding of spatial and temporal distributions of microorganisms at Station ALOHA, a site representative of the North Pacific Subtropical Gyre (NPSG). Station ALOHA is unique from other long-term time series sites and global scale surveys in that surface waters are perennially stratified and span a narrower range of environmental variability across time. However, the NPSG represents the largest circulation feature on Earth and therefore understanding its biological dynamics is imperative (Karl and Lukas,
1996; Karl and Church, 2014). Below I summarize key findings from my research and discuss ways in which the work can inform future microbial oceanography surveys at Station Aloha. I also highlight several avenues for future research that could be addressed with experimental manipulations.

**Variability in Surface Planktonic Communities**

Few microbial time-series studies have been conducted in open ocean habitats having low seasonal variability such as the North Pacific Subtropical Gyre (NPSG), where surface waters experience comparatively mild seasonal variation (Church *et al.*, 2013). In chapter two we observed that microbial richness at 25m was positively correlated with the average wind speed of days prior to sampling, suggesting that wind-driven mixing or dust deposition may be impacting microbial diversity. We also observed that bulk microbial community composition at 25m exhibited significant correlations with solar irradiance. Many bacterial groups whose relative abundances varied with solar irradiation corresponded to taxa known to exhibit strong seasonality in other oceanic regions.

The study described in Chapter 2, spanned an approximately 2-year period, however sampling has continued after the boundaries of this study. Analyzing data from additional years should help to determine whether wind and sunlight are consistently the dominant environmental correlates of richness and composition at Station ALOHA. Alternatively the strength of the relationship could depend on inter-annual environmental fluctuations such as the El Nino Southern Oscillation. The wind speeds we observed varied from close to zero to approximately 11 m/s. These wind conditions can be describe as ranging from calm to a strong breeze. Our results lead to the prediction that sampling immediately after stronger wind events will yield even greater increases in microbial richness.

The increase in richness we observed corresponded to an increase in chlorophyll concentrations but not an increase in autotrophic biomass, suggesting that cells with higher chlorophyll concentrations located at the deep chlorophyll maximum layer (DCM) had been entrained upward. At the Bermuda Atlantic Time Series, diversity is highest
during the period of deep water mixing, which at BATS is a predicable seasonal occurrence, unlike at Station ALOHA (Vergin et al., 2013). We did not observe a relationship between deep water mixing and increased diversity, but perhaps the mix layer deepened then shoaled before observations were made. The potential wind effects observed in our study could be confirmed by sampling at regular frequencies across a transition period from calm to windy conditions. This would be particularly informative if samples are collected above and below the mixed layer depth and well as above and below the DCM, to allow for tracking both taxa and water vertical transport.

It was reasonable that solar irradiance is the strongest correlate to community composition at 25 m, as sunlight fuels the base of the food chain, impacts ocean hydrology and mediates many interactions between taxa (Moran and Miller, 2007; Ruiz-González et al., 2013). Despite being less seasonal than other study sites, primary production at station ALOHA is linked to irradiance and highest during summer months (Karl and Church, 2014). It is intriguing that the Tara Oceans Project found microbial community composition in surface waters was more strongly related to water temperature than photosynthetically available radiation (Sunagawa et al., 2015). The Tara Ocean’s dataset however incorporated geographically and oceanographically different sites and biomes, which spanned greater differences in temperature and solar radiation than our study, and were not developed to sample a seasonal time series. This likely may explains why the potential strength of temperature versus sunlight on community composition differed between studies. Mesocosm experiments could be used to isolate the effects of varying solar radiation or temperature on microbial community composition from other potential abiotic factors. Mesocosm experiments have successfully been employed to study the impact of grazers on bacterial community composition, the impact of temperature and ultraviolet radiation on plankton community structure and the impact of ocean acidification on marine dissolved organic matter over an entire seasonal cycle (Baltar et al., 2016; Zark et al., 2015; Vidussi et al., 2011).

**Microbial Communities Across the Depth Gradient**
In Chapter 2 we observed that compared to 25 m, microbial communities at 500 m had higher richness, less seasonality and shared few taxonomic groups with surface waters. This is consistent with previous observations and the dramatic physical, geochemical and biological differences between the ocean’s sunlit layer and the darker, colder waters below. The nature of the microbial adaptive and physiological transitions from shallow to deep surface water communities and potential stratification within the euphotic and aphotic zones however is not well understood. In chapter three we conducted a time-series survey of picoplankton communities from surface waters to 1000m depth. We observed that microbial richness was generally higher in deeper samples, although it peaked at 125 m and 200 m and then declined with depth. We also observed that samples partitioned into clear clusters that consisted of all 25, 75 and 125 m samples above the DCM, 125 m samples below the DCM, all 200 m samples and all 500, 770 and 1000m samples.

Our work suggests that the DCM delineates a sharp transition between the taxonomic and functional composition of communities found in the euphotic and aphotic zones. Future survey studies at station ALOHA should target the communities directly above and below the DCM to confirm these observations, better define the boundaries of the transition in communities and improve our understanding of the interplay between upwelling nutrients and down welling light that shapes this feature. Our results also demonstrate that mesopelagic microbial communities are not in fact homogenous across depths. Future survey studies aimed at characterizing communities within the mesopelagic should sample at multiple depths, in particular at both 200 and 500 m, across this zone. Communities directly below the euphotic zone compared to deeper in the mesopelagic could be shaped by differing access to sinking particulate matter. Therefore, measuring particulate carbon concentrations and respiration rates while profiling microbial communities at both these depths would be valuable.

Several isolates of abundant bacterioplankton clades from oligotrophic ocean surface waters have been reported to have small, GC poor genomes, including *Prochlorococcus*, SAR11 and SAR86 (Yooseph et al., 2010; Swan et al., 2013; Giovannoni et al., 2005; Dupont et al., 2012; Rocap et al., 2003). Consistent with
previous studies of isolates and single amplified genomes, our gene-based metagenomic analyses described in Chapter 3 revealed that low GC content and small genomes were prevalent across diverse microbial taxa in surface samples at all time points. However, samples spanning the depth gradient also revealed a sharp discontinuity in these genomic traits across the DCM. A number of hypotheses have been proposed to explain potential environmental pressures on GC content, but no well-supported mechanisms have emerged (Agashe and Shankar, 2014). McEwan and colleagues (1998) noted that GC pairs have one more nitrogen atom than AT pairs and observed aerobic nitrogen fixers in two lineages had higher genomic GC content than their non-nitrogen fixing relatives, suggesting that higher GC is related to reduced nitrogen limitation through nitrogen fixation (McEwan et al., 1998). Since most bacteria cannot fix nitrogen this hypothesis is very limited in scope.

Two later studies noted that high GC codons tend to code for amino acids with larger numbers of nitrogen atoms in their side chain (Baudouin-Cornu et al., 2004; Bragg and Hyder, 2004). In addition, AT pairs have one more carbon molecule than GC pairs and high AT codons tend to code for amino acids with larger numbers of C atoms in their side chains. Together this suggests that the structure of the codon table allows evolutionary processes to simultaneously tune an organism’s genome and proteome in response to carbon and nitrogen limitation, but that there is a trade off between optimizing for carbon versus nitrogen conservation. Consistent with this, we observed that the average number of nitrogen atoms per amino acid residue side chain was higher below DCM and the average number of carbon atoms per amino acid residue side chain was higher above the DCM.

Together the shifts in GC and other genome traits we report in Chapter 3 across a diversity of microbial taxa in our spatiotemporal metagenomic survey, are most congruous with the hypothesis that organisms above the DCM have evolved to conserve nitrogen in their macromolecular pools under continuous nitrogen limitation of ocean surface waters. The higher GC and larger genomes found in the organisms that reside below the DCM suggests relaxation of this selective pressure near the nitrocline. It would be difficult however to demonstrate experimentally
that low GC and preferential incorporation of low nitrogen amino acids provides an advantage under low nitrogen regimes, as the GC of organisms evolves over long time spans, and is entangled with other life history traits such as adaption to different light levels and ability to take up different forms of nitrogen (Biller et al., 2014). Interestingly in this regard, many diverse taxa with fundamentally different physiological strategies and life histories, all appear to follow the same GC content versus depth trend. Collecting environmental proteomic data across the depth gradient would allow for a direct comparison of protein nitrogen use between communities and diverse taxa, and a better understanding of how the genomic features we observed translate to potential elemental conservation.

Microbes on plastic debris

Our findings described in Chapter 4 suggested that plastic debris forms a habitat for complex microbial assemblages that have organisms, lifestyles and metabolic pathways that are distinct from those of free-living planktonic microbial communities. The communities we observed on larger particles were net autotrophic and had high densities of chlorophyll compared to the surrounding seawater.

At the time of writing this thesis, this paper was one of only five studies that used molecular techniques to characterize microbial communities residing on buoyant plastic debris surveyed from open ocean habitats (Oberbeckmann et al., 2014; Zettler et al., 2013; Amaral-Zettler et al., 2015; Oberbeckmann et al., 2016). All the studies including ours either suffered from a limited number of samples, limited geographic range, or both. Additional surveys that cover larger expanses of the oceans are needed to get a rudimentary picture of the spatial and temporal distribution of microbial clades on plastic debris.

Plastic debris refers to items spanning several orders of magnitude in size (Barnes et al., 2009). In Chapter 4 we did not observe differences in microbial composition between the two size classes of microplastics we investigated. We did
however observe differences in chlorophyll a concentrations between the 0.2 – 2.0-mm size class and the > 2-mm size classes. Future survey work should consider including additional plastic size classes on both ends of the size spectrum, as ecological theory predicts that larger particles should be more diverse. The size of the debris will also impact how eukaryotes interact with the plastic. On the smaller end of the spectrum (< 1mm) plastic particles may frequently pass through the digestive system of zooplankton (Cole et al., 2016). Larger pieces can also pass through digestive systems of marine macro-fauna. However, a larger size range of organisms can also settle on larger pieces of plastic debris. Our observations in Chapters 4 and 5 as well as observations in the North Sea suggest that interactions between sessile eukaryotic and prokaryotic organisms play important roles in shaping the microbial communities and therefore observing macro- and microorganisms while studying on plastic debris is important to understanding the system (Oberbeckmann et al., 2016).

In Chapter 5 we observed some of the successional steps that may lead to the complex communities on plastic debris. One of the unique aspects of plastic debris compared to natural particulate matter is its ability to persistent in surface waters for long periods of time and be transported long distances across ocean basins. This has led to concerns that plastic debris is acting as a vector for invasive species and microbial pathogens (Zettler et al., 2013; Rech et al., 2016). We observed in Chapter 5 that bacterial communities at different time points were less similar than communities within the same time point. This suggests that microbial turnover could occur at high enough rates that plastic-mediated invasion is not a high risk. However, some OTUs were shared between time points and the methods we used to characterize microbial communities only captured relative abundances of organisms. A next step is to use a method such as real-time PCR to quantify the absolute change in abundance of cells between time points. This will clarify the extent to which microbial composition diverged because additional taxa are being added or because taxa are being replaced.

In addition, future experiments observing succession on plastic particles while manipulating the abiotic environment surrounding the plastic and the pool of
organisms available to potentially colonize the plastic would improve our understanding of the drivers of and temporal scales at which microbial communities on plastic turn over. For example, it would be valuable to directly test how water temperature or nutrient concentration perturbations surrounding a piece of colonized plastic, similar to changes a plastic particle would experiences while being transported from a coastal to oligotrophic gyre environment, potentially changes the successional trajectory of the community. In addition it would be possible to test a biofilm community’s ability to resistance invasion by transplanting already colonized plastic into media containing a new pool of microbes. Such questions are not well explored but will improve our understanding of microbial community assembly on inert surfaces and be key to evaluating the risk plastic debris poses to dispersing of microbes into new environments (Burmølle et al., 2006; Nadell et al., 2015).

Here we studied both planktonic and particulate microbial communities across space and time on the scale of days and millimeters on plastic debris to months and meters in planktonic communities. We observed changes in taxonomic groups, functional potential (protein coding genes) and genomic traits across these scales, which revealed potential ecological and environmental drivers shaping these communities. The roles of nitrogen availability in particular proved complex. While we observed in Chapter 3, that planktonic communities likely adapt to nitrogen limitation by reducing genomic and proteomic nitrogen use, work in Chapter 4 suggests communities on plastic debris instead may be able to survive low nitrogen conditions in the surface waters in part because some members fix nitrogen. Future work investigating planktonic and particulate (natural and anthropogenic) niches along multiple spatial and temporal scales in one study will provide a less fragmented, more integrated view of microbial community variability. Also including experimental based studies along side survey work will help validate hypotheses generated from observing natural changes in the environment. Such multifaceted deployments will be ambitious, but becoming more feasible, making it an exciting time for the field of marine microbial ecology.
References


Gilbert JA, O’Dor R, King N, Vogel TM, O’Dor R, Fennel K, et al. (2011). The importance of metagenomic surveys to microbial ecology: or why Darwin would have been a metagenomic scientist. Microb Inform Exp 1: 5.


