Microhabitats Shape Bacterial Community Composition, Ecosystem Function, and Genome Traits

by

Marian Louise Schmidt

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Ecology and Evolutionary Biology)
in the University of Michigan
2018

Doctoral Committee:

Assistant Professor Vincent J. Denef, Chair
Associate Professor Gregory Dick
Professor Deborah Goldberg
Professor George W. Kling
Professor Patrick D. Schloss

Marian Louise Schmidt

marschmi@umich.edu

ORICID iD: 0000-0002-2866-4496

Dedication

To my partner, Dan, who helped give me the strength to pursue my dreams.

Acknowledgments

My dissertation would not have been possible without the encouragement from countless people over the last six years.

My committee members, Vincent Denef, Greg Dick, Deborah Goldberg, George Kling, and Pat Schloss all played immense roles in my development as a scientist. I am grateful for my advisor, Vincent, for his dedication, patience, enthusiasm, openness, and prompt responses. Vincent has helped develop me into a better writer and researcher while allowing me the space to grow in other professional realms. Thanks to Greg for his great advice and expertise in our D^3 meetings, always being available to help me hone in on my scientific questions and pick the right samples to answer them. I am so appreciative of the guidance from Deborah who was always willing to develop my scientific questions into a conceptual diagram and think about the big picture. I am also thankful for Deborah's recognition of my growth. Her words of affirmation kept me going when things got tough. George played an extremely important role in helping me develop hypotheses, put my work in the broader context of the field, and create effective and straightforward presentations. I appreciate the Klingelhoffer lab meetings that I attended and the many hours that George generously gave me in the beginning of my PhD, which were critical in my early development as a graduate student. Finally, thanks to Pat who opened his lab meetings and office for discussions on microbial ecology, programming workflows, and open science. I am grateful for the supportive local Software Carpentry group that Pat (and Meg Duffy) created, which gave me the opportunity to develop as a programmer and teacher.

I would like to extend appreciation to everyone in my department who shared their research, experiences, and a drive to change things for the better. An important thank you goes to John Gallias who was always there in the moment when I was in tears with major computer problems. His compassion and thorough help allowed me to continue doing my work to the final hour of a deadline. I am thankful for Jane Sullivan who gave me logistical support and helped me when things got tough during pre-candidacy. Thanks to Cindy Carl who was always around to help with logistics and provide thorough and prompt support. I also appreciate the help that Kati Ellis, Carol Solomon, Kendra Renner, Stephanie Blackmore, and Rob Heller gave me.

Special thanks to current and former members of the Denef and Duhaime Labs (including Melissa Duhaime) for their career and life advice, coding wisdom, ice cream breaks, and offering an ear: Jacob Evans, Sara Jackrel, Rachel Cable, Edna Chiang, Ruben Props, Michelle Berry, Morgan Lindback, Nikesh Dahal, Amadeus Twu, Kathryn Schmidt, Rachna Pannu, and Nate Arringdale. I am grateful for Jeff White who had the patience during my first day in the field, which ended up with him saving our pickup truck and boat that I accidentally drove off the boat launch at 1am. I appreciate the helpful field assistants I had over the years: Justin Millar, Edna Chiang, Amelia Waters, Seana Florida, Daniel Katz, and Masanori Fujimoto.

I am thankful for the company and support from my family: mom, dad, Avery and Molly, Carla, and Uncle Terry. I am grateful for my friends and roommates who have given me the so much after a difficult day: Steven Silvis, Kyle Edwards, and especially my dear friend Sofia Dabrowski. I am indebted to the love, dedication, scientific advice, and encouragement that I have received from Daniel Katz since my first week in graduate school. I am grateful to Kathleen Gorski for her long friendship, to Mike Walker and Sierra Petersen for their helpful lunches, to Brooke Ayles, Masa Nohara, Genna Cohen, Rachel Cable, Jill & Mike Kiepura, Lauren Opra, Feli Mittereder, Dani Hochfellner, Marisa Spradlin, Chris Marinos, and Kay Olivarez for crushing it at climbing over the years, and to Josh Morrison, Sean Donovan, and Kady Kidd for the years of fabulous Ru Paul Drag Race watching, critiquing and entertainment. Thanks to the amazing friends who helped keep me healthy and happy throughout this process by running countless miles with me: Morgan Lindback, Clarisse Betancourt, Tiffany Carey, and Lucas Joel. Finally, I am lucky to have met such amazing people and scientists: John Paul Piso Balmonte, Tsinatkeab Hailu, Ben Miller, Pame Martinez, Dea Thomaz, Jordan Bemmels, Oscar Vargas, Byron Smith, Anat Belasen, and Chrissy Tingle.

I am grateful for the community of scholars who are dedicated to the current and ongoing efforts related to diversity, equity, and inclusion in science that will help make the lives of people of all backgrounds and social identities in our scientific community better. I appreciate the EEB Diversity Committee, especially Jo Kurdziel and graduate student representatives Teal Harrison, Kristel Sanchez, and Tamara Milton, my fellow GREEBs (*i.e.*, Graduate Representatives in Ecology and Evolutionary Biology) peer mediator from 2016-2018, Gordon Fitch, and the 2017-2018 GREEBs president, Jill Meyers, and Katherine Crocker who all helped to initiate (and continue) conversations related to climate and inclusion within our department. I am also thankful to the Diversity, Equity, and Inclusion Certificate program, which Rackham Graduate School created in 2017 and I had the honor of being part of the inaugural cohort. Throughout my certification, Deborah Willis and Rebeca Villegas helped me grow into a better facilitator of difficult conversations. I am also thankful to the many female scientists who I worked with in Females Excelling More in Math, Engineering, and Science (FEMMES) whose efforts will help to increase representation of minorities in science and spark scientific interest in young girls. I have learned immensely from the above individuals, groups, and programs, which give me hope for a more inclusive future in science and whose ongoing efforts will have lasting and positive impacts on our scientific community.

This work was made possible by support from the National Science Foundation (NSF) Graduate Research Fellowship Program (Grant No. DGE 1256260), the NSF Doctoral Dissertation Improvement Grant (Award ID: 1701467), the Margaret Ayers Host Award for Women Graduate Students, and annual travel and research funding from the University of Michigan Department of Ecology & Evolutionary Biology and Rackham Graduate School. The financial support, especially from the NSF and the Ayers Host awards, had an immense and positive impact on my research and well-being.

The support that I have received over the years means the world. My support team has empowered me to learn and achieve so much and I hope to do the same for others.

Table of Contents

List of Tables

List of Figures

List of Supporting Information

Abstract

This dissertation helps to integrate bacteria into the broader field of ecology by investigating bacterial community composition and diversity as it relates to ecosystem function in microhabitats within freshwater systems of the Great Lakes Region. Here, I combine field- and laboratory-based measurements of observational data collected from three major types of lake ecosystems: inland lakes, a freshwater estuary (Muskegon Lake), and a Great Lake (Lake Michigan). First, to determine the primary controls on lake bacterial community composition, I assessed the influence of lake layer (*i.e.* stratification), lake productivity, and particle-association on the bacterial community across 11 inland lakes with varying productivity in Southwestern Michigan. I found that particle-association very strongly structures freshwater bacterial community composition. Second, I studied a freshwater estuarine lake, Muskegon Lake, which has a large spatio-temporal variation in bacterial heterotrophic productivity, to test whether there was an association between heterotrophic production and bacterial biodiversity (defined as the number of taxa and taxon abundance). I specifically focused on two co-occurring freshwater habitats that my first chapter showed to be populated by very distinct communities: particle-associated and free-living. Positive biodiversity-heterotrophic productivity relationships were found only in particles. Third, I performed a genome-based analysis of free-living specialists, particle-associated bacterial specialists, and generalists to characterize the genomic architecture and genetic traits that are associated with adaptations to these specific habitats. The genomes of particle-associated specialist bacteria were about twice the size of the genomes of free-living specialists and generalists, which had streamlined genomes. Fourth, to identify the bacterial taxa driving heterotrophic productivity across the large set of lake samples, I found that high nucleic acid (*i.e.,* HNA) functional groups identified by flow cytometry can serve as a proxy for freshwater bacterial heterotrophic productivity, whereas low nucleic acid (*i.e.,* LNA) functional groups cannot. Then, I used a machine learning approach to identify bacterial taxa associated with HNA and LNA. This allowed me to identify the bacterial taxa, which were often members of the Phylum Bacteroidetes, that are associated heterotrophic productivity. Finally, I investigated patterns of lake specificity and phylogenetic conservation of taxonomic groups. Throughout my dissertation, I found that there was very deep (Class to Phylum-level) phylogenetic conservation of which bacteria lived in which habitats, but not of what bacterial taxa contributed to HNA and LNA functional groups, and thus heterotrophic productivity. Positive biodiversity-heterotrophic productivity relationships only existed in particle-associated, and not free-living communities, and communities composed of more phylogenetically related organisms were more productive per-capita. These differences in biodiversity-ecosystem function relationships may in part be explained by particle-associated bacteria having larger genomes, higher nitrogen content, and more unique genes that provide the potential for niche complementarity. The taxa that drove HNA and LNA cell numbers, and by proxy heterotrophic productivity, were lake and time-specific and indicated that taxa could switch between the two functional groups. Overall, my dissertation elucidates the ecological and evolutionary effects of microhabitat structure on bacterial communities and genomes in natural systems.

Chapter I:

Introduction

Bacteria play a fundamental role in shaping the ecosystems of our planet. Due to the diversity of their metabolism, their large abundances, and the ubiquity of microbial cells across all ecosystems, microbes pump the global biogeochemical cycles of elements (Falkowski et al. 2008) and influence the Earth's climate (Singh et al. 2010). And yet, we still lack an understanding of many fundamental aspects of bacterial ecology.

Over the last few decades, new sequencing technologies have opened up a window with novel views into the bacterial world. While initial research efforts were designed to answer basic questions, such as "who is there?" and "what are they doing?" (**Figure 1.1A**), basic community ecology relationships still remain poorly understood in bacterial systems. Therefore, there is a current focus in the fields of microbiology and ecology to more thoroughly integrate the two fields. There are many questions that remain to be answered in microbial community ecology, specifically by breaking down the community into different aspects (**Figure 1.1B**), such as abundance, diversity, and composition. For example, some important questions include: What are the effects of microbial abundance, diversity, and composition on function? What genomic characteristics allow bacteria to thrive in where they do? What are the environmental and ecological drivers to patterns of bacterial diversity? The work in this thesis helps to answer some of these questions. While there are many ways to approach these relevant questions, one of the initial steps is to determine the best ways to identify and measure bacterial taxa from the environment.

Methods to measure bacterial taxa

Traditionally, bacteria were cultivated and physiologically characterized, or later on, used to perform molecular cloning, Sanger sequencing of the complete 16S rRNA gene, and placed into the tree of life. However, these approaches are impeded by the fact that most bacteria are currently unable to be cultured in the lab (Stewart 2012). Therefore, culture-independent techniques have been more commonly used recently to measure bacterial diversity in environmental samples. The most widespread approach is to perform DNA extraction of a sample and then do high-throughput sequencing of the 16S rRNA gene in a sample of interest where universal primers are used to amplify a section of a universal marker, the 16S rRNA gene, and more specifically (typically) the V4 hypervariable region (Kozich et al. 2013). Next, the sequences can be matched up with a database (*e.g.,* the SILVA database (Quast et al. 2013a) or an ecosystem specific database (*e.g.,* the freshwater bacterial database (Rohwer et al. 2017a)) and the taxonomy of the organism can be associated with the sequence. The major benefit of this method is that the relative abundance of each organism can be calculated assuming or using a copy number value of the 16S rRNA gene from a database (Stoddard et al. 2015). [The problem of correcting based on 16S rRNA gene copy is still debated as some show that it improves estimates of microbial diversity (Kembel et al. 2012) while others maintain that it is an unresolved issue (Louca et al. 2018).] Relative shifts of the same OTU across samples are however not influenced by this copy number variation.

Beyond 16s rRNA copy number variation, there are other important drawbacks of this molecular-based approach. For example, the 16S rRNA gene is unable to resolve ecologically relevant units of taxonomy. This method also requires a relatively large sample, preservation in the field, DNA extraction in the lab, library preparation and sequencing, and data analysis. This leads to a relatively long turnaround time between sample collection and results. Alternative methods have instead applied single cell approaches to assess community diversity. This ranges from microscopic imaging, including the use of taxon-specific fluorescently-labeled DNA probes (Amann et al. 1990, Amann and Fuchs 2008) to methods combining such taxonomic identification with functional assays like nano-scale secondary ion mass spectrometry (*i.e.,* NanoSIMS, Herrmann et al. 2007). While these methods provide a high-resolution view into bacterial community structure and function, which has important implications for scaling up biogeochemical processes fueled by microbes (Popa et al. 2007, Dekas et al. 2009, 2016, Finzi-Hart et al. 2009), these single-cell, microscopy-based methods are time consuming and have low-throughput. Recently, another high-throughput single-cell based diversity assessment method has been developed using flow cytometry, a method that I use to define broad functional groups in **Chapter V**. This method uses flow cytometry measurements of several dimensions of phenotypic attributes of cells which are provided within minutes and require only 1 mL of sample. These data can then be processed through a pipeline that fits bivariate kernel density functions to phenotypic parameter combinations of an entire community (Props et al. 2016). This method has been shown to correlate with the 16S rRNA gene survey approaches mentioned above (Props et al. 2017a, 2017b) and can therefore be used as a quick method to measure microbial community diversity but not composition.

Some methods have been developed to connect 16S rRNA gene-based taxa to ecological, functional or phylogenetically relevant information. Some examples include Oligotyping (Eren et al. 2013), PICRUSt (Langille et al. 2013), operational ecological units (Preheim et al. 2013), and ecotypes (Koeppel and Wu 2014). These methods (including the measure of phenotypic diversity) are essentially limited to measuring overall shifts in diversity and cannot directly connect to metabolic or functional meanings, like nanoSIMS. Therefore, to allow the assessment of how taxonomic and phenotypic diversity shifts translate into shifts in functional diversity, microbial ecologists are increasingly turning to whole genome-based (*i.e.*, "genome centric") approaches. This is inspired by both the difficulty of measuring traditional traits of bacteria *in situ*, and the relative ease to deeply sample the genomic makeup of natural communities. Therefore, microbial ecologists often use the method of metagenomics (and its derivatives metatranscriptomics (RNA), metaproteomics (protein), and metametabolomics (metabolites)) to look deeper into the genetic underpinnings of changes in microbial abundance and predict the metabolic potential of organisms in the environment. In metagenomics, all of the DNA from an environmental sample is extracted and then sequenced using high-throughput sequencing. Next, bioinformatic assembly tools piece these small DNA sequences into larger contiguous DNA sequences (*i.e.*, "contigs"). The contigs are then clustered together to "bin" the samples into similar units based on tetra-nucleotide signatures and abundance measures (Alneberg et al. 2014, Wu et al. 2014, Kang et al. 2015, Laczny et al. 2015, Sczyrba et al. 2017). After quality control (*e.g.,* removal of bins that are composed of fragments originating from multiple populations), these bins can be considered as “metagenome assembled genomes” (*a.k.a.* MAGs). Importantly, a MAG is a population-level measure of a genome as each sequencing fragment likely originated from a different cell. A MAG thus reflects the average genome of a population, although it is often incomplete and can be contaminated by genomes from other populations. Thanks to recent developments in binning and curation tools (Eren et al. 2015, Parks et al. 2015, Olm et al. 2017, Sieber et al. 2018) the quality of bins can be more readily addressed and after appropriate quality filtering, a MAG can then be used for taxonomic and functional annotation of the organisms. Thus, MAGs can be used for more specific questions about why an organism lives where it was found and what its metabolic potential might be. This allows a further integration of a community diversity metric to ecologically relevant functions.

Biodiversity impacts community function

The number of species on our planet is dramatically decreasing (Thomas et al. 2004, Wake and Vredenburg 2008) and a major question in biology is: What is the impact of this reduction in species on ecosystem functioning (Loreau et al. 2001, Hooper et al. 2005)? In response to this question, there has been a large body of work in the field of biodiversity-ecosystem function (BEF) relationships. This research has found many trends showing that primary production increases as the number of unique plant species are added to a community (*e.g.*, as plant species richness increases; Tilman et al. 2014, Grace et al. 2016, Liang et al. 2016, Emmett Duffy et al. 2017). BEF relationships are generally positive and asymptotic. Therefore, biodiversity loss causes a small change in ecosystem function at first and then, at some tipping point, there is a dramatic decrease in function (Cardinale et al. 2012, Hooper et al. 2012, Tilman et al. 2014). Such trends have largely been observed in plant communities specifically focused on plant richness and plant biomass (used as a proxy for primary production) as the ecosystem function of interest. However, studies in other systems have shown a positive diversity-function relationship (Hillebrand and Cardinale 2004, Tylianakis et al. 2008, Duffy et al. 2016, Zeppilli et al. 2016) indicating that this trend extends across many ecosystem types.

Bacterial biodiversity

"While the first rationale for concern over biodiversity should apply to microbes, they lack charisma" ~ *J. Schimel* (1995)

Microbial ecologists are interested in whether these BEF relationships developed in the broader field of ecology apply to bacterial communities? Trying to answer this question poses a few challenges: First, there is a lack of a consensus regarding a bacterial species concept (Koeppel and Wu 2014). Second, typical bacterial communities have a larger range of diversity (hundreds to thousands of species) compared to macro-organismal communities (tens to hundreds of species) and the observed diversity depends on the methodology used. Third, microbes, especially bacteria, arguably have a broader range of ecosystem contributions than macroorganisms do, as their functioning is closely intertwined with all biogeochemical transformations (*e.g.*, sulfate reduction, heterotrophic respiration, iron reduction, manganese oxidation, etc; Falkowski et al. 2008) and it is undetermined if BEF theory is universal to these various bacterial functions. The same reasons that make addressing the question challenging, make it important to tackle research on BEF theory in bacterial and other microbial systems. Doing so explores how universal and generalizable these patterns are in the field of ecology across the tree of life. Yet, relative to eukaryotic organisms, fewer studies have tackled this question in a bacterial context, although some have done so across various experimental and observational ecosystems (Schimel 1995, Griffiths et al. 2000, Wohl et al. 2004, Bell et al. 2005, Fierer et al. 2007, Langenheder et al. 2010, Delgado-Baquerizo et al. 2016).

The lack of a bacterial species concept

While there are many ways to delineate species, the most common definition is the biological species concept, which defines species as organisms that have the ability to interbreed in nature and create fertile offspring (Mayr 1942). However, this concept has tenuous applications to bacteria and archaea that reproduce asexually. While homologous recombination between lineages could serve as a proxy for a microbial species (Eppley et al. 2007) and appears to decline with increasing genetic distance (Didelot and Maiden 2010), there are a few complicating factors. Recombination rates can be difficult to measure *in situ* and are variable across a genome leading to irregular rates of 'sexual' isolation across a genome (Retchless and Lawrence 2007, 2010). The rampant non-homologous recombination rates (Fraser et al. 2007) further complicate species definitions and, thus, a universal bacterial biological species concept has been difficult to attain. Instead, bacterial and archaeal species have traditionally been defined using operational definitions such as a 70% cut-off of pairwise genomic DNA-DNA hybridization (Stackebrandt and Goebel 1994). With the advent of sequencing of marker genes, particularly the 16S rRNA gene, microbial ecologists converged on the use of an operational taxonomic unit (OTU) based on 97% sequence similarity of the 16S rRNA gene. While this approach tends to group organisms with broad functional similarities, it often combines ecologically distinct populations into one OTU (Acinas et al. 2004, Hunt et al. 2008a, Denef et al. 2010b, Sharon et al. 2013). This has become increasingly clear through the application of genomics where individuals that are originally grouped as one OTU actually have highly divergent gene content (Acinas et al. 2004, Denef et al. 2010b, Morowitz et al. 2011, Shapiro et al. 2012, Shapiro and Polz 2014). Nonetheless, broad correspondence between 16S rRNA gene divergence and genomic divergence is the norm (Konstantinidis and Tiedje 2005). While the concept of a bacterial species (Fraser et al. 2007, Shapiro and Polz 2014) and the appropriate operational units to be used remain in flux (Berry et al. 2017, Callahan et al. 2017), OTUs defined by >97% 16S rRNA gene sequence identity remains the most commonly used metric in microbial ecology and will be used extensively throughout this dissertation (**Chapter II, Chapter III, & Chapter V**). However, for investigation of more detailed traits I use a genome-centric approach in **Chapter IV**.

Freshwater systems

Role of microorganisms in freshwater systems

Bacteria are the major engines that drive earth's biogeochemical cycles (Falkowski et al. 2008), and specifically in freshwater systems bacteria play a pivotal role in biogeochemical cycling of nutrients (Cotner and Biddanda 2002) and are an important source of food for organisms of higher trophic levels (Azam and Graf 1983, Azam and Malfatti 2007, Pomero et al. 2007). There are many habitats both vertically and horizontally within a lake in which different bacteria can make a living (Shade et al. 2008, Jones et al. 2012a). Specifically, the work within this dissertation analyzes the impact of three general habitat classifications on bacterial community composition: different lake layers, nutrient levels, and particulate matter (**Chapter II**), with a major focus on bacteria attached to particulate matter compared to free-living bacteria (**Chapters II, III, IV**).

Environmental forces in freshwater lakes

In many lakes, seasonal thermal stratification leads to the formation of a discrete layer in the surface, called the *epilimnion*, and a layer near the bottom, called the *hypolimnion*, which harbor contrasting bacterial communities (Garcia et al. 2013, Köllner et al. 2013). The environmental differences between the epilimnion and hypolimnion occur very rapidly in high-nutrient lakes where waters of the hypolimnion can be dramatically lower in pH and dissolved oxygen concentrations due both to microbial and physical processes. On the other hand, environmental conditions in the epilimnion and hypolimnion of low-nutrient lakes change more gradually and often light can penetrate deep into the hypolimnion. Additionally, suspended particulate matter from both abiotic (i.e., terrestrial sediment via run-off or littoral-zone clastics mixed up by wave action) and biotic (i.e., phytoplankton, zooplankton, other organisms and their detritus) sources create unique habitats for bacteria to inhabit and help to maintain bacterial diversity within freshwater lakes (Grossart 2010, Stocker 2012). Work in marine and freshwater particle-associated bacterial communities, including work presented in this dissertation, has indicated that these communities are taxonomically and functionally distinct from free-living bacteria (Zeigler Allen et al. 2012, Rösel et al. 2012, Parveen et al. 2013, Smith et al. 2013a, Bižić-Ionescu et al. 2014, Ganesh et al. 2014a, Jackson et al. 2014, Simon et al. 2014).

Particles in aquatic ecosystems have been called “hotspots of microbial activity” (Azam 1998) and “hotbeds for genome reshuffling” (Ganesh et al. 2014a) and may provide microheterogeneity in the water column that helps sustain high levels of taxonomic and metabolic diversity in aquatic microbial communities (Hunt et al. 2008a, Grossart 2010, Stocker 2012, Salcher 2014). While outnumbered by free-living bacteria (Caron et al. 1982), particle-associated bacteria are disproportionately active in organic matter mineralization, as particulate organic carbon is more readily bioavailable than the dissolved organic carbon pool on which free-living bacteria rely (Crump et al. 1998, Lemarchand et al. 2006, Ghiglione et al. 2009, Grossart 2010, Schmidt et al. 2017). Input of terrestrial runoff to aquatic ecosystems, and the formation and removal of particulate matter in aquatic ecosystems have been changing dramatically due to human activities, particularly land use change, eutrophication, and the introduction of invasive species. These human-induced changes may influence the nature, abundance or the relative importance of particles in aquatic ecosystems. As bacteria are the first biological organisms to respond to disturbances, the vulnerability of lake ecosystems to environmental change particularly depends on the bacterial community (Paerl et al. 2003). **My dissertation analyzes how particle-associated and free-living bacterial community composition and diversity varies and its corresponding effect on bacterial heterotrophic production.**

Overview of dissertation

In **Chapter II**, I determine some of the primary environmental controls on lake bacterial community composition. I assess the influence of lake layer (*i.e.* stratification), lake productivity, and particle-association on the patterns of bacterial community composition across 11 inland lakes with varying productivity in Southwestern Michigan. Next in **Chapter III**, I analyze data from a freshwater estuarine lake, Muskegon Lake, which has large spatio-temporal variation in bacterial heterotrophic productivity, to test whether there was an association between heterotrophic production and biodiversity (defined as the number of taxa and taxon abundance). I specifically focus on two co-occurring freshwater habitats, particle-associated and free-living, that chapter II shows to be populated by very distinct communities. In **Chapter IV**, I perform a genome-centric analysis of free-living and particle-associated bacterial specialists and generalists to see which genomic traits were associated with adaptations to these specific habitats. Finally, in **Chapter V**, I identified the bacterial taxa driving two aquatic function groups (one of which is important for heterotrophic productivity) across the large set of lake samples I collected throughout my dissertation. These assessments of aquatic bacterial diversity, and their genomic underpinnings, will help provide a mechanistic view into drivers of aquatic bacterial community diversity and composition and their influence on metabolic function.

References

Acinas, S. G., V. Klepac-Ceraj, D. E. Hunt, C. Pharino, I. Ceraj, D. L. Distel, and M. F. Polz. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430:551–554.

Alneberg, J., B. S. Bjarnason, I. De Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. 2014. Binning metagenomic contigs by coverage and composition. Nature Methods 11:1144–1146.

Amann, R., and B. M. Fuchs. 2008. Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. Nature Reviews Microbiology 6:339–348.

Amann, R. I., L. Krumholz, and D. A. Stahl. 1990. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. Journal of Bacteriology 172:762–770.

Azam, F. 1998. Microbial Control of Oceanic Carbon Flux: The Plot Thickens. Science 280:694–696.

Azam, F., and J. S. Graf. 1983. The Ecological Role of Water-Column Microbes in the Sea. Marine Ecology 10:257–263.

Azam, F., and F. Malfatti. 2007. Microbial structuring of marine ecosystems. Nature reviews. Microbiology 5:782–791.

Bell, T., J. A. Newman, B. W. Silverman, S. L. Turner, and A. K. Lilley. 2005. The contribution of species richness and composition to bacterial services. Nature 436:1157–1160.

Berry, M. A., J. D. White, T. W. Davis, S. Jain, H. Thomas, G. J. Dick, O. Sarnelle, and V. J. Denef. 2017. Are oligotypes meaningful ecological and phylogenetic units? A case study of Microcystis in freshwater lakes 8:1–7.

Bižić-Ionescu, M., M. Zeder, D. Ionescu, S. Orlić, B. M. Fuchs, H.-P. Grossart, and R. Amann. 2014. Comparison of bacterial communities on limnic versus coastal marine particles reveals profound differences in colonization. Environmental microbiology:1–36.

Callahan, B. J., P. J. McMurdie, and S. P. Holmes. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME Journal 11:2639–2643.

Cardinale, B. J., J. E. Duffy, A. Gonzalez, D. U. Hooper, C. Perrings, P. Venail, A. Narwani, G. M. Mace, D. Tilman, D. A.Wardle, A. P. Kinzig, G. C. Daily, M. Loreau, J. B. Grace, A. Larigauderie, D. S. Srivastava, and S. Naeem. 2012. Biodiversity loss and its impact on humanity. Nature 489:326–326.

Caron, D. A., P. G. Davis, L. P. Madin, and J. M. Sieburth. 1982. Heterotrophic bacteria and bacterivorous protozoa in oceanic macroaggregates. Science 218:795–797.

Cotner, J., and B. Biddanda. 2002. Small Players , Large Role : Microbial Influence on Biogeochemical Processes in Pelagic Aquatic Ecosystems. Ecosystems 5:105–121.

Crump, B. C., J. A. Baross, and C. A. Simenstad. 1998. Dominance of particle-attached bacteria in the Columbia River estuary, USA. Aquatic Microbial Ecology 14:7–18.

Dekas, A. E., S. A. Connon, G. L. Chadwick, E. Trembath-Reichert, and V. J. Orphan. 2016. Activity and interactions of methane seep microorganisms assessed by parallel transcription and FISH-NanoSIMS analyses. ISME Journal 10:678–692.

Dekas, A. E., R. S. Poretsky, and V. J. Orphan. 2009. Deep-Sea archaea fix and share nitrogen in methane-consuming microbial consortia. Science 326:422–426.

Delgado-Baquerizo, M., L. Giaramida, P. B. Reich, A. N. Khachane, K. Hamonts, C. Edwards, L. A. Lawton, and B. K. Singh. 2016. Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. Journal of Ecology 104:936–946.

Denef, V. J., L. H. Kalnejais, R. S. Mueller, P. Wilmes, B. J. Baker, B. C. Thomas, N. C. VerBerkmoes, R. L. Hettich, and J. F. Banfield. 2010. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proceedings of the National Academy of Sciences of the United States of America 107:2383–90.

Didelot, X., and M. C. J. Maiden. 2010. Impact of recombination on bacterial evolution. Trends in Microbiology 18:315–322.

Duffy, J. E., J. S. Lefcheck, R. D. Stuart-Smith, S. A. Navarrete, and G. J. Edgar. 2016. Biodiversity enhances reef fish biomass and resistance to climate change. Proceedings of the National Academy of Sciences 113:6230–6235.

Emmett Duffy, J., C. M. Godwin, and B. J. Cardinale. 2017. Biodiversity effects in the wild are common and as strong as key drivers of productivity. Nature 549:261–264.

Eppley, J. M., G. W. Tyson, W. M. Getz, and J. F. Banfield. 2007. Genetic exchange across a species boundary in the archaeal genus ferroplasma. Genetics 177:407–416.

Eren, A. M., Ö. C. Esen, C. Quince, J. H. Vineis, H. G. Morrison, M. L. Sogin, and T. O. Delmont. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319.

Eren, A. M., L. Maignien, W. J. Sul, L. G. Murphy, S. L. Grim, H. G. Morrison, and M. L. Sogin. 2013. Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods in Ecology and Evolution 4:1111–1119.

Falkowski, P. G., T. Fenchel, and E. F. Delong. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–9.

Fierer, N., M. a. Bradford, and R. B. Jackson. 2007. Toward an ecological classification of soil bacteria. Ecology 88:1354–1364.

Finzi-Hart, J. A., J. Pett-Ridge, P. K. Weber, R. Popa, S. J. Fallon, T. Gunderson, I. D. Hutcheon, K. H. Nealson, and D. G. Capone. 2009. Fixation and fate of C and N in the cyanobacterium Trichodesmium using nanometer-scale secondary ion mass spectrometry. Proceedings of the National Academy of Sciences 106:6345–6350.

Fraser, C., W. P. Hanage, and B. G. Spratt. 2007. Recombination and the nature of bacterial speciation. Science 315:476–480.

Ganesh, S., D. J. Parris, E. F. DeLong, and F. J. Stewart. 2014. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. The ISME journal 8:187–211.

Garcia, S. L., I. Salka, H. P. Grossart, and F. Warnecke. 2013. Depth-discrete profiles of bacterial communities reveal pronounced spatio-temporal dynamics related to lake stratification. Environmental Microbiology Reports 5:549–555.

Ghiglione, J. F., P. Conan, and M. Pujo-Pay. 2009. Diversity of total and active free-living vs. particle-attached bacteria in the euphotic zone of the NW Mediterranean Sea. FEMS Microbiology Letters 299:9–21.

Grace, J. B., T. M. Anderson, E. W. Seabloom, E. T. Borer, P. B. Adler, W. S. Harpole, Y. Hautier, H. Hillebrand, E. M. Lind, M. Pärtel, J. D. Bakker, Y. M. Buckley, M. J. Crawley, E. I. Damschen, K. F. Davies, P. A. Fay, J. Firn, D. S. Gruner, A. Hector, J. M. H. Knops, A. S. MacDougall, B. A. Melbourne, J. W. Morgan, J. L. Orrock, S. M. Prober, and M. D. Smith. 2016. Integrative modelling reveals mechanisms linking productivity and plant species richness. Nature 529:390–393.

Griffiths, B. S., K. Ritz, R. D. Bardgett, R. Cook, S. Christensen, F. Ekelund, S. J. Sorensen, E. Baath, J. Bloem, P. C. de Ruiter, J. Dolfing, and B. Nicolardot. 2000. Ecosystem response of pasture soil communities to fumigation-induced microbial diversity reductions: an examination of the biodiversity-ecosystem function relationship. Oikos 90:279–294.

Grossart, H. P. 2010. Ecological consequences of bacterioplankton lifestyles: Changes in concepts are needed. Environmental Microbiology Reports 2:706–714.

Herrmann, A. M., K. Ritz, N. Nunan, P. L. Clode, J. Pett-Ridge, M. R. Kilburn, D. V. Murphy, A. G. O'Donnell, and E. A. Stockdale. 2007. Nano-scale secondary ion mass spectrometry - A new analytical tool in biogeochemistry and soil ecology: A review article. Soil Biology and Biochemistry 39:1835–1850.

Hillebrand, H., and B. J. Cardinale. 2004. Consumer effects decline with prey diversity. Ecology Letters 7:192–201.

Hooper, D. U., E. C. Adair, B. J. Cardinale, J. E. K. Byrnes, B. A. Hungate, K. L. Matulich, A. Gonzalez, J. E. Duffy, L. Gamfeldt, and M. I. Connor. 2012. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. Nature 486:105–108.

Hooper, D. U., F. S. Chapin, J. J. Ewel, A. Hector, P. Inchausti, S. Lavorel, J. H. Lawton, D. M. Lodge, M. Loreau, S. Naeem, B. Schmid, H. Setälä, a J. Symstad, and D. a Wardle. 2005. Effects of Biodiversity on Ecosystem Functioning: A Consensus of Current Knowledge. Ecological Monographs 75:3–35.

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320:1081–1085.

Jackson, C. R., J. J. Millar, J. T. Payne, and C. A. Ochs. 2014. Free-Living and Particle-Associated Bacterioplankton in Large Rivers of the Mississippi River Basin Demonstrate Biogeographic Patterns. Applied and Environmental Microbiology 80:7186–7195.

Jones, S. E., T. a. Cadkin, R. J. Newton, and K. D. McMahon. 2012. Spatial and temporal scales of aquatic bacterial beta diversity. Frontiers in Microbiology 3:1–10.

Kang, D. D., J. Froula, R. Egan, and Z. Wang. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165.

Kembel, S. W., M. Wu, J. A. Eisen, and J. L. Green. 2012. Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. PLoS Computational Biology 8:16–18.

Koeppel, A. F., and M. Wu. 2014. Species matter: the role of competition in the assembly of congeneric bacteria. The ISME journal 8:531–40.

Köllner, K., D. Carstens, C. Schubert, J. Zeyer, and H. Bürgmann. 2013. Impact of particulate organic matter composition and degradation state on the vertical structure of particle-associated and planktonic lacustrine bacteria. Aquatic Microbial Ecology 69:81–92.

Konstantinidis, K. T., and J. M. Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences of the United States of America 102:2567–2572.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Applied and Environmental Microbiology 79:5112–5120.

Laczny, C. C., T. Sternal, V. Plugaru, P. Gawron, A. Atashpendar, H. H. Margossian, S. Coronado, L. V. der Maaten, N. Vlassis, and P. Wilmes. 2015. VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3:1–7.

Langenheder, S., M. T. Bulling, M. Solan, and J. I. Prosser. 2010. Bacterial Biodiversity-Ecosystem Functioning Relations are Modified by Environmental Complexity. PLoS ONE 5.

Langille, M. G. I., J. Zaneveld, J. G. Caporaso, D. McDonald, D. Knights, J. a Reyes, J. C. Clemente, D. E. Burkepile, R. L. Vega Thurber, R. Knight, R. G. Beiko, and C. Huttenhower. 2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature biotechnology 31:814–21.

Lemarchand, C., L. Jardillier, J. F. Carrias, M. Richardot, D. Debroas, T. Sime-Ngando, and C. Amblard. 2006. Community composition and activity of prokaryotes associated to detrital particles in two contrasting lake ecosystems. FEMS Microbiology Ecology 57:442–451.

Liang, J., T. W. Crowther, N. Picard, S. Wiser, M. Zhou, G. Alberti, E. D. Schulze, A. D. McGuire, F. Bozzato, H. Pretzsch, S. De-Miguel, A. Paquette, B. Hérault, M. Scherer-Lorenzen, C. B. Barrett, H. B. Glick, G. M. Hengeveld, G. J. Nabuurs, S. Pfautsch, H. Viana, A. C. Vibrans, C. Ammer, P. Schall, D. Verbyla, N. Tchebakova, M. Fischer, J. V. Watson, H. Y. H. Chen, X. Lei, M. J. Schelhaas, H. Lu, D. Gianelle, E. I. Parfenova, C. Salas, E. Lee, B. Lee, H. S. Kim, H. Bruelheide, D. A. Coomes, D. Piotto, T. Sunderland, B. Schmid, S. Gourlet-Fleury, B. Sonké, R. Tavani, J. Zhu, S. Brandl, J. Vayreda, F. Kitahara, E. B. Searle, V. J. Neldner, M. R. Ngugi, C. Baraloto, L. Frizzera, R. Bałazy, J. Oleksyn, T. Zawiła-Niedźwiecki, O. Bouriaud, F. Bussotti, L. Finér, B. Jaroszewicz, T. Jucker, F.

Valladares, A. M. Jagodzinski, P. L. Peri, C. Gonmadje, W. Marthy, T. O'Brien, E. H. Martin, A. R. Marshall, F. Rovero, R. Bitariho, P. A. Niklaus, P. Alvarez-Loayza, N. Chamuya, R. Valencia, F. Mortier, V. Wortel, N. L. Engone-Obiang, L. V. Ferreira, D. E. Odeke, R. M. Vasquez, S. L. Lewis, and P. B. Reich. 2016. Positive biodiversity-productivity relationship predominant in global forests. Science 354.

Loreau, M., S. Naeem, P. Inchausti, J. Bengtsson, J. P. Grime, A. Hector, D. U. Hooper, M. A. Huston, D. Raffaelli, B. Schmid, D. Tilman, and D. A. Wardle. 2001. Biodiversity and ecosystem functioning: current knowledge and future challenges. Science 294:804–808.

Louca, S., M. Doebeli, and L. W. Parfrey. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6:1–12.

Mayr, E. 1942. Systematics and the origin of species. Columbia University Press.

Morowitz, M. J., V. J. Denef, E. K. Costello, B. C. Thomas, V. Poroyko, D. A. Relman, and J. F. Banfield. 2011. Strain-resolved community genomic analysis of gut microbial colonization in a premature infant. Proceedings of the National Academy of Sciences 108:1128–1133.

Olm, M. R., C. T. Brown, B. Brooks, and J. F. Banfield. 2017. DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME Journal 11:2864–2868.

Paerl, H. W., J. Dyble, P. H. Moisander, R. T. Noble, M. F. Piehler, J. L. Pinckney, T. F. Steppe, L. Twomey, and L. M. Valdes. 2003. Microbial indicators of aquatic ecosystem change: Current applications to eutrophication studies. FEMS Microbiology Ecology 46:233–246.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25:1043–1055.

Parveen, B., I. Mary, A. Vellet, V. Ravet, and D. Debroas. 2013. Temporal dynamics and phylogenetic diversity of free-living and particle-associated Verrucomicrobia communities in relation to environmental variables in a mesotrophic lake. FEMS Microbiology Ecology 83:189–201.

Pomero, L., P. Williams, F. Azam, and J. Hobbie. 2007. The microbial loop. Oceanography 20:28–33.

Popa, R., P. K. Weber, J. Pett-Ridge, J. A. Finzi, S. J. Fallon, I. D. Hutcheon, K. H. Nealson, and D. G. Capone. 2007. Carbon and nitrogen fixation and metabolite exchange in and between individual cells of Anabaena oscillarioides. ISME Journal 1:354–360.

Preheim, S. P., A. R. Perrott, A. M. Martin-Platero, A. Gupta, and E. J. Alm. 2013. Distribution-based clustering: Using ecology to refine the operational taxonomic unit. Applied and Environmental Microbiology 79:6593–6603.

Props, R., F. M. Kerckhof, P. Rubbens, J. De Vrieze, E. H. Sanabria, W. Waegeman, P. Monsieurs, F. Hammes, and N. Boon. 2017a. Absolute quantification of microbial taxon abundances. ISME Journal 11:584–587.

Props, R., P. Monsieurs, M. Mysara, L. Clement, and N. Boon. 2016. Measuring the biodiversity of microbial communities by flow cytometry. Methods in Ecology and Evolution 7:1376–

1385.

Props, R., M. L. Schmidt, J. Heyse, H. A. Vanderploeg, N. Boon, and V. J. Denef. 2017b. Flow cytometric monitoring of bacterioplankton phenotypic diversity predicts high population-specific feeding rates by invasive dreissenid mussels. Environmental Microbiology 00.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41:590–596.

Retchless, A. C., and J. G. Lawrence. 2007. Temporal Fragmentation of Speciation in Bacteria. Science 317:1093–1096.

Retchless, A. C., and J. G. Lawrence. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. Proceedings of the National Academy of Sciences 107:11453–11458.

Rohwer, R. R., J. J. Hamilton, R. J. Newton, and K. D. McMahon. 2017. TaxAss: Leveraging Custom Databases Achieves Fine-Scale Taxonomic Resolution. bioRxiv:214288.

Rösel, S., M. Allgaier, and H.-P. Grossart. 2012. Long-term characterization of free-living and particle-associated bacterial communities in Lake Tiefwaren reveals distinct seasonal patterns. Microbial ecology 64:571–583.

Salcher, M. M. 2014. Same same but different: Ecological niche partitioning of planktonic freshwater prokaryotes. Journal of Limnology 73:74–87.

Schimel, J. P. 1995. Ecosystem consequences of microbial diversity and community structure. Pages 239–251 *in* F. S. Chapin III and C. Korner, editors. Arctic and alpine biodiversity: Patterns, causes and ecosystem consequences. Springer-Verlag.

Schmidt, M. L., B. A. Biddanda, A. D. Weinke, E. Chiang, F. Januska, R. Props, and V. J. Denef. 2017. Microhabitats shape diversity-productivity relationships in freshwater bacterial communities. bioRxiv:231688.

Sczyrba, A., P. Hofmann, P. Belmann, D. Koslicki, S. Janssen, J. Dröge, I. Gregor, S. Majda, J. Fiedler, E. Dahms, A. Bremges, A. Fritz, R. Garrido-Oter, T. S. Jørgensen, N. Shapiro, P. D. Blood, A. Gurevich, Y. Bai, D. Turaev, M. Z. Demaere, R. Chikhi, N. Nagarajan, C. Quince, F. Meyer, M. Balvočiutė, L. H. Hansen, S. J. Sørensen, B. K. H. Chia, B. Denis, J. L. Froula, Z. Wang, R. Egan, D. Don Kang, J. J. Cook, C. Deltel, M. Beckstette, C. Lemaitre, P. Peterlongo, G. Rizk, D. Lavenier, Y. W. Wu, S. W. Singer, C. Jain, M. Strous, H. Klingenberg, P. Meinicke, M. D. Barton, T. Lingner, H. H. Lin, Y. C. Liao, G. G. Z. Silva, D. A. Cuevas, R. A. Edwards, S. Saha, V. C. Piro, B. Y. Renard, M. Pop, H. P. Klenk, M. Göker, N. C. Kyrpides, T. Woyke, J. A. Vorholt, P. Schulze-Lefert, E. M. Rubin, A. E. Darling, T. Rattei, and A. C. McHardy. 2017. Critical Assessment of Metagenome Interpretation - A benchmark of metagenomics software. Nature Methods 14:1063–1071.

Shade, A., S. E. Jones, and K. D. McMahon. 2008. The influence of habitat heterogeneity on freshwater bacterial community composition and dynamics. Environmental Microbiology 10:1057–1067.

Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabo, M. F.

Polz, and E. J. Alm. 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336:48–51.

Shapiro, B. J., and M. F. Polz. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends in Microbiology 22:235–247.

Sharon, I., M. J. Morowitz, B. C. Thomas, E. K. Costello, D. A. Relman, and J. F. Banfield. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. Genome Research 23:111–120.

Sieber, C. M. K., A. J. Probst, A. Sharrar, B. C. Thomas, M. Hess, S. G. Tringe, and J. F. Banfield. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 3:836–843.

Simon, H. M., M. W. Smith, and L. Herfort. 2014. Metagenomic insights into particles and their associated microbiota in a coastal margin ecosystem. Frontiers in Microbiology 5:466.

Singh, B. K., R. D. Bardgett, P. Smith, and D. S. Reay. 2010. Microorganisms and climate change: terrestrial feedbacks and mitigation options. Nature Reviews Microbiology Microbiology 8:779–90.

Smith, M. W., L. Z. Allen, A. E. Allen, L. Herfort, and H. M. Simon. 2013. Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. Frontiers in Microbiology 4:1–20.

Stackebrandt, E., and B. M. Goebel. 1994. Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. International Journal of Systematic and Evolutionary Microbiology 44:846–849.

Stewart, E. J. 2012. Growing unculturable bacteria. Journal of Bacteriology 194:4151–4160.

Stocker, R. 2012. Marine microbes see a sea of gradients. Science 338:628–33.

Stoddard, S. F., B. J. Smith, R. Hein, B. R. K. Roller, and T. M. Schmidt. 2015. rrnDB: Improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Research 43:D593–D598.

Thomas, C. D., A. Cameron, R. E. Green, M. Bakkenes, L. J. Beaumont, Y. C. Collingham, B. F. N. Erasmus, M. F. De Siqueira, A. Grainger, L. Hannah, L. Hughes, B. Huntley, A. S. Van Jaarsveld, G. F. Midgley, L. Miles, M. A. Ortega-Huerta, A. T. Peterson, O. L. Phillips, and S. E. Williams. 2004. Extinction risk from climate change. Nature 427:145–8.

Tilman, D., F. Isbell, and J. M. Cowles. 2014. Biodiversity and Ecosystem Functioning. Annual Review of Ecology, Evolution, and Systematics 45:471–493.

Tylianakis, J. M., T. A. Rand, A. Kahmen, A. M. Klein, N. Buchmann, J. Perner, and T. Tscharntke. 2008. Resource heterogeneity moderates the biodiversity-function relationship in real world ecosystems. PLoS Biology 6:0947–0956.

Wake, D. B., and V. T. Vredenburg. 2008. Are we in the midst of the sixth mass extinction? A view from the world of amphibians. Proceedings of the National Academy of Sciences 105:11466–11473.

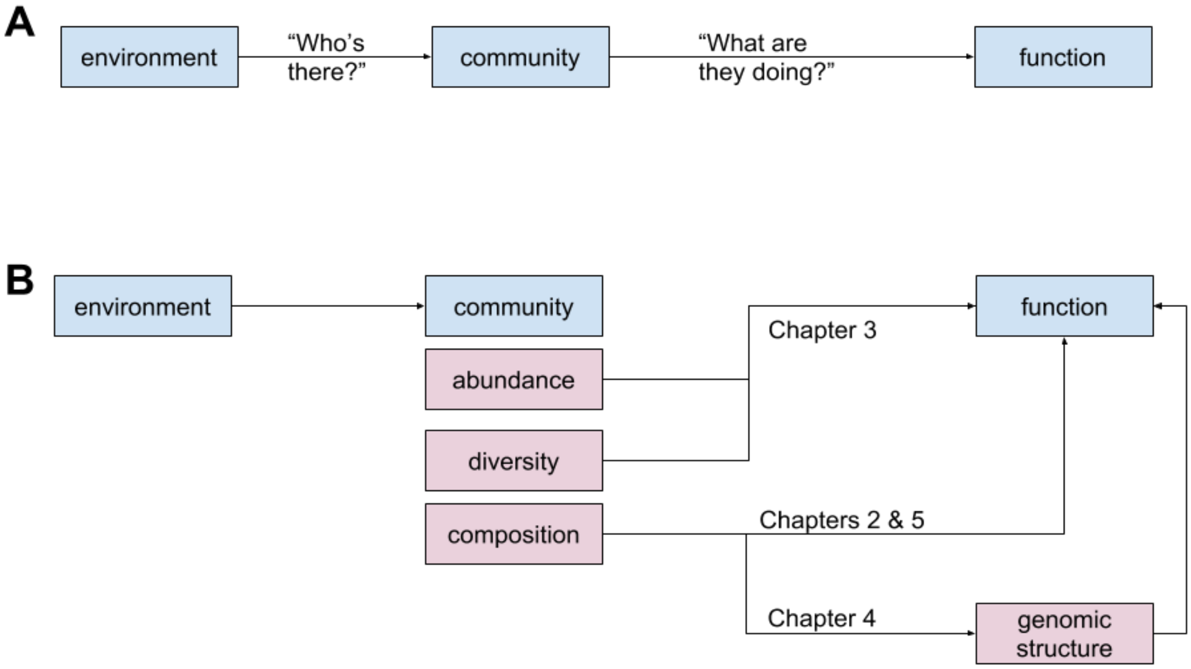
Wohl, D. L., S. Arora, and J. R. Gladstone. 2004. Functional redundancy supports biodiversity and ecosystem function in a closed and constant environment. Ecology 85:1534–1540.

Wu, Y. W., Y. H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer. 2014. MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:1–18.

Zeigler Allen, L., E. E. Allen, J. H. Badger, J. P. McCrow, I. T. Paulsen, L. D. Elbourne, M. Thiagarajan, D. B. Rusch, K. H. Nealson, S. J. Williamson, J. C. Venter, and A. E. Allen. 2012. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME Journal 6:1403–1414.

Zeppilli, D., A. Pusceddu, F. Trincardi, and R. Danovaro. 2016. Seafloor heterogeneity influences the biodiversity–ecosystem functioning relationships in the deep sea. Scientific Reports 6:26352.

Figure 1.1. Conceptual overview of this dissertation.



Chapter II:
Phylogenetic conservation of freshwater lake habitat preference varies between abundant bacterioplankton[1]

ABSTRACT

Despite their homogeneous appearance, aquatic systems harbor heterogeneous habitats resulting from nutrient gradients, suspended particulate matter, and stratification. Recent reports suggest phylogenetically conserved habitat preferences among bacterioplankton, particularly for particle-associated (PA) and free-living (FL) habitats. Here, we show that independent of lake nutrient level and layer, PA and FL abundance-weighted bacterial community composition (BCC) differed and that inter-lake BCC varied more for PA than FL fractions. In low-nutrient lakes, BCC differences between PA and FL fractions were larger than between lake layers. The reverse was true for high-nutrient lakes. Nutrient level affected BCC more in hypolimnia than in epilimnia, likely due to hypolimnetic hypoxia in high-nutrient lakes. In line with previous reports, we observed within-phylum OTU habitat preference conservation, though not for all phyla, including the phylum with the highest average relative abundance across all habitats (*Bacteroidetes*). Consistent phylum level habitat preferences may indicate that the functional traits that underpin ecological adaptation of freshwater bacteria to lake habitats can be phylogenetically conserved, though levels of conservation are phylum-dependent. Resolving taxa preferences for freshwater habitats sets the stage for identification of traits that underpin habitat specialization and associated functional traits that influence differences in biogeochemical cycling across freshwater lake habitats.

[1] Published as: Schmidt ML, White JD, Denef VJ. Phylogenetic conservation of freshwater lake habitat preference varies between abundant bacterioplankton phyla. *Environ Microbiol* 2016; **18**: 1212–1226.

Introduction

In freshwater lakes, bacteria are key players in organic carbon processing, nutrient remineralization, and form the base of the microbial food web (Wetzel 2001). While the pelagic zone of lakes may appear to be uniform and unstructured and is often sampled as if it were (Grossart 2010), bacterial community composition (BCC) varies both horizontally and vertically within a lake (Shade et al. 2008, Jones et al. 2012b). In temperate lakes, seasonal thermal stratification lead to the formation of discrete water masses (i.e., epilimnion and hypolimnion) that harbor contrasting bacterial communities (Garcia et al. 2013, Köllner et al. 2013, Paganin et al. 2013). Lake nutrient levels and resulting productivity strongly impact the environmental conditions within lake layers, especially in the hypolimnia of more productive lakes where the accumulation of organic matter can lead to oxygen depletion and decreasing pH by bacteria and changes in the BCC (Lindström 2000, Yannarell et al. 2003, Jankowski et al. 2014).

Within these large spatial habitats (lakes with different nutrient levels and lake layers), particulate matter provides additional habitat heterogeneity in the water column and thus, helps sustain high levels of diversity in aquatic microbial communities (Hunt et al. 2008a, Grossart 2010, Salcher 2014). Studies in marine and freshwater systems indicate that these communities are taxonomically and functionally distinct from free-living bacteria (Grossart 2010, Zeigler Allen et al. 2012, Smith et al. 2013a, Ganesh et al. 2014b, Mohit et al. 2014, Simon et al. 2014). Moreover, recent observations in bathypelagic marine systems suggest that the preference of bacteria between a particle-associated (particulate organic matter) and free-living (dissolved organic matter) habitats is highly phylogenetically conserved, up to the class to phylum level (Salazar et al. 2015).

In this study, we investigated two main questions. First, how is freshwater bacterioplankton community composition simultaneously shaped by three habitat types: filter fraction [free-living (FL; 0.22-3 μm) or particle-associated (PA; 3-20 μm], lake layer (epilimnion or hypolimnion), and nutrient level (high- or low-nutrient, based on total phosphorus)? Second, are habitat preferences phylogenetically conserved? Using a dataset from 11 north-temperate freshwater lakes varying widely in productivity, we characterized BCC in these habitats at one time point during the summer-stratified period via high throughput Illumina sequencing of the V4 hyper-variable region of the 16S rRNA gene. Our findings contribute to a growing understanding of the importance of the habitat heterogeneity encountered within the freshwater pelagic environment (Salcher 2014) and expands recent efforts to determine the extent to which habitat specialization is a phylogenetically conserved trait among aquatic bacteria.

Results

Limnological Data

Based on epilimnetic total phosphorus (TP) concentrations (range: 7.6 - 47.8 µg/L), we categorized our lakes as four low-nutrient (TP ≤ 10 µg/L) and seven high-nutrient (TP > 10 µg/L) lakes (**Table 2.1**). This is slightly lower than the cutoff proposed between oligotrophic and eutrophic lakes (12 µg/L; Carlson 1977), though corresponds to the distinction between lakes with and without significant hypolimnetic hypoxia in our dataset **(Figure SI 2.6)**. Ten of the lakes exhibited mid-summer stratification. At the time of sampling, Sherman Lake was isothermal due to artificial de-stratification by aerators, so we treated it separately, classifying it as “Mixed”. Vertical profiles of the stratified lakes indicated summer gradients of temperature, pH, and dissolved oxygen (DO) typical of their respective nutrient levels (Wetzel 2001**; Table 2.1 & Figure SI 2.6)**. In all of the stratified high-nutrient lakes, DO concentrations were depleted (0 - 1 mg/L) throughout the hypolimnion, whereas DO concentrations were always ≥ 5 mg/L at the sampling depths in the four low-nutrient lakes **(Figure SI 2.6)**.

Differences in Bacterial Community Composition

We calculated how many OTUs were unique to each habitat and shared between habitat types. The highest number of unique OTUs was in high-nutrient lakes, which was significantly greater than in low-nutrient lakes (Chi-sq, $p < 2.2 \times 10^{-16}$; **Figure 2.1**). Hypolimnia harbored significantly more unique OTUs than epilimnia (Chi-sq, $p < 2.2 \times 10^{-16}$). The FL fraction had significantly more unique OTUs than compared with the PA fraction (Chi-sq, $p < 2 \times 10^{-6}$; **Figure 2.1 left panel**). The number of shared OTUs was significantly higher in PA and FL communities as compared to the other two comparisons (Chi-sq, $p < 2.2 \times 10^{-16}$).

While the majority of the OTUs in each habitat were unique, they were quite rare in terms of their abundance (**Figure 2.1, right panel**). OTUs that were shared between sample types represented 98% of reads in the free-living (2,741 single and doubletons) and particle-associated samples (2,252 single and doubletons), 96% of reads in low-nutrient lakes (1,625 single or doubletons), 91% of reads in high-nutrient lakes (3,500 single and doubletons), 97% of reads in the epilimnion (2,315 single and doubletons), and 93% of reads in the hypolimnion (2,829 single and doubletons).

To discover the habitat that was the most important determinant of the BCC, we created a non-metric multidimensional scaling (NMDS) plot and performed nested PERMANOVA tests with the Sørensen and Bray-Curtis dissimilarity metrics. When considering OTU presence or absence based on Sørensen dissimilarity (**Figure 2.1A**), samples clustered primarily by lake layer (NMDS1, 13.3% PERMANOVA, $p < 0.001$, **Table 2.2**) and secondarily by nutrient level (NMDS2, 11.4% PERMANOVA, $p < 0.001$). However, when considering the relative abundance of taxa based on Bray-Curtis dissimilarity (**Figure 2.2B**), samples clustered primarily by lake layer (NMDS1, 15.8% PERMANOVA, $p <0.001$) and secondarily by PA and FL filter fractions (NMDS2, 12.7% PERMANOVA, $p < 0.001$). In terms of the relative abundance of OTUs, lake layer and PA and FL fractions explained similar and independent amounts of variation (12.7 – 15.8%, **Table 2.2**) of the community composition. In contrast, when considering OTU presence or absence only, lake layer explained 13.3% while PA and FL fractions only explained 4.5% of the variation. Lake layer explained a higher proportion of BCC variation in high-nutrient (30.1% PERMANOVA, $p < 0.001$) lakes as compared to low-nutrient lakes (14.7% PERMANOVA, $p < 0.001$), primarily due to differences in the presence or absence of OTUs (**Table 2.2**). Similarly, nutrient level explained a higher proportion of variation between lake hypolimnia (32.4% PERMANOVA, $p < 0.001$) than between lake epilimnia (11.2 % PERMANOVA, $p < 0.001$) due to the presence or absence of OTUs (**Table 2.2**). In the hypolimnion and in high-nutrient lakes, filter fraction significantly differed based on OTU relative abundance (16.8% and 13.8%, repectively, PERMANOVA, $p < 0.001$) but was not as strongly influenced by OTU presence or absence (**Table 2.2**). Dissolved oxygen, temperature, and pH all co-varied with our defined habitats and each only explained an additional 2.6 – 9.9% of BCC variation between all samples and was usually insignificant (**Table 2.2**).

We assessed whether lake-to-lake variability in BCC differed depending on lake habitat. In terms of OTU presence or absence (**Figure 2.3, top**), lake-to-lake variability was significantly lower in the high-nutrient hypolimnion PA and FL samples (KW; $p = 2.4 \times 10^{-7}$). In terms of relative abundance of OTUs (**Figure 2.3, bottom**), lake-to-lake variability was higher for PA than for FL BCC, however this difference was only significant in low-nutrient lakes within the hypolimnion (KW; $p = 3.5 \times 10^{-5}$).

Significant Changes in BCC at the Phylum Level

On average, *Bacteroidetes* was the most abundant phylum across all samples (**Figure 2.4; Figure SI 2.10; Figure SI 2.11, right panel**) and showed limited differential abundance between the three types of habitat comparisons (filter fraction, lake layer, and nutrient level). (*See **Figure 2.4** for a general overview of differences in phylum relative abundance between habitat types, e.g., all FL vs all PA samples; **Figure SI 2.10** for differences in phylum relative abundance between all specific comparisons that control for variation in the other habitat types, e.g., FL vs. PA in the epilimnion of low nutrient lakes; and **Figure SI 2.11** for the statistical evaluation between specific comparisons within **Figure SI 2.10***). *Actinobacteria* was the most frequently differentially abundant phylum with significant differences in 8 of the 13 comparisons (across the x-axis of **Figure SI 2.11**). They were consistently more prevalent in FL fractions compared to PA fractions, low-nutrient relative to high-nutrient lakes, and in lake epilimnia relative to hypolimnia (**Figure 2.4**).

In line with the significant BCC differences between PA and FL communities (**Figure 2.3; Table 2.2**) differentially abundant phyla were found between PA and FL fractions. *Cyanobacteria* and *Planctomycetes* were differentially abundant in PA relative to FL fractions, while the reverse was true for *Actinobacteria* and *Betaproteobacteria* (**Figure 2.4;** DEseq2 $p < 0.01$). When refining our analysis by comparing PA and FL fractions in each combination of lake layer and nutrient level, additional differentially represented phyla were identified (**Figure SI 2.11**). For example, *Chloroflexi* was differentially abundant in the particle-associated fractions in the hypolimnion of high-nutrient lakes and Candidate Division OD1 was differentially abundant in the free-living fractions in the epi- and hypolimnion of high-nutrient lakes and the hypolimnion of low-nutrient lakes. Interestingly, there were no differentially abundant phyla within either filter fraction of the epilimnion of low-nutrient lakes.

Differential abundance between high- and low-nutrient lakes as well as between hypolimnia and epilimnia was mainly detected for phyla with low average relative abundance across the entire dataset. For example, *Lentisphaerae* was differentially abundant within the high-nutrient (relative to low-nutrient) and hypolimnion (relative to epilimnion) lake habitats while *Armatimonadetes* was differentially abundant within low-nutrient (relative to high-nutrient) and hypolimnion (relative to epilimnion) lake habitats (**Figure 2.4**). When refining our analyses by controlling for variation within the specific habitat types, the more abundant phyla were found to be differentially abundant as well (**Figure SI 2.10; Figure SI 2.11**). For example, *Actinobacteria* and *Alphaproteobacteria* were differentially abundant in low-nutrient hypolimnion samples (both PA and FL) and within the epilimnion of high nutrient lakes (both PA and FL). Many of the same phyla that were differentially abundant between high-nutrient lake layers were also differentially abundant between high- and low-nutrient hypolimnia. Most notable, *Verrucomicrobia* and *Armatimonadetes* were significantly over-represented in low-nutrient (relative to high-nutrient) hypolimninia while many phyla such as *Lentisphaerae, Chlorobi, Firmicutes, Spirochaetae*, NPL-UPA2, *Deinococcus-Thermus,* Candidate Division OP3 and Candidate Division SR1 were over-represented in all high-nutrient relative to low-nutrient hypolimnia (**Figure SI 2.8; Figure SI 2.9**).

Significant Changes in BCC at the OTU-Level

We calculated the fraction of OTUs that showed significant preference for specific habitats (e.g., PA vs. FL fraction of low-nutrient epilimnia) and summed these for the three types of habitat comparisons (*e.g.*, all PA vs. all FL; **Figure 2.5, left panel**). OTUs with habitat preference accounted for 0.2 – 27% of all OTUs within a specific phylum (**Figure 2.5, left panel**; Genus-level differential abundance for each specific habitat comparison is presented in **Figure SI 2.12**). When weighing these fractions with the relative abundance of the OTUs, OTUs with habitat preference accounted for up to 97% of the relative abundance of a phylum in a lake habitat (**Figure 2.5, right panel**). For example, *Armatimonadetes* had 1 significant OTU of 16 total

OTUs in the low-nutrient lakes, however, this significant OTU accounted for 95% of *Armatimonadetes* in low-nutrient lakes.

In most cases, if a phylum contained OTUs with habitat preference, the habitat that was preferred was consistent for all differentially abundant OTUs within that phylum (**Figure 2.5**). For example, *Chloroflexi* had significant OTUs in the PA (6%), high nutrient (15.6%) and hypolimnion (24.4%) samples. When including the relative abundance of the OTUs, the proportion of OTUs that showed significant habitat preferences was variable depending on the phylum and the habitat comparison that was made. For example, most *Actinobacteria* OTUs showed preference for the FL fraction, though very little preference between lake layers and nutrient levels; **Figure 2.5**).

It is important to note that there was a very large overlap between the lake layer and nutrient comparisons. There were 258 significant OTUs that were differentially abundant in the high- relative to low-nutrient samples and all except 8 of these OTUs were also differentially abundant in hypolimnion relative to epilimnion samples (486 significant OTUs total). In addition, the most differentially abundant OTU in the high- relative to low-nutrient samples, which belonged to the *Betaproteobacteria* tribe betI-A, was also the most differentially abundant OTU in the hypolimnion samples (**Table 2.3**). Therefore, many of the OTUs that were differentially abundant in high-nutrient lakes were only differentially abundant within the hypolimnion of these lakes.

Discussion

Spatial variation is one of the factors that help explain the “paradox of the plankton,” i.e., why there are so many co-existing plankton species despite a limited range of resources in the water column (Hutchinson 1961, Chesson 2000). However, when studying drivers of bacterioplankton diversity in freshwater lakes, sampling strategies generally do not differentiate between the spatial habitats existing within a lake (Grossart 2010). In addition to spatial lake habitats shaping BCC, large changes in BCC have been reported between aquatic habitats defined by in-line filter pore sizes in marine and freshwater systems (Rösel and Grossart 2012, Rösel et al. 2012, Parveen et al. 2013, Smith et al. 2013, D'Ambrosio et al. 2014, Bižić-Ionescu et al. 2014, Ganesh et al. 2014, 2015, Simon et al. 2014). In this study, we determined habitat preferences of bacterial taxa between free-living (FL; 0.22-3 μm) and particle-associated (PA; 3-20 μm) filter fractions within the epi- and hypolimnion of north-temperate lakes spanning a large nutrient gradient. Our results help integrate previous studies that focused on comparing one of these habitat pairs (Allgaier and Grossart 2006, Shade et al. 2008, Kolmonen et al. 2011, Rösel and Grossart 2012, Rösel et al. 2012, Garcia et al. 2013). The existence of a series of OTUs that are significantly differentially abundant between these spatial habitats indicates taxon-specific habitat preferences that help explain the high numbers of coexisting bacteria in the pelagic zones of freshwater lakes.

For differentially abundant OTUs, we showed that OTU level habitat preferences of abundant OTUs between lake layers, filter fractions, and nutrient levels was highly conserved at the phylum level for a majority of phyla. This is counter to what is predicted based on the polyphyletic distribution of most bacterial traits (Martiny et al. 2012) and the metabolic versatility that occurs even below the OTU level (Hunt et al. 2008a, Hoefman et al. 2014). However, phylum to class level conservation of habitat preference between particle-associated and free-living fractions has been recently reported for multiple phyla in marine bacterioplankton (Salazar et al. 2015). Complex traits tend to be more phylogenetically conserved, suggesting that some traits underpinning habitat partitioning are complex, i.e., encoded by multiple interdependent genes (Martiny et al. 2012). It also has to be noted that most OTUs did not show differential abundance between habitats, either because we lacked the statistical power to show differential abundance (as most of these OTUs were rare for most phyla) or because these OTUs are generalists across habitats as we defined them.

For the large number of phyla differentially abundant in the oxygen-depleted hypolimnia of high-nutrient lakes, the most important shared trait is likely their ability to grow anaerobically, which can be considered a complex trait (Martiny et al. 2012). Many cultured representatives of the phyla enriched in this habitat have indeed been shown to be capable of anaerobic metabolism (Krieg et al. 2012). The differential abundance in low-nutrient hypolimnia of six of the seven most abundant phyla (*Bacteriodetes, Cyanobacteria, Verrucomicrobia, Actinobacteria, Planctomycetes,* and *Alphaproteobacteria*; middle column; **Figure SI 2.11**) could be attributed to the absence of taxa enriched in high-nutrient hypolimnia, thus increasing the relative abundance of these ubiquitous and mostly aerobic freshwater lineages (Newton et al. 2011a).

The similarity of epilimnetic BCC between lakes with varying nutrient levels is consistent with the environmental similarity of surface waters of the high- and low-nutrient lakes we sampled, and has been observed in previous studies (Allgaier and Grossart 2006). The greater dissimilarity between epilimnetic and hypolimnetic communities in high-nutrient lakes (**Figure 2.2; Figure 2.3**) in turn reflects the fact that the hypolimnia of high-nutrient lakes are typically more environmentally distinct from their epilimnia as compared to low-nutrient lakes due to pronounced vertical gradients in dissolved oxygen, pH, and photosynthetically active radiation. The distinct environmental conditions in nutrient-enriched hypolimnia also translated into significantly higher observed richness (**Figure SI 2.8**), which has been suggested to be due to nutrient accumulation in the hypolimnion (Kara et al. 2013) and higher habitat heterogeneity (Shade et al. 2008, Jankowski et al. 2014). The presence or absence of OTUs detected in high-nutrient hypolimnia between lakes was more similar than compared to any other inter-lake comparison of habitats (**Figure 2.3, top**), which may indicate strong species sorting at the local scale due to hypoxia and lower pH and photosynthetically active radiation (Van der Gucht et al. 2007).

The mixed lake offered an interesting opportunity to observe the impact of a press disturbance on a lake ecosystem, i.e., the continuous aeration and de-stratification of lake layers. Previous studies have investigated the impact of short-term water column mixing pulse disturbances on bacterial communities (Shade et al. 2010, 2012). They found that communities were rapidly altered, leading to more similar communities across the water column, increased richness in the surface waters and the appearance of unique OTUs after artificial mixing (Shade et al. 2012). However, the BCC reverted to the pre-disturbance state within 7 (epilimnion) and 11 (hypolimnion) days following the return of stratification (Shade et al. 2012). While we only sampled a single artificially mixed high-nutrient lake, we also found that mixing increased OTU richness relative to high-nutrient lake epilimnia and vertically homogenized the BCC (**Figure SI 2.8**). The mixed lake had higher deep-water temperatures and lower near-surface dissolved oxygen concentrations than the other lakes (**Figure SI 2.6**), which may contribute to the increase in observed richness.

Both technical and biological factors affect the extensive OTU overlap observed between PA and FL fractions in this and previous studies (**Figure 2.1**; Bižić-Ionescu et al. 2014). The inline filtration method used to physically discriminate PA from FL fractions may be only a crude indicator of microbial habitat preferences. This may be due to (1) the filter pore size used affects the results, and (2) PA bacteria can dislodge from particles during sample filtration, and inversely FL bacteria can be trapped on higher pore size filters (Hunt et al. 2008a, Simon et al. 2014). While the pore size to discriminate PA and FL bacteria varies, the 3 µm cutoff is frequently used (Bidle and Fletcher 1995, Crump et al. 1998, Acinas et al. 1999, Besemer et al. 2005, Eloe et al. 2011, Jackson et al. 2014), including in the recent worldwide Tara Oceans survey of marine microbial communities, on which our sampling procedures were based (Pesant et al. 2015, Sunagawa et al. 2015). The OTU overlap can be explained by biological factors as well. The release of hydrolytic enzymes from PA bacteria can help liberate small molecules from particles and in turn, stimulate the growth of their FL counterparts on these newly dissolved nutrients (Long and Azam 2001, Ghiglione et al. 2009). However, it would not enable maintenance of large population sizes for species requiring nutrients only available on particles. The FL fraction may also capture cells migrating between the ephemeral habitats provided by particulate matter (Fraser et al. 2009) or cells that have a generalist behavior and alternate between the FL and PA habitats.

Despite the overlapping presence of many OTUs, a series of OTUs were found to be differentially abundant between PA and FL fractions. One trait determining differential abundance between FL and PA fractions is cell size. For example, ubiquitous freshwater bacteria such as AcI *(Actinobacteria)* and *Polynucleobacter (Betaproteobacteria)* were consistently overrepresented in FL habitats and are both reported ultramicrobacteria that are rarely larger than 1 µm (Pernthaler 2013). Cell size and morphology are dependent on a small number of genes and both traits are phylogenetically dispersed (Margolin 2009). Cell size is also a phenotypically plastic trait (Margolin 2009). Thus, other traits such as the presence of genes enabling extracellular digestion of particulate matter may be more important than size for explaining the relatively high differential abundance between PA and FL fractions. The presence of excreted enzymes such as glycoside hydrolases and proteases have indeed been observed for PA bacteria (Grossart et al. 2006). Excreted enzymes are relatively simple traits and thus are not predicted to be highly phylogenetically conserved. Yet, a recent study has shown specific classes of glycoside hydrolases to be phylogenetically conserved (Amend et al. 2015). *Planctomycetes* likely have larger cell sizes due to the presence of unique intracellular structures (nucleoid, anammoxosome) and budding cell division (Lee et al. 2009, Fuerst and Sagulenko 2011). Filamentous forms of *Cyanobacteria* such as *Anabaena*, *Planktothrix,* and *Pseudoanabaena,* as well as grouped- or paired-cells or microcolonies like *Snowella* and *Synechococcus* (Callieri 2010) most likely would not pass through the 3 μm filter.

Our observation that inter-lake variability is greater for PA bacteria than for FL bacteria, especially within the hypolimnion, (**Figure 2.3, bottom**) suggests that PA bacteria may be important determinants of differences in bacterially mediated ecosystem processes between lakes. While PA bacteria are typically outnumbered by FL bacteria (Caron et al. 1982), in certain systems PA bacteria can be disproportionally active (Azam 1998, Crump et al. 1998, Grossart et al. 2006, Lemarchand et al. 2006, Ghiglione et al. 2007). Further insights into the functional potential of PA bacteria and their activity relative to FL bacteria will be needed to assess their contributions to freshwater biogeochemical cycling. Application of metagenomic analyses, as has been recently applied to PA and FL communities in marine systems (Smith et al. 2013b, Ganesh et al. 2014b, Simon et al. 2014) are a logical next step in freshwater lake systems.

Particle-associated bacterial communities may also be more sensitive to global change stressors (land use change, species invasions, changes in geochemical cycles (Chapin et al. 2000). First, the changes in runoff patterns due to urban and agricultural land development and climate change could alter the nature and amount of suspended solids delivered to freshwater lakes (Adrian et al. 2009, Nõges et al. 2011, Michalak et al. 2013). Second, an invasive dreissenid mussel that has major impacts on aquatic systems in North America and Eastern Europe (Dame and Olenin 2005, Ruesink et al. 2006, Higgins and Zanden 2010) was recently shown to preferentially remove PA bacteria from freshwater systems through size-selective filter feeding (Cotner et al. 1995, Denef et al. 2017). The fact that the specific taxa that are differentially abundant between PA and FL fractions are dependent on lake layer and nutrient level indicates that (1) global change could significantly alter PA communities and (2) the impacts on BCC, and potentially community functioning, could be large and strongly ecosystem-dependent.

In this study, we documented how freshwater lake bacterioplankton communities differ between free-living and particle-associated fractions, lake layers, and nutrient levels. Our results differentiated taxa that are generalists across these habitats from more specialized taxa. OTU-level habitat partitioning for specialized taxa was highly conserved at the phylum level for multiple phyla, indicating that complex multi-gene traits may underpin ecological adaptation to these distinct habitats. Our findings contribute to a growing understanding of how habitat heterogeneity helps sustain high bacterial diversity. Importantly, our data identified taxa to pursue with genomic approaches to help identify traits that underpin habitat specialization and co-occurring functional traits that determine the unique characteristics of bacterially mediated ecosystem processes in these different freshwater habitats.

Methods

Lake Sampling and Sample Processing

Samples were collected on 5-8 August 2013 from 11 lakes across a productivity gradient (7.6 - 47.8 μg/L total phosphorus, TP) in southwestern Michigan (**Table 2.1**). Sampling was conducted over the deepest basin of each lake. A vertical lake profile was taken from the surface to the bottom at 1-2 m intervals for temperature, pH, and dissolved oxygen (mg/L) with a mutli-parameter Hydrolab sonde (Hach Hydromet). The epilimnion and hypolimnion (if present) were identified from the temperature profile and duplicate water samples were collected with a 3.2 L horizontal Van Dorn bottle from the middle of each lake layer, except for Gull Lake which was sampled towards the top of the hypolimnion. Water samples were immediately pre-filtered through a 210 and 20 μm nitex cloth (WildCo.) to remove large phyto- and zooplankton and stored in a cooler until processed in the lab within 6 hours. An additional sample of lake water was collected from the epilimnion without filtration for TP. For our analyses, we only used epilimnetic data for Wintergreen Lake as the thermocline extended to the bottom. At the time of sampling, Sherman Lake was artificially de-stratified ("mixed", **Table 2.1**) with aerators installed by a local watershed association (www.shermanlakemi.com).

Lakes were categorized as "low-nutrient" (TP ≤ 10 µg/L) or "high-nutrient (TP > 10 µg/L) based on total phosphorus (TP). This cutoff corresponds to the depth of the thermocline, the distinction between lakes with and without significant hypolimnetic hypoxia in our dataset (**Figure SI 2.6**), and is similar to the 12 µg/L cutoff proposed by Carlson (1977). Subsamples of unfiltered lake water were poured off after thorough mixing and frozen for later analysis of TP, which was performed using standard colorimetric techniques (molybdenum-blue method) and long pathlength spectrophotometry following persulfate digestion of organic matter in an autoclave (Murphy J and Riley JP 1962, Menzel and Corwin 1965).

Microbial biomass for particle-associated (3-20 µm fraction) and free-living (0.22-3 µm fraction) bacterial communities were collected by sequential in-line filtration of 20 µm pre-filtered lake water samples (350 mL - 2 L) through a 3 µm isopore polycarbonate membrane filter (TSTP, 47 mm diameter, Millipore, Billerica, MA, USA) and a 0.22 µm Express Plus polyethersulfone membrane filter (47 mm diameter, Millipore, Billerica, MA, USA) held in line with a 47 mm polycarbonate in-line filter holder (Pall Corporation, Ann Arbor, MI, USA). Filtration was performed using an E/S portable peristaltic pump with an easy-load L/S peristaltic pump head (Masterflex®, Cole Parmer Instrument Company, Vernon Hills, IL, USA). Filters were then submersed in RNAlater (Ambion) in 2 mL cryovials, frozen in liquid nitrogen and transferred to a -80 ^{o}C freezer until DNA extraction.

DNA Extraction

DNA extractions were performed using an optimized method based on the AllPrep DNA/RNA/miRNA Universal kit (Qiagen; McCarthy et al. 2015). In summary, filters were first washed with phosphate buffered saline (PBS; pH 7.4) while folded (cell-side in) to prevent against cell-loss and to remove RNAlater, which inhibits DNA yields. Then filters were placed in a 2 mL tube with 125 µL of lysozyme (8 mg/L; Sigma) and incubated for 5 minutes at 37 ^{o}C. Next, 600 µL of buffer RLT plus (Qiagen) and 6 µL of β-mercaptoethanol was added to each tube and incubated for 90 minutes at room temperature using a rotisserie motor. After incubation, tubes were vortexed on high for 10 minutes. The lysate was transferred to a QiaShredder column (Qiagen), 300 µL of 100% ethanol was added to the lysate and then transferred to a DNA column (DNeasy Blood and Tissue Kit, Qiagen) and washed with 350 µL of buffer AW1 (Qiagen). Next, 80 µL of proteinase K solution was added to the DNA column and incubated at room temperature for 5 minutes. The DNA column was washed with buffer AW1 and buffer AW2. DNA was eluted using 2 x 30 µL elution buffer (buffer EB, Qiagen) into a fresh 1.5 mL centrifuge tube and stored at 4 ^{o}C until processed for sequencing.

DNA Sequencing and Processing

Extracted DNA was sequenced using Illumina MiSeq v2 chemistry 2x250 (500 cycles) of dual index-labeled primers that targeted the V4 hypervariable region of the 16S rRNA gene (515F/806R; (Caporaso et al. 2012, Kozich et al. 2013) at the University of Michigan Medical School on December 20th, 2013. RTA v1.17.28 and MCS v2.2.0 software were used to generate data. Fastq files were submitted to NCBI sequence read archive under BioProject PRJNA304344, SRA accession number SRP066777. We analyzed the sequence data using mothur v.1.36.1 (Schloss et al. 2009a) based on the MiSeq standard operating procedure accessed on November 3rd, 2015. We used the *cluster.split* command in mothur to assign sequences to OTUs at 97% similarity using the average neighbor algorithm and *classify.seqs* command in mothur using the Wang method implemented in the RDP classifier to assign the taxonomy of the OTUs with a combination of the Silva Database (release 119) and the freshwater 16S rRNA database (Newton et al. 2011; available at https://github.com/mcmahon-uw/FWMFG) for taxonomic classification (Pruesse et al. 2012).

Statistical Analyses

Further analysis of sequence data was performed in R version 3.2.2 (August 2015) using the phyloseq (McMurdie and Holmes 2013) and vegan (Oksanen et al. 2013) R-packages. All figures were made using the ggplot2 R-package (Wickham 2009). All input files and code are available at https://github.com/DenefLab/Final_PAFL_Trophicstate. OTU- and taxonomy tables produced in mothur along with categorical and measured environmental variables were imported into phyloseq. We pruned out all non-bacterial and chloroplast sequences and then merged replicate samples by summing using the *merge_samples* function in phyloseq.

Differences in Bacterial Community Composition

We transformed the sequence read depth of each of the summed replicate samples by taking the proportion of each OTU and scaling it to the minimum sequence read depth in the data set (14,925 sequences), and then rounding to the nearest integer (McMurdie and Holmes 2014). We determined the number of shared OTUs between filter fractions, nutrient levels and lake layers using the *anti_join* and *semi_join* functions in the dplyr R-package (Wickham et al. 2018). We used the *chisq.test* function within the stats R-package (R Core Team 2018) to perform a Chi squared test to test whether there was a significant difference in the number of unique OTUs detected in each environment.

We calculated two non-metric multidimensional scaling (NMDS) ordinations based on (1) Sørensen (unweighted; OTU presence or absence) and (2) Bray-Curtis (abundance-weighted; OTU relative abundance) dissimilarity using the *metaMDS* function (vegan) with 2 dimensions (k = 2), a square root transformation, and Wisconsin double standardization. To test if filter fraction, nutrient level, lake layer, DO, temperature, and pH could significantly explain variation in the bacterial community composition, we used *adonis* (vegan) to run a permutational ANOVA (PERMANOVA; Anderson 2001). We tested our variables nested within each other to account for co-variation. To test for significant differences in the variance of the Sørensen and Bray-Cutis dissimilarities between lakes, a Kruskall-Wallis test was performed using the *kruskal.test* function within the stats package (R Core Team 2018) along with the post-hoc tests using the *kruskalmc* function (pgirmess R-package; Giraudoux 2018). Significant differences were visually added using the *multcompLetters* function within the mutlcompView R-package (Fig 3; (Graves et al. 2015).

Differentially abundant phyla and OTUs

We identified significantly differentially abundant phyla and OTUs between habitats by calculating the $\log_2$-fold ratio using the negative binomial generalized linear model framework of the *DESeq* function in the DESeq2 R-package (Love et al. 2014, McMurdie and Holmes 2014). P-values were adjusted for multiple tests with a Benjamini-Hochberg false discovery rate-correction and a threshold p-value of 0.01 was used to prevent the likelihood of false-positives.

We used the $\log_2$-fold ratio of the relative abundance ($\log_2$ of odds ratio) to create a heatmap representing differentially abundant taxa between PA and FL fractions, high- and low-nutrient lakes, and between hypolimnia and epilimnia (**Figure 2.3, Figure SI 2.11, & Figure SI 2.12**).

To determine how phylogenetically conserved habitat preference of each phylum was at the OTU level, we counted the number of significant OTUs from the *DESeq* output using the *group_by* and *summarize* functions (dplyr R-package; Wickham 2011) and combined this data with the total number of OTUs within each of the habitats. We also calculated the within phylum relative abundance of each of the OTUs in order to detect what proportion of the phylum abundance was made up by the significant OTUs.

Acknowledgements

Work by MLS was supported by the National Science Foundation Graduate Research Fellowship. O. Sarnelle (Michigan State University) and staff at the W.K. Kellogg Biological Station, especially A. Fogiel and T. Brownell, provided tremendous logistical support and lent equipment. The Gull Lake Quality Organization contributed additional funds to support the lake sampling. We are grateful to members of the Denef, Dick, and Duhaime laboratories (University of Michigan) as well as the three anonymous reviewers for help with improving previous drafts of this manuscript.

References

Acinas, S. G., J. Antón, and F. Rodríguez-Valera. 1999. Diversity of Free-Living and Attached Bacteria in Offshore Western Mediterranean Waters as Depicted by Analysis of Genes Encoding 16S rRNA. Applied and Environmental Microbiology 65:514–522.

Adrian, R., C. M. O. Reilly, H. Zagarese, S. B. Baines, D. O. Hessen, W. Keller, D. M. Livingstone, R. Sommaruga, D. Straile, E. Van Donk, G. A. Weyhenmeyer, and M. Winder. 2009. Lakes as sentinels of climate change. Limnology and Oceanography 54:2283–2297.

Allen, L. Z., E. E. Allen, J. H. Badger, J. P. McCrow, I. T. Paulsen, L. D. Elbourne, M. Thiagarajan, D. B. Rusch, K. H. Nealson, S. J. Williamson, J. C. Venter, and A. E. Allen. 2012. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME Journal 6:1403–1414.

Allgaier, M., and H. P. Grossart. 2006. Seasonal dynamics and phylogenetic diversity of free-living and particle-associated bacterial communities in four lakes in northeastern Germany. Aquatic Microbial Ecology 45:115–128.

Ambrosio, L. D., K. Ziervogel, B. Macgregor, A. Teske, and C. Arnosti. 2014. Composition and enzymatic function of particle-associated and free-living bacteria : a coastal / offshore comparison:2167–2179.

Amend, A. S., A. C. Martiny, S. D. Allison, R. Berlemont, M. L. Goulden, Y. Lu, K. K. Treseder, C. Weihe, and J. B. H. Martiny. 2015. Microbial response to simulated global change is phylogenetically conserved and linked with functional potential. The ISME Journal:1–10.

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32–46.

Azam, F. 1998. Microbial Control of Oceanic Carbon Flux: The Plot Thickens. Science 280:694–696.

Besemer, K., M. Moeseneder, J. Arrieta, G. Herndl, and P. Peduzzi. 2005. Complexity of bacterial communities in a river-floodplain system edited by foxit reader. Applied and environmental microbiology 71:609–620.

Bidle, K. D., and M. Fletcher. 1995. Comparison of free-living and particle-associated bacterial communities in the chesapeake bay by stable low-molecular-weight RNA analysis. Applied and Environmental Microbiology 61:944–952.

Bižić-Ionescu, M., M. Zeder, D. Ionescu, S. Orlić, B. M. Fuchs, H.-P. Grossart, and R. Amann. 2014. Comparison of bacterial communities on limnic versus coastal marine particles reveals profound differences in colonization. Environmental microbiology:1–36.

Callieri, C. 2010. Single cells and microcolonies of freshwater picocyanobacteria: A common ecology. Journal of Limnology 69:257–277.

Caporaso, J. G., C. L. Lauber, W. A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer, S. M. Owens, J. Betley, L. Fraser, M. Bauer, N. Gormley, J. A. Gilbert, G. Smith, and R. Knight. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6:1621–1624.

Carlson, R. E. 1977. A trophic state index for lakes. Limnology and Oceanography 22:361–369.

Caron, D. A., P. G. Davis, L. P. Madin, and J. M. Sieburth. 1982. Heterotrophic bacteria and bacterivorous protozoa in oceanic macroaggregates. Science 218:795–797.

Chapin, F. S., E. S. Zavaleta, V. T. Eviner, R. L. Naylor, P. M. Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E. Hobbie, M. C. Mack, and S. Díaz. 2000. Consequences of changing biodiversity. Nature 405:234–42.

Chesson, P. L. 2000. Mechanisms of maintenance of species diversity. Annual Review of Ecological Systematics 31:343–366.

Cotner, J. B., W. S. Gardner, J. R. Johnson, R. H. Sada, J. F. Cavaletto, and R. T. Heath. 1995. Effects of Zebra Mussels (Dreissena polymorpha) on Bacterioplankton: Evidence for Both Size-Selective Consumption and Growth Stimulation. Journal of Great Lakes Research 21:517–528.

Crump, B. C., J. A. Baross, and C. A. Simenstad. 1998. Dominance of particle-attached bacteria in the Columbia River estuary, USA. Aquatic Microbial Ecology 14:7–18.

Dame, R. F., and S. Olenin. 2005. The comparative roles of suspension-feeders in ecosystems: proceedings of the NATO advanced research workshop on the comparative roles of suspension-feeders in ecosystems. Springer Science and Business Media, Nida, Lithuania.

Denef, V. J., H. J. Carrick, J. Cavaletto, E. Chiang, T. H. Johengen, and H. A. Vanderploeg. 2017. Lake Bacterial Assemblage Composition Is Sensitive to Biological Disturbance Caused by an Invasive Filter Feeder. mSphere 2:e00189-17.

Eloe, E. A., C. N. Shulse, D. W. Fadrosh, S. J. Williamson, E. E. Allen, and D. H. Bartlett. 2011. Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. Environmental Microbiology Reports 3:449–458.

Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt, W. P. Hanage, and T. Bacteria. 2009. The Bacterial Species Challenge : Ecological Diversity:741–746.

Fuerst, J. a, and E. Sagulenko. 2011. Beyond the bacterium: planctomycetes challenge our concepts of microbial structure and function. Nature Reviews Microbiology 9:403–413.

Ganesh, S., L. a Bristow, M. Larsen, N. Sarode, B. Thamdrup, and F. J. Stewart. 2015. Size-fraction partitioning of community gene transcription and nitrogen metabolism in a marine oxygen minimum zone. The ISME Journal 9:2682–2696.

Ganesh, S., D. J. Parris, E. F. DeLong, and F. J. Stewart. 2014. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. The ISME journal 8:187–211.

Garcia, S. L., I. Salka, H. P. Grossart, and F. Warnecke. 2013. Depth-discrete profiles of bacterial communities reveal pronounced spatio-temporal dynamics related to lake stratification. Environmental Microbiology Reports 5:549–555.

Ghiglione, J. F., P. Conan, and M. Pujo-Pay. 2009. Diversity of total and active free-living vs. particle-attached bacteria in the euphotic zone of the NW Mediterranean Sea. FEMS Microbiology Letters 299:9–21.

Ghiglione, J. F., G. Mevel, M. Pujo-Pay, L. Mousseau, P. Lebaron, and M. Goutx. 2007. Diel and seasonal variations in abundance, activity, and community structure of particle-attached and free-living bacteria in NW Mediterranean Sea. Microbial Ecology 54:217–231.

Giraudoux, P. 2018. pgirmess: Spatial Analysis and Data Mining for Field Ecologists.

Graves, S., H.-P. Piepho, and L. S. with help from Sundar Dorai-Raj. 2015. multcompView: Visualizations of Paired Comparisons.

Grossart, H. P. 2010. Ecological consequences of bacterioplankton lifestyles: Changes in concepts are needed. Environmental Microbiology Reports 2:706–714.

Grossart, H. P., T. Kiorboe, K. W. Tang, M. Allgaier, E. M. Yam, and H. Ploug. 2006. Interactions between marine snow and heterotrophic bacteria: aggregate formation and microbial dynamics. Aquatic Microbial Ecology 42:19–26.

Van der Gucht, K., K. Cottenie, K. Muylaert, N. Vloemans, S. Cousin, S. Declerck, E. Jeppesen, J.-M. Conde-Porcuna, K. Schwenk, G. Zwart, H. Degans, W. Vyverman, and L. De Meester. 2007. The power of species sorting: local factors drive bacterial community composition over a wide range of spatial scales. Proceedings of the National Academy of Sciences of the United States of America 104:20404–20409.

Higgins, S., and M. Zanden. 2010. What a difference a species makes: a meta-analysis of dreissenid mussel impacts on freshwater ecosystems. Ecological monographs 80:179–196.

Hoefman, S., D. van der Ha, N. Boon, P. Vandamme, P. De Vos, and K. Heylen. 2014. Niche differentiation in nitrogen metabolism among methanotrophs within an operational taxonomic unit. BMC microbiology 14:83.

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320:1081–1085.

Hutchinson, G. 1961. The paradox of the plankton. American Naturalist 95:137–145.

Jackson, C. R., J. J. Millar, J. T. Payne, and C. A. Ochs. 2014. Free-Living and Particle-Associated Bacterioplankton in Large Rivers of the Mississippi River Basin Demonstrate Biogeographic Patterns. Applied and Environmental Microbiology 80:7186–7195.

Jankowski, K., D. E. Schindler, and M. C. Horner-Devine. 2014. Resource availability and spatial heterogeneity control bacterial community response to nutrient enrichment in lakes. PLoS ONE 9.

Jones, S. E., T. a. Cadkin, R. J. Newton, and K. D. McMahon. 2012. Spatial and temporal scales of aquatic bacterial beta diversity. Frontiers in Microbiology 3:1–10.

Kara, E. L., P. C. Hanson, Y. H. Hu, L. Winslow, and K. D. McMahon. 2013. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. The ISME journal 7:680–684.

Köllner, K., D. Carstens, C. Schubert, J. Zeyer, and H. Bürgmann. 2013. Impact of particulate organic matter composition and degradation state on the vertical structure of particle-associated and planktonic lacustrine bacteria. Aquatic Microbial Ecology 69:81–92.

Kolmonen, E., K. Haukka, A. Rantala-Ylinen, P. Rajaniemi-Wacklin, L. Lepistö, and K. Sivonen. 2011. Bacterioplankton community composition in 67 Finnish lakes differs according to trophic status. Aquatic Microbial Ecology 62:241–250.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Applied and Environmental Microbiology 79:5112–5120.

Krieg, N. R., J. T. Staley, D. R. Brown, B. P. Hedlund, B. J. Paster, N. L. Ward, W. Ludwig, and W. B. Whitman. 2012. Bergey's Manual of Systematic Bacteriology. Springer, New York, NY.

Lee, K.-C., R. I. Webb, and J. a Fuerst. 2009. The cell cycle of the planctomycete Gemmata obscuriglobus with respect to cell compartmentalization. BMC cell biology 10.

Lemarchand, C., L. Jardillier, J. F. Carrias, M. Richardot, D. Debroas, T. Sime-Ngando, and C. Amblard. 2006. Community composition and activity of prokaryotes associated to detrital particles in two contrasting lake ecosystems. FEMS Microbiology Ecology 57:442–451.

Lindström, E. 2000. Bacterioplankton Community Composition in Five Lakes Differing in Trophic Status and Humic Content. Microbial ecology 40:104–113.

Long, R. A., and F. Azam. 2001. Antagonistic interactions among marine pelagic bacteria. Applied and Environmental Microbiology 67:4975–4983.

Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:1–21.

Margolin, W. 2009. Sculpting the Bacterial Cell. Current Biology 19:R812–R822.

Martiny, A. C., K. K. Treseder, and G. Pusch. 2012. Phylogenetic conservatism of functional traits in microorganisms. The ISME journal 7:830–838.

McCarthy, A., E. Chiang, M. L. Schmidt, and V. J. Denef. 2015. RNA Preservation Agents and Nucleic Acid Extraction Method Bias Perceived Bacterial Community Composition. Plos One 10:e0121659.

McMurdie, P. J., and S. Holmes. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217.

McMurdie, P. J., and S. Holmes. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Computational Biology 10:e1003531.

Menzel, D. W., and N. Corwin. 1965. The measurement of total phosphorus in seawater based on liberation of organically bound fractions by persulfate oxidation. Limnology & Oceanography 10:280–282.

Michalak, A. M., E. J. Anderson, D. Beletsky, S. Boland, N. S. Bosch, T. B. Bridgeman, J. D. Chaffin, K. Cho, R. Confesor, I. Daloglu, J. V. DePinto, M. A. Evans, G. L. Fahnenstiel, L. He, J. C. Ho, L. Jenkins, T. H. Johengen, K. C. Kuo, E. LaPorte, X. Liu, M. R. McWilliams, M. R. Moore, D. J. Posselt, R. P. Richards, D. Scavia, A. L. Steiner, E. Verhamme, D. M. Wright, and M. A. Zagorski. 2013. Record-setting algal bloom in Lake Erie caused by agricultural and meteorological trends consistent with expected future conditions. Proceedings of the National Academy of Sciences 110:6448–6452.

Mohit, V., P. Archambault, N. Toupoint, and C. Lovejoy. 2014. Phylogenetic differences in attached and free-living bacterial communities in a temperate coastal lagoon during summer, revealed via high-throughput 16S rRNA gene sequencing. Applied and Environmental Microbiology 80:2071–2083.

Murphy J, and Riley JP. 1962. A modified single solution method for the determination of phosphate in natural waters. Analytica Chimica Acta 27:31–36.

Newton, R. J., S. E. Jones, A. Eiler, K. D. McMahon, and S. Bertilsson. 2011. A guide to the natural history of freshwater lake bacteria. Page Microbiology and molecular biology reviews.

Nõges, P., T. Nõges, M. Ghiani, F. Sena, R. Fresner, M. Friedl, and J. Mildner. 2011. Increased nutrient loading and rapid changes in phytoplankton expected with climate change in stratified South European lakes: Sensitivity of lakes with different trophic state and catchment properties. Hydrobiologia 667:255–270.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. 2013. vegan: Community Ecology Package.

Paganin, P., L. Chiarini, A. Bevivino, C. Dalmastri, A. Farcomeni, G. Izzo, A. Signorini, C. Varrone, and S. Tabacchioni. 2013. Vertical distribution of bacterioplankton in Lake Averno in relation to water chemistry. FEMS Microbiology Ecology 84:176–188.

Parveen, B., I. Mary, A. Vellet, V. Ravet, and D. Debroas. 2013. Temporal dynamics and phylogenetic diversity of free-living and particle-associated Verrucomicrobia communities in relation to environmental variables in a mesotrophic lake. FEMS Microbiology Ecology 83:189–201.

Pernthaler, J. 2013. Freshwater Microbial Communities. Pages 97–112 *in* E. Rosenberg, E. F. DeLong, S. Lory, E. Stackebrandt, and F. Thompson, editors. The Prokaryotes - Prokaryotic Communities and Ecophysiology. Berlin, Germany.

Pesant, S., F. Not, M. Picheral, S. Kandels-Lewis, N. Le Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, C. Dimier, S. Searson, S. G. Acinas, P. Bork, E. Boss, C. Bowler, C. De Vargas, M. Follows, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, S. Kandels-Lewis, L. Karp-Boss, E. Karsenti, U. Krzic, F. Not, H. Ogata, S. Pesant, J. Raes, E. G. Reynaud, C. Sardet, M. Sieracki, S. Speich, L. Stemmann, M. B. Sullivan, S. Sunagawa, D. Velayoudon, J. Weissenbach, and P. Wincker. 2015. Open science resources for the discovery and analysis of Tara Oceans data. Scientific Data 2:150023.

Pruesse, E., J. Peplies, and F. O. Glöckner. 2012. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Rösel, S., M. Allgaier, and H.-P. Grossart. 2012. Long-term characterization of free-living and particle-associated bacterial communities in Lake Tiefwaren reveals distinct seasonal patterns. Microbial ecology 64:571–583.

Rösel, S., and H. P. Grossart. 2012. Contrasting dynamics in activity and community composition of free-living and particle-associated bacteria in spring. Aquatic Microbial Ecology 66:169–181.

Ruesink, J. L., B. E. Feist, C. J. Harvey, J. S. Hong, A. C. Trimble, and L. M. Wisehart. 2006. Changes in productivity associated with four introduced species: Ecosystem transformation of a "pristine" estuary. Marine Ecology Progress Series 311:203–215.

Salazar, G., F. M. Cornejo-Castillo, E. Borrull, C. Díez, E. Lara, D. Vaqué, J. M. Gasol, and S. G. Acinas. 2015. Particle-association lifestyle is a phylogenetically conserved trait in bathypelagic prokaryotes. Molecular Ecology:doi: 10.1111/mec.13419.

Salcher, M. M. 2014. Same same but different: Ecological niche partitioning of planktonic freshwater prokaryotes. Journal of Limnology 73:74–87.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. a. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75:7537–7541.

Shade, A., C. Y. Chiu, and K. D. McMahon. 2010. Seasonal and episodic lake mixing stimulate differential planktonic bacterial dynamics. Microbial Ecology 59:546–554.

Shade, A., S. E. Jones, and K. D. McMahon. 2008. The influence of habitat heterogeneity on freshwater bacterial community composition and dynamics. Environmental Microbiology 10:1057–1067.

Shade, A., J. S. Read, N. D. Youngblut, N. Fierer, R. Knight, T. K. Kratz, N. R. Lottig, E. E. Roden, E. H. Stanley, J. Stombaugh, R. J. Whitaker, C. H. Wu, and K. D. McMahon. 2012. Lake microbial communities are resilient after a whole-ecosystem disturbance. The ISME journal 6:2153–2167.

Simon, H. M., M. W. Smith, and L. Herfort. 2014. Metagenomic insights into particles and their associated microbiota in a coastal margin ecosystem. Frontiers in Microbiology 5:466.

Smith, M. W., L. Z. Allen, A. E. Allen, L. Herfort, and H. M. Simon. 2013a. Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. Frontiers in Microbiology 4:1–20.

Smith, M. W., L. Zeigler Allen, A. E. Allen, L. Herfort, and H. M. Simon. 2013b. Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. Frontiers in microbiology 4:120.

Sunagawa, S., L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. D'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, C. Bowler, C. de

Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, E. Boss, C. Bowler, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. Sieracki, and D. Velayoudon. 2015. Structure and function of the global ocean microbiome. Science 348:1261359.

Wetzel, R. G. 2001. Limnology: Lake and River Ecosystems. Page Journal of Phycology.

Wickham, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York.

Wickham, H. 2011. The Split-Apply-Combine Strategy for Data Analysis. Journal of Statistical Software 40:1–29.

Wickham, H., R. François, L. Henry, and K. Müller. 2018. dplyr: A Grammar of Data Manipulation.

Yannarell, a C., a D. Kent, G. H. Lauster, T. K. Kratz, and E. W. Triplett. 2003. Temporal patterns in bacterial communities in three temperate lakes of different trophic status. Microbial ecology 46:391–405.

Table 2.1. Limnological data for the lakes and samples included in this study.

LAKE NAME	TP (µG/L)	NUTRIENT LEVEL	MIXING STATUS	TOTAL DEPTH (M)	GPS COORDINATES NORTH	GPS COORDINATES WEST	EPILIMNION				HYPOLIMNION			
							Sample Depth (m)	Temp (Celsius)	Dissolved Oxygen (mg/L)	pH	Sample Depth (m)	Temp (Celsius)	Dissolved Oxygen (mg/L)	pH
GULL	7.6	Low	Stratified	31.1	42.39651	85.40936	4	22.92	10.45	8.25	17	9.11	10.21	7.75
LITTLE LONG	8.0	Low	Stratified	7.8	42.41803	85.44348	3	23.37	10.97	8.31	7	19.77	7.99	7.4
SIXTEEN	8.8	Low	Stratified	23.6	42.56518	85.61352	3	23.21	12.00	8.28	16	7.265	5.15	7.37
LEE	9.0	Low	Stratified	15.2	42.17991	85.11844	3	23.38	10.24	8.38	11	9.23	2.41	7.29
PAYNE	11.2	High	Stratified	13.7	42.63749	85.52143	2	23.27	10.97	8.3	11	7.76	0.82	7.05
SHERMAN	13.6	High	Mixed	11.3	42.35212	85.38545	2	24.17	8.48	8.1	9	23.83	7.37	7.78
BRISTOL	13.8	High	Stratified	14.0	42.48737	85.24799	2	22.60	10.03	8.2	11	8.14	0.34	7.2
BASSETT	19.9	High	Stratified	10.2	42.66509	85.48509	1	21.90	14.28	8.63	8	7.62	0.41	7.3
BAKER	28.0	High	Stratified	7.5	42.64643	85.50279	1	22.53	9.71	8.25	5.5	9.21	0.63	7.21 - 7.19
BASELINE	36.1	High	Stratified	14.3	42.42421	85.85677	2	22.88	9.48	8.29	11	9.38	0.034	6.85
WINTERGREEN	47.8	High	Stratified	6.5	42.39757	85.38536	2	22.79	8.58	8.01	5	18.01	0.38	7.27

Table 2.2. R^2 values and P-values from PERMANOVA on the entire and sub-setted data sets.

The mixed lake is omitted from all analyses. BC = Bray-Curtis dissimilarity; S = Sørenson dissimilarity. * = p < 0.05; ** = p < 0.01; *** = p < 0.001, NS = not significant.

Nested PERMANOVA	**All Samples (n = 37)**		**Particle-Associated (n = 18)**		**Free-Living (n = 19)**		**Epilimnion (n = 19)**		**Hypolimnion (n = 18)**		**High- Nutrient (n = 22)**		**Low -Nutrient (n = 15)**	
Individual	**BC**	**S**	**BC**	**S**	**BC**	**S**	**BC**	**S**	**BC**	**S**	**BC**	**S**	**BC**	**S**
Lake Layer	0.158 (0.001 ***)	0.133 (0.001 ***)	0.181 (0.002 **)	0.151 (0.002 ***)	0.234 (0.001 ***)	0.162 (0.001 ***)					0.333 (0.001 ***)	0.301 (0.001 ***)	0.189 (0.001 ***)	0.147 (0.001 ***)
Particle-Association	0.127 (0.001 ***)	0.045 (0.013 *)					0.182 (0.001 ***)	0.065 (0.037 *)	0.168 (0.001 ***)	0.074 (0.031 *)	0.138 (0.001 ***)	0.053 (**NS** (0.072)	0.172 (0.002 **)	0.082 (0.040 *)
Nutrient Level	0.094 (0.001 ***)	0.114 (0.001 ***)	0.129 (0.007 **)	0.148 (0.005 **)	0.110 (0.006 **)	0.122 (0.001 ***)	0.099 (0.013 *)	0.112 (0.001 ***)	0.320 (0.001 ***)	0.324 (0.001 ***)				
Dissolved Oxygen	0.039 (0.010 **)	0.054 (0.003 **)	0.058 (**NS** 0.227)	0.067 (**NS** 0.102)	0.057 (**NS** 0.155)	0.074 (0.046 *)	0.072 (**NS** 0.060)	0.082 (0.007 **)	0.065 (0.027 *)	0.076 (0.022 *)	0.049 (0.043 *)	0.063 (0.027 *)	0.083 (0.048 *)	0.096 (0.010 **)
Temperature	0.031 (**NS** 0.056)	0.027 (**NS** 0.115)	0.042 (**NS** 0.592)	0.042 (**NS** 0.589)	0.043 (**NS** 0.375)	0.034 (**NS** 0.825)	0.042 (**NS** 0.380)	0.064 (**NS** 0.054)	0.072 (0.008 **)	0.064 (0.048 *)	0.035 (**NS** 0.168)	0.044 (**NS** 0.151)	0.088 (0.027 *)	0.085 (0.032 *)
pH	0.026 (**NS** 0.110)	0.027 (**NS** 0.115)	0.036 (**NS** 0.654)	0.039 (**NS** 0.604)	0.037 (**NS** 0.498)	0.038 (**NS** 0.558)	0.089 (0.024 *)	0.078 (0.006 **)	0.049 (**NS** 0.0813)	0.061 (**NS** 0.056)	0.042 (**NS** 0.109)	0.039 (**NS** 0.183)	0.075 (**NS** 0.063)	0.099 (0.007 **)
Residuals	*0.524*	*0.601*	*0.555*	*0.553*	*0.520*	*0.569*	*0.517*	*0.599*	*0.326*	*0.401*	*0.403*	*0.500*	*0.393*	*0.490*

Table 2.3. Top 5 most differentially abundant OTUs within each habitat, sorted by rank.

SEE NEXT PAGE

PHYLUM	GENUS	OTU	LOG2-FOLD RATIO	HABITAT PREFERENCE (SPECIFIC HABITAT WHERE OTU IS DIFFERENTIALLY ABUNDANT)
FREE-LIVING				
VERRUCOMICROBIA	unclassified	Otu00303	5.4	FL (Epilimnion Low-Nutrient)
ACTINOBACTERIA	acIV-A	Otu00094	4.4	FL (Epilimnion Low-Nutrient)
ACTINOBACTERIA	acTH1-A	Otu00046	3.9	FL (Epilimnion Low-Nutrient)
BETAPROTEOBACTERIA	Pnec	Otu00013	3.9	FL (Epilimnion Low-Nutrient)
BETAPROTEOBACTERIA	betIII-A	Otu00189	3.8	FL (Epilimnion Low-Nutrient)
PARTICLE-ASSOCIATED				
CHLOROFLEXI	Anaerolinea	Otu00454	2.3	PA (Epilimnion Low-Nutrient)
CHLOROFLEXI	unclassified	Otu00551	2.5	PA (Epilimnion Low-Nutrient)
UNCLASSIFIED	unclassified	Otu00375	2.5	PA (Epilimnion Low-Nutrient)
ACTINOBACTERIA	Myco	Otu00313	2.6	PA (Epilimnion High-Nutrient)
CHLOROFLEXI	unclassified	Otu00114	2.6	PA (Epilimnion Low-Nutrient)
LOW-NUTRIENT				
BACTEROIDETES	unclassified	Otu00188	10.6	Low-Nutrient (FL Epilimnion)
BACTEROIDETES	unclassified	Otu00185	9.4	Low-Nutrient (FL Epilimnion)
DELTAPROTEOBACTERIA	OM27_clade	Otu00187	9.2	Low-Nutrient (PA Hypolimnion)
UNCLASSIFIED	unclassified	Otu00382	9.1	Low-Nutrient (FL Epilimnion)
CYANOBACTERIA	unclassified	Otu00270	9.0	Low-Nutrient (PA Epilimnion)
HIGH-NUTRIENT				
BETAPROTEOBACTERIA	betI-A	Otu00004	2.5	High-Nutrient (PA Hypolimnion)
VERRUCOMICROBIA	unclassified	Otu00005	3.0	High-Nutrient (PA Hypolimnion)
BACTEROIDETES	unclassified	Otu00118	3.5	High-Nutrient (FL Hypolimnion)
BETAPROTEOBACTERIA	Dechloromonas	Otu00123	3.9	High-Nutrient (PA Hypolimnion)
BACTEROIDETES	unclassified	Otu00120	3.9	High-Nutrient (PA Hypolimnion)
EPILIMNION				
BACTEROIDETES	unclassified	Otu00188	10.0	Epilimnion (Low-Nutrient FL)
BACTEROIDETES	unclassified	Otu00492	8.6	Epilimnion (Low-Nutrient FL)
VERRUCOMICROBIA	Candidatus Xiphinematobacter	Otu00521	8.3	Epilimnion (Low-Nutrient FL)
BACTEROIDETES	unclassified	Otu00334	8.2	Epilimnion (Low-Nutrient FL)
BETAPROTEOBACTERIA	unclassified	Otu00087	8.1	Epilimnion (Low-Nutrient PA)
HYPOLIMNION				

BETAPROTEOBACTERIA	betI-A	Otu00004	2.1	Hypolimnion (High-Nutrient FL)
ACTINOBACTERIA	acI-C	Otu00239	3.1	Hypolimnion (High-Nutrient FL)
CHLOROFLEXI	unclassified	Otu00097	3.1	Hypolimnion (High-Nutrient FL)
NPL-UPA2	unclassified	Otu00843	3.4	Hypolimnion (High-Nutrient FL)
BETAPROTEOBACTERIA	betI-A	Otu00282	3.7	Hypolimnion (High-Nutrient PA)

Figure 2.1. Unique and shared OTUs between lake habitats

(Left) Cumulative number of unique and shared OTUs detected across all lake habitats after transformation of the sequence depth to 14,925 sequences based on the scaling method mentioned in the methods. **(Right)** The relative abundance of the unique and shared OTUs based on their abundance within each habitat. *** Indicates a significant difference in the proportion of unique OTUs based on a Chi-squared test ($p < 0.0001$). Numbers in parentheses along the x-axis represent sample sizes of each lake habitat.

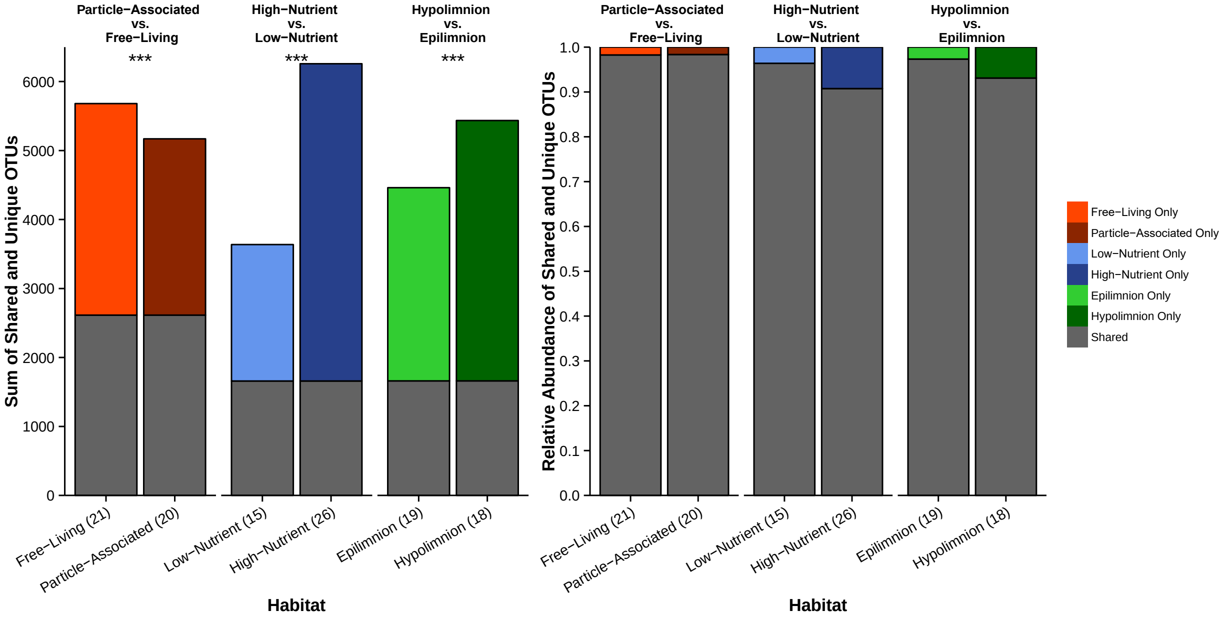


Figure 2.2. NMDS Ordinations of lake habitat influence on bacterial community composition.

Non-Metric Multidimensional Scaling (NMDS) Ordinations visualizing differences in bacterial community composition based on **(A)** Sørensen dissimilarity (OTU presence or absence) and **(B)** Bray-Curtis dissimilarity (OTU relative abundance). Data points are colored by filter fraction, shaped by lake layer, and filled in by the nutrient level of each sample.

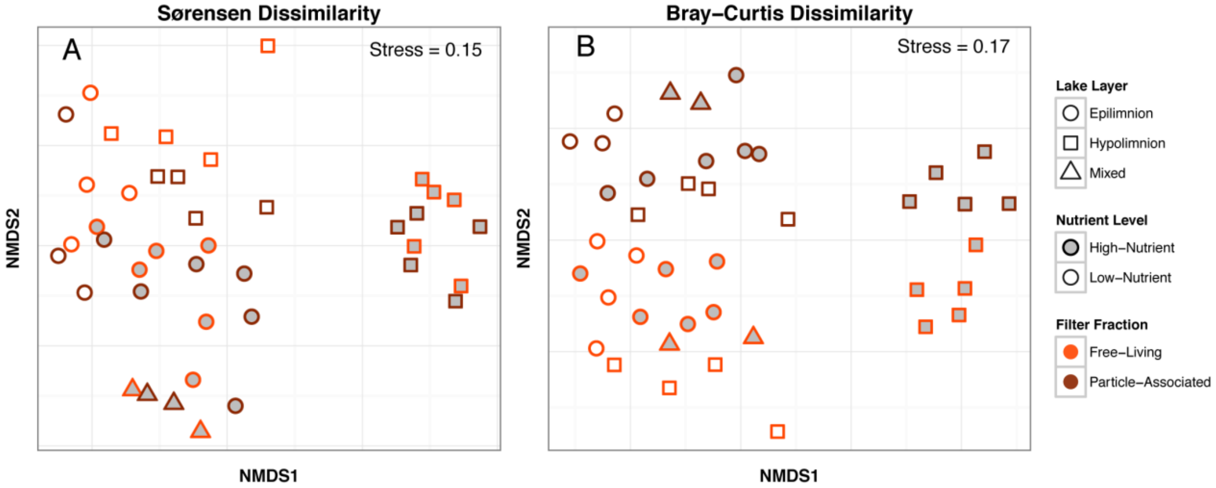


Figure 2.3. Box and whisker plots of the variance in the Sørensen and Bray-Curtis dissimilarity metrics between lakes.

(Top) Sørensen dissimilarity (OTU presence or absence; KW: p = 2.4 x 10^{-7}) and **(Bottom)** Bray-Curtis dissimilarity (OTU relative abundance; KW: p = 3.5 x 10^{-5}) between samples within the same habitat in different lakes. Letter(s) next to data points indicate groups of samples that are significantly different in their degree of lake-to-lake dissimilarity. Numbers in parentheses along the x-axis represent sample sizes of each lake habitat.

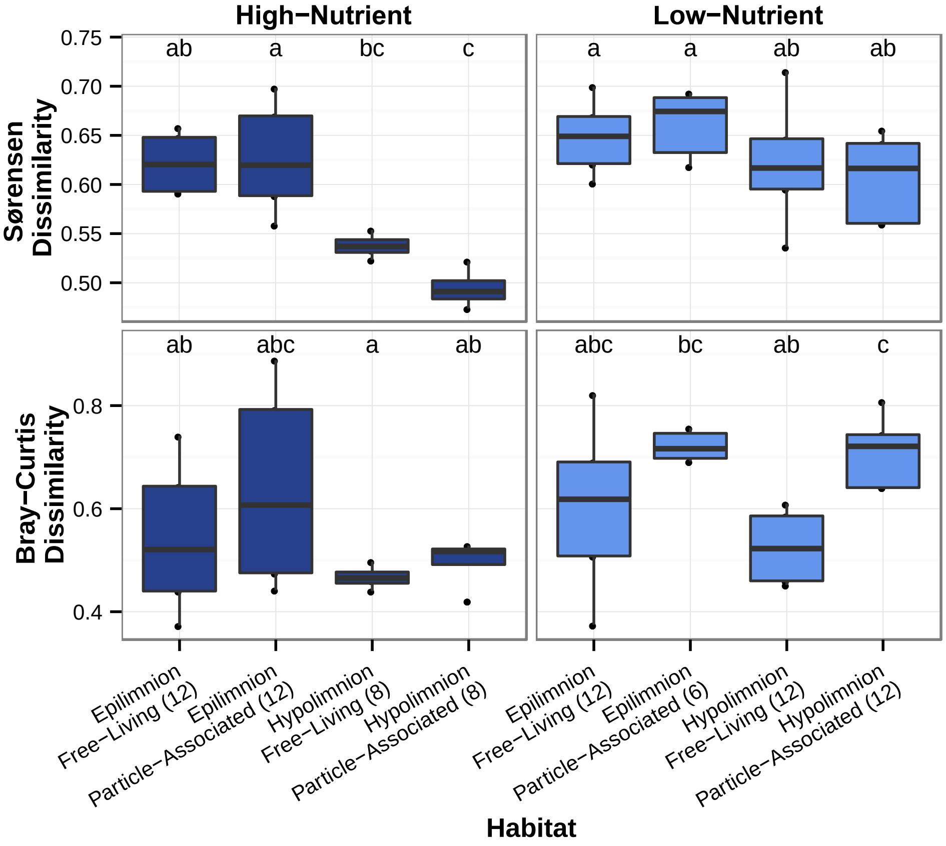


Figure 2.4. Phylum abundance of the six lake habitats of the 17 most abundant phyla and classes of *Proteobacteria*.

Box and whisker plots of the phylum abundance across all samples grouped based on **(top)** filter fraction, **(middle)** nutrient level, or **(bottom)** lake layer. Numbers in parentheses within the legend represent sample sizes of each lake habitat. Red stars represent significant differentially abundant phyla between filter fraction, productivity level, or lake layers as calculated by DESeq.

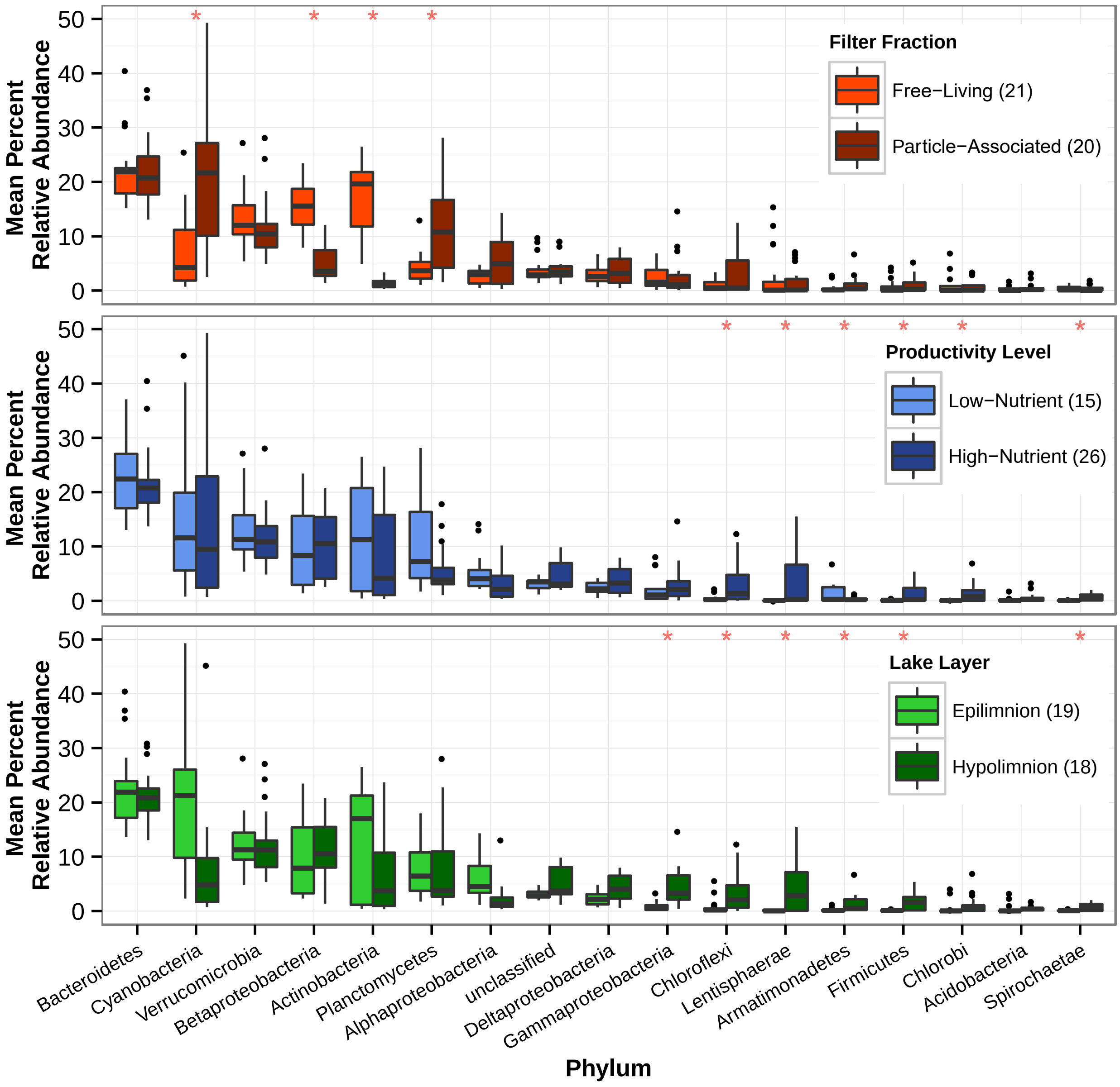
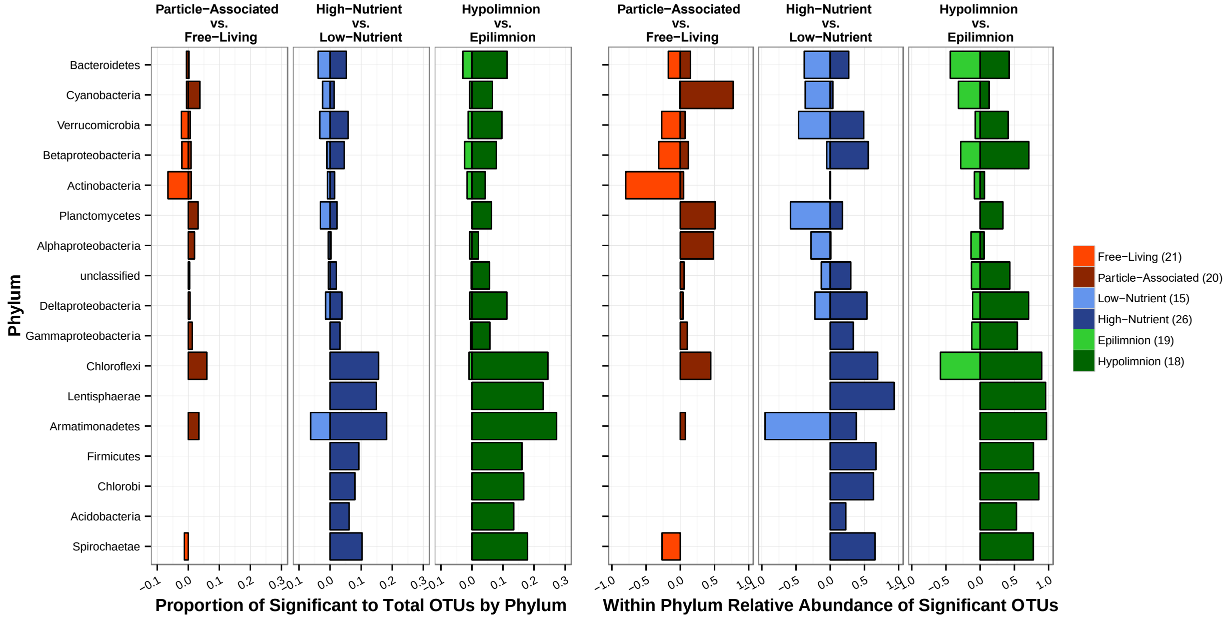


Figure 2.5. Significant differentially abundant OTUs in the 17 most abundant phyla and classes of *Proteobacteria*.

(Left) Proportion of the total number of OTUs that were significantly differentially abundant divided by the total number of OTUs present within each phylum and habitat. **(Right)** Within-phylum relative abundance of the significant OTUs within each phylum and habitat. Numbers in parentheses within the legend represent sample sizes of each lake habitat.



Supporting Information 2.A. Measuring Within Sample Diversity

Experimental Procedures

Within-sample diversity

We measured the within-sample diversity in two ways. First, each sample was randomly sub-sampled to 14,924 sequences without replacement 100 times using the *rarefy_even_depth* function in phyloseq. Second, the sequencing read counts were transformed by taking the proportion of each OTU in the samples and scaling it to the minimum sequence depth in the data set (14,925 sequences), and then rounding to the nearest integer (McMurdie & Holmes, 2014). As the sequencing read counts were transformed by scaling for all the between-sample diversity comparisons, a within-sample diversity figure that was directly comparable to the rest of the results was created. As the total range of sequencing reads after scaling was 521 sequences (Figure SI.2.2), I thought it was appropriate for this data, however, caution should be taken when applying this method to other datasets as scaling is not the standard method for calculating within-sample diversity metrics.

The inverse Simpson index and observed OTU richness were calculated using the *estimate_richness* function in phyloseq. To remove the impact of richness, the inverse Simpson index was divided by the observed richness to obtain Simpson's measure of evenness (Magurran, 2004), which is not sensitive to species richness and ranges from 0 (uneven) to 1 (even).

Results

Within Sample Diversity

For the sub-sampled data the mean Inverse Simpson value ranged from 8.62 to 81.0 (Fig. S3, top-left panel), Simpson's measure of evenness ranged from 0.0135 to 0.0917 (Fig. S3, middle-left panel), observed bacterial richness ranged from 276 to 1,219 OTUs and (Fig. S3, bottom-left panel. For the scaled data (range was 521 sequencing reads; Fig. S2), the mean Inverse Simpson value ranged from 8.44 to 81.2 (Fig. S3, top-right panel), Simpson's measure of evenness ranged from 0.0124 to 0.0918 (Fig. S3, middle-right panel), and observed bacterial richness ranged from 2746 to 1,377 OTUs (Fig. S3, bottom-right panel). In both types of sample transformations the community evenness as measured by the Inverse Simpson reflected a very similar pattern as the observed richness where the values were greatest in the hypolimnia of high-nutrient lakes (both PA and FL communities). Therefore, the Simpson measure of evenness, which showed a pattern not skewed by richness, was also calculated. Based on Simpson's evenness the high-nutrient, particle-associated epilimnion samples were the most uneven communities and the low-nutrient particle-associated samples were the most variable between lakes. The particle-associated and free-living communities of the mixed lake samples did not differ from each other in any diversity metric.

Figure SI. 2.6. Full lake profiles for each lake.

(Top) High-nutrient and **(Bottom)** low-nutrient lake profiles. Colored points represent the raw value at each depth. Lines connect the discrete points within each lake. Sherman Lake was the mixed lake.

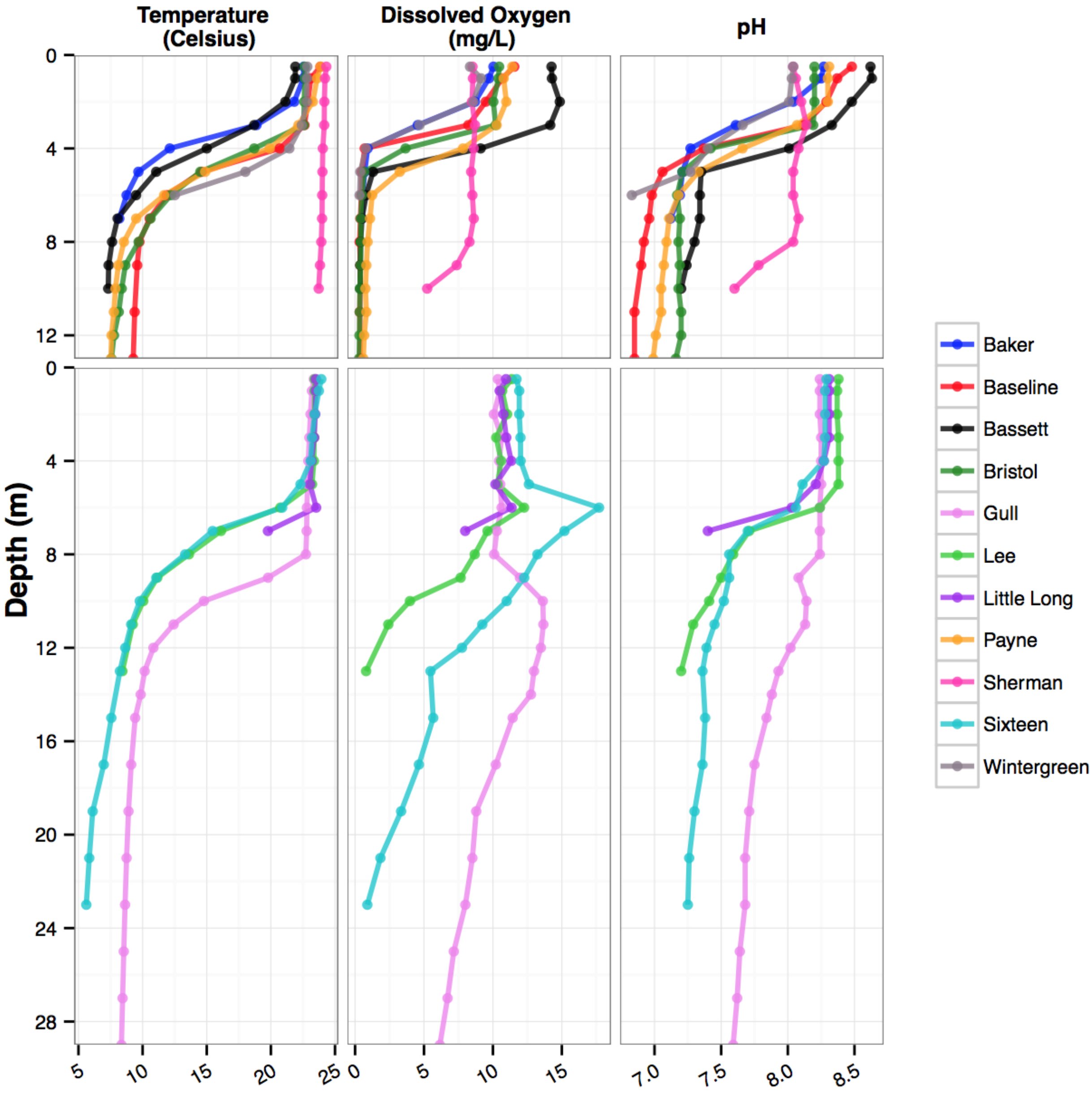


Figure SI. 2.7. Histogram of the total number of sequencing reads per sample.

(Left) Sequencing reads per sample after summing the raw number of sequencing reads between replicate samples. **(Right)** Sequencing reads per sample after transforming the sequence depth of each of the summed replicate samples by taking the proportion of each OTU in samples and scaling it to the minimum sequence depth in the data set (14,925 sequences), and then rounding to the nearest integer. Dotted vertical lines represent the mean sequencing read depth of the free-living (orange) and particle-associated samples. The bin-width represents the number of sequencing reads that are binned within each bar within the histogram

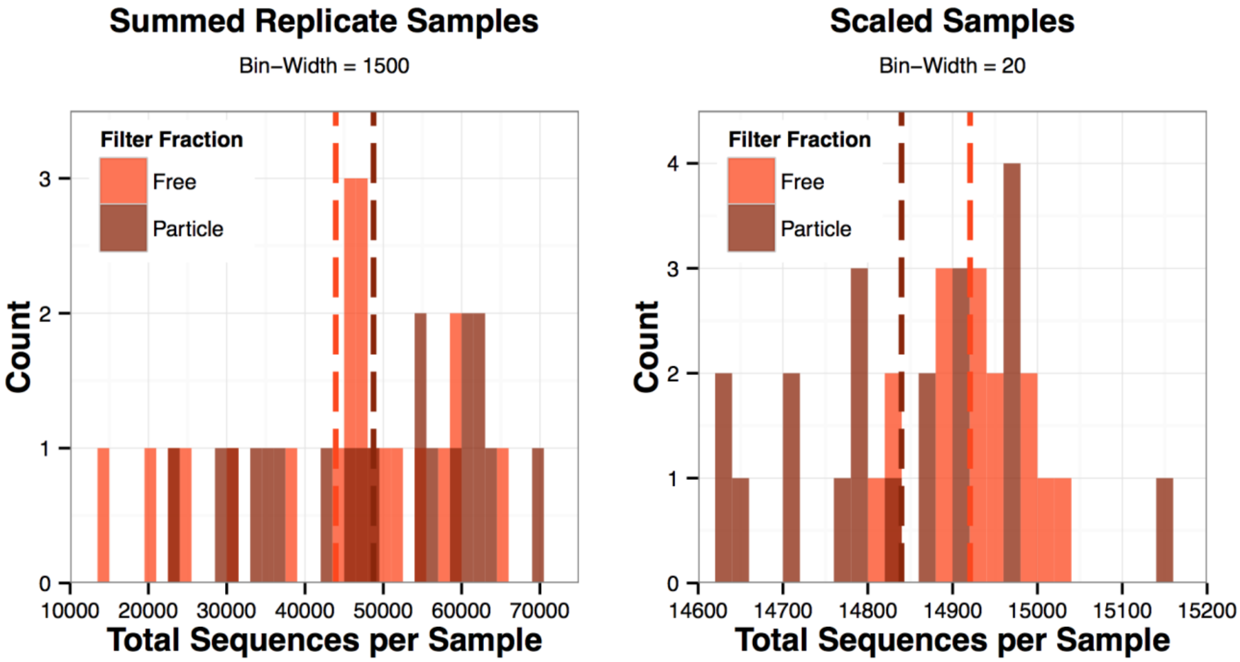
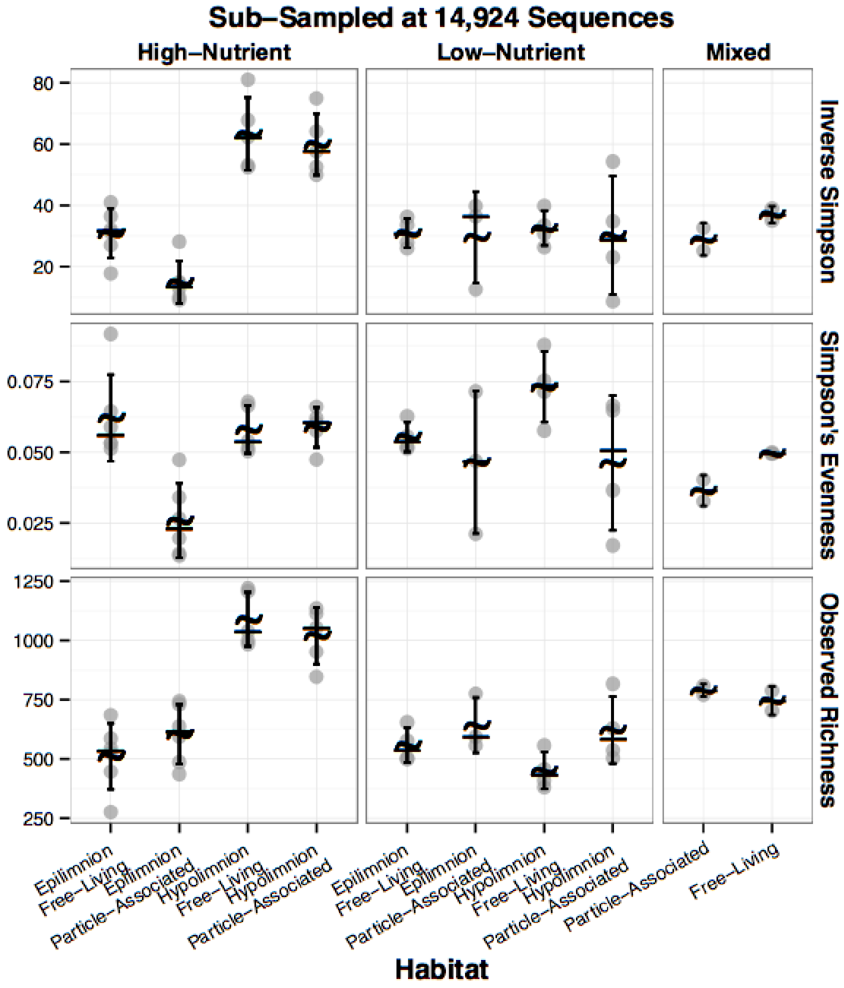


Figure SI. 2.8. Within-sample OTU richness and evenness

Next Page: Within-sample OTU richness and evenness impacted by **(Left panel)** sub-sampling (rarefy-ing) the data 100 times without replacement at 14,924 sequencing reads and **(right panel)** by scaling the reads to 14,925 sequencing reads as describe within the methods and Figure S2, right panel legend (sequencing depth range is 521 sequences). As the scaling method (right panel) is not the standard method for alpha diversity metrics, caution must be taken when interpreting the plot. Raw (gray points), mean (black tilde) and median (black horizontal bar) values and standard deviation (error bars) for **(Top panel)** the Inverse Simpson **(Middle panel)** Simpson's measure of evenness and **(Bottom panel)** OTU richness within each lake habitat sampled.


Sub–Sampled at 14,924 Sequences
High–Nutrient
Low–Nutrient
Mixed
Inverse Simpson
Simpson's Evenness
Observed Richness
80
60
40
20
0.075
0.050
0.025
1250
1000
750
500
250
Epilimnion Free–Living
Epilimnion Particle–Associated
Hypolimnion Free–Living
Hypolimnion Particle–Associated
Particle–Associated
Free–Living
Habitat

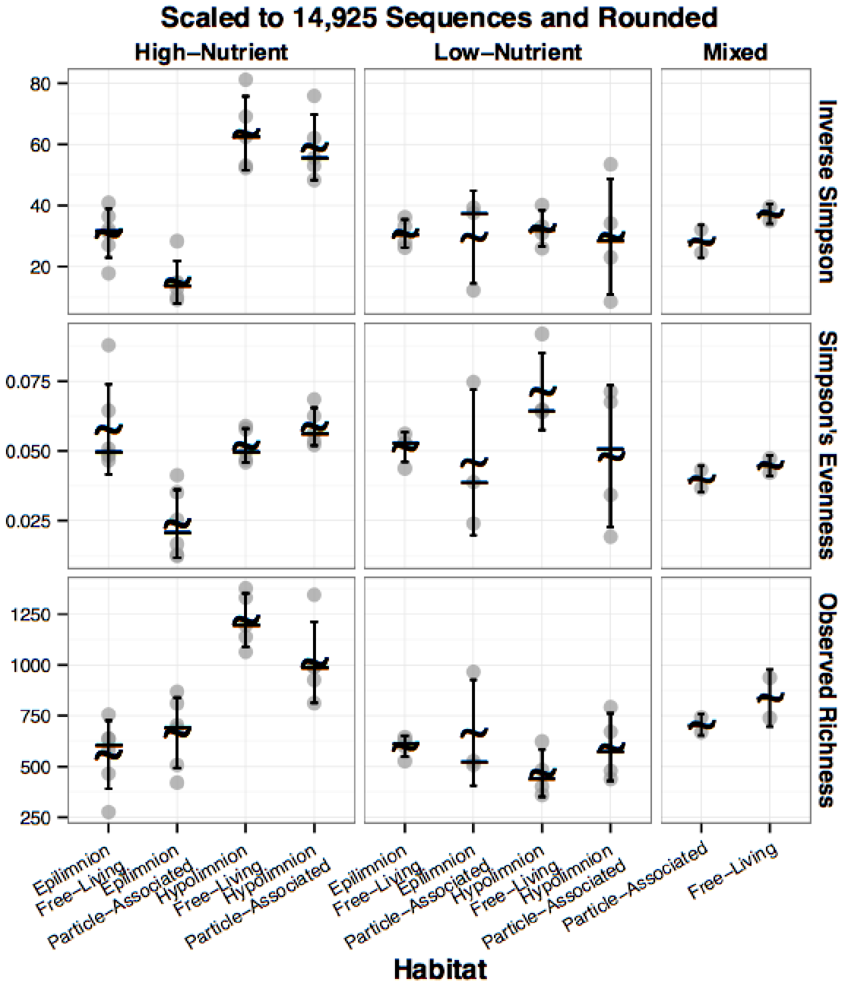

Scaled to 14,925 Sequences and Rounded
High–Nutrient
Low–Nutrient
Mixed
Inverse Simpson
Simpson's Evenness
Observed Richness
80
60
40
20
0.075
0.050
0.025
1250
1000
750
500
250
Epilimnion Free–Living
Epilimnion Particle–Associated
Hypolimnion Free–Living
Hypolimnion Particle–Associated
Particle–Associated
Free–Living
Habitat

Figure SI. 2.9. Phylum abundance of the 3 lake layer habitats.

Including the mixed lake, of the 17 most abundant phyla and classes of *Proteobacteria*. Box and whisker plots of the phylum abundance across the epilimnion, hypolimnion and mixed lake samples. Numbers in parentheses within the legend represent sample sizes of each lake habitat.

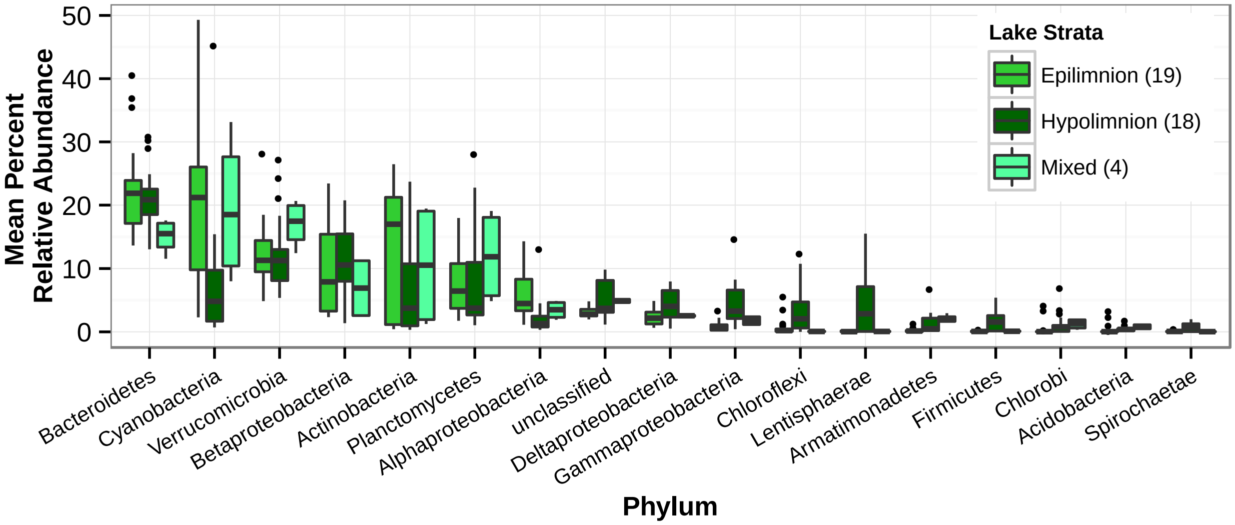


Figure SI. 2.10. Mean percent relative abundance of each taxonomic group.

Mean abundance of each phylum or class is plotted with error bars representing the standard error of the mean.

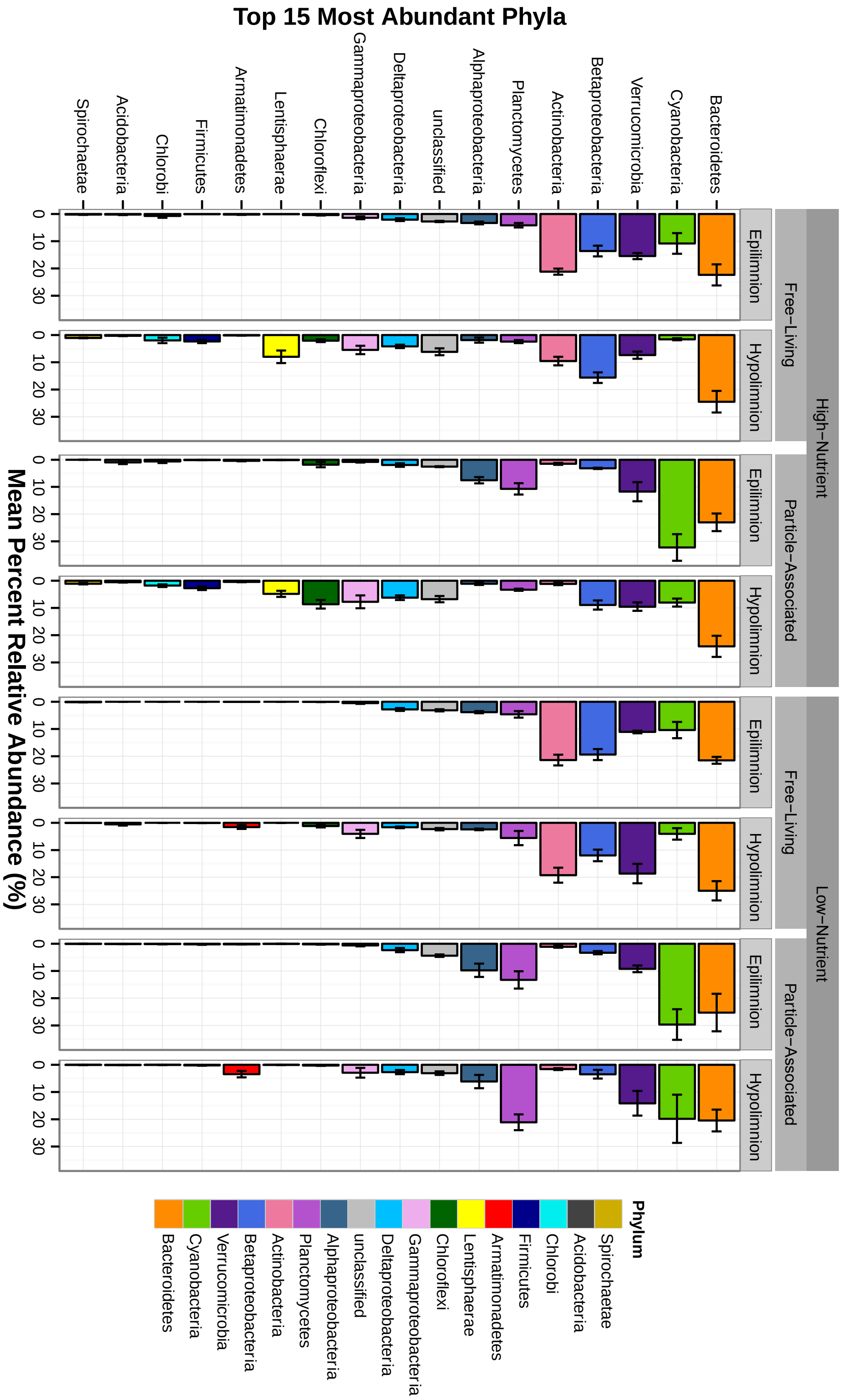


Figure SI. 2.11. Differentially abundant phyla across all specific lake habitats.

(Left) Heat map of the significantly differentially abundant (based on $\log_2$-ratios of relative abundance data) bacterial phyla and classes of the *Proteobacteria* between filter fractions, nutrient levels, and lake layers (titles). Blue represents a significant differential abundance of a phylum or class in the particle-associated fraction, high-nutrient lakes, or hypolimnion. Yellow represents a significant differential abundance of a phylum or class in the free-living fraction, low-nutrient lakes, or epilimnion. Only significant phyla identified were included. (**Right)** Bar graph of the mean relative abundance of the bacterial phylum or class of the *Proteobacteria* across the entire data set sorted from high to low relative abundance. Error bars represent the standard error of the mean.

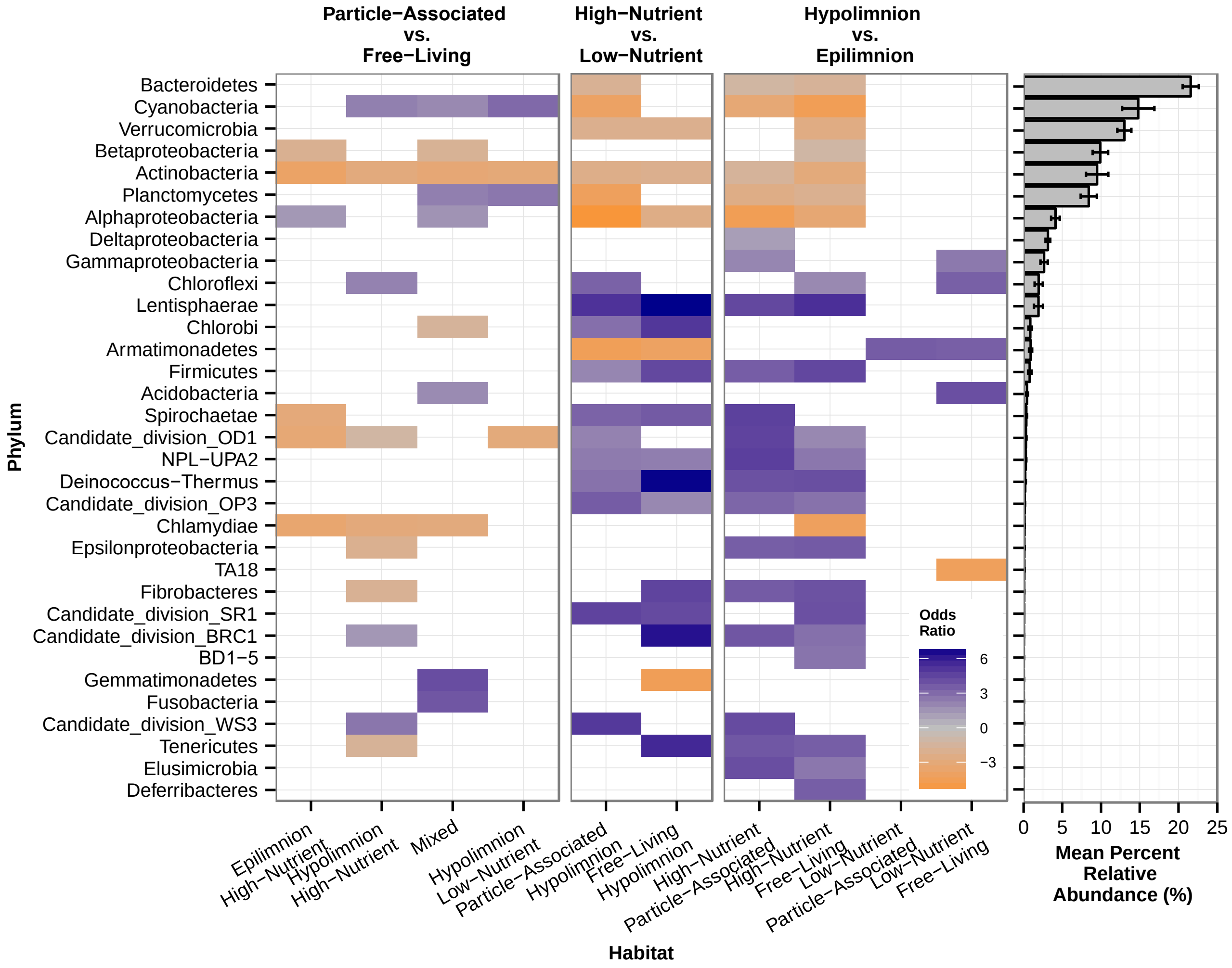


Figure SI. 2.12. Differentially abundant Genera or Freshwater Tribes.

Next Page: Based on the average Log$_2$-ratio of significant OTUs, across lake habitats. Heat map of the significantly differentially abundant (based on log$_2$-ratios of relative abundance data within each genera) bacterial OTUs within each genera between filter fractions, nutrient levels, and lake layers (titles). Blue represents a significant differential abundance of a genus or freshwater tribe in the particle-associated fraction, high-nutrient lakes, or hypolimnion. Yellow represents a significant differential abundance of a genus in the free-living fraction, low-nutrient lakes, or epilimnion. Only habitats with significant differentially abundant OTUs identified were included. Only genera or freshwater tribes with significant differentially abundant OTUs within the dataset are included.

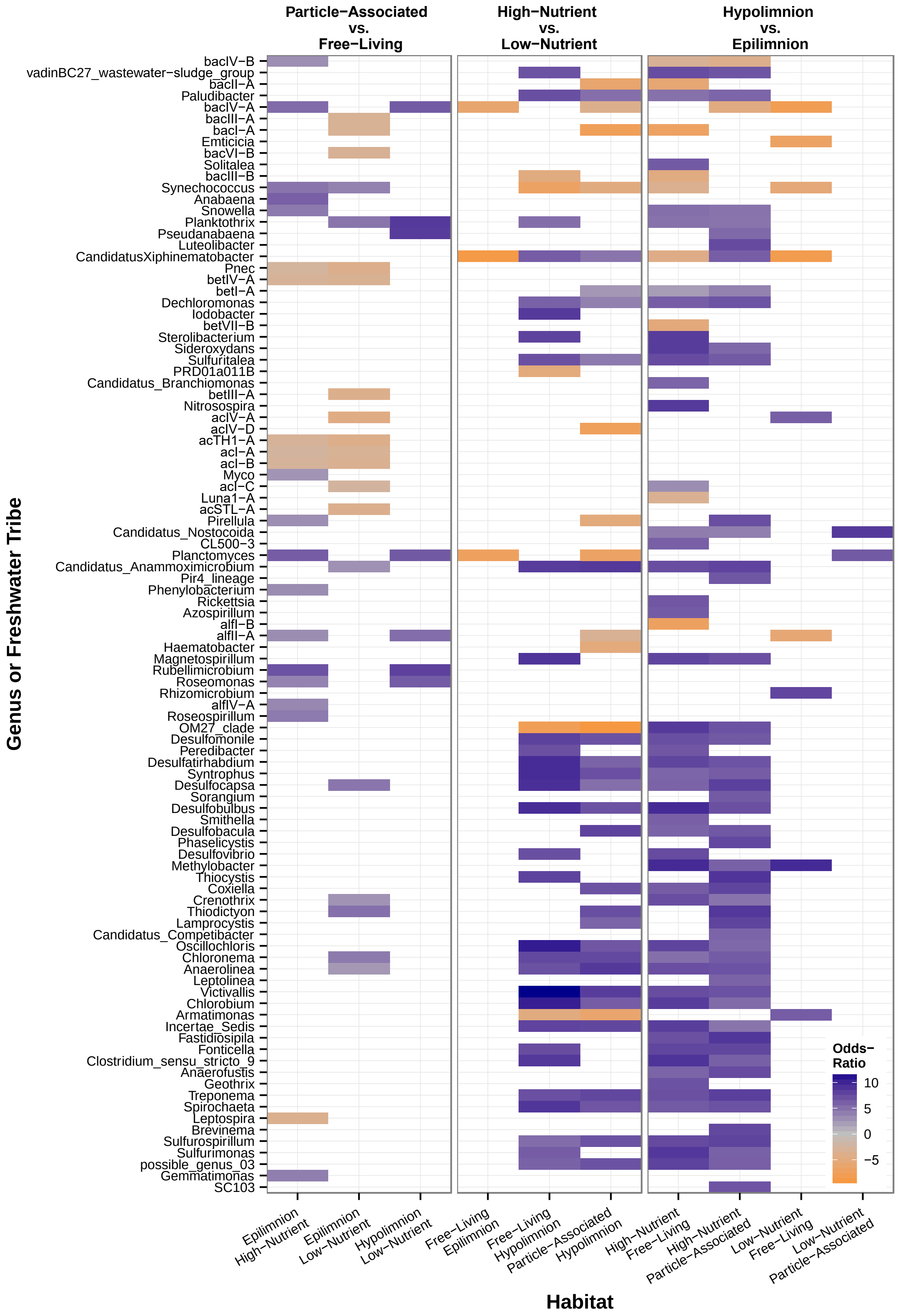
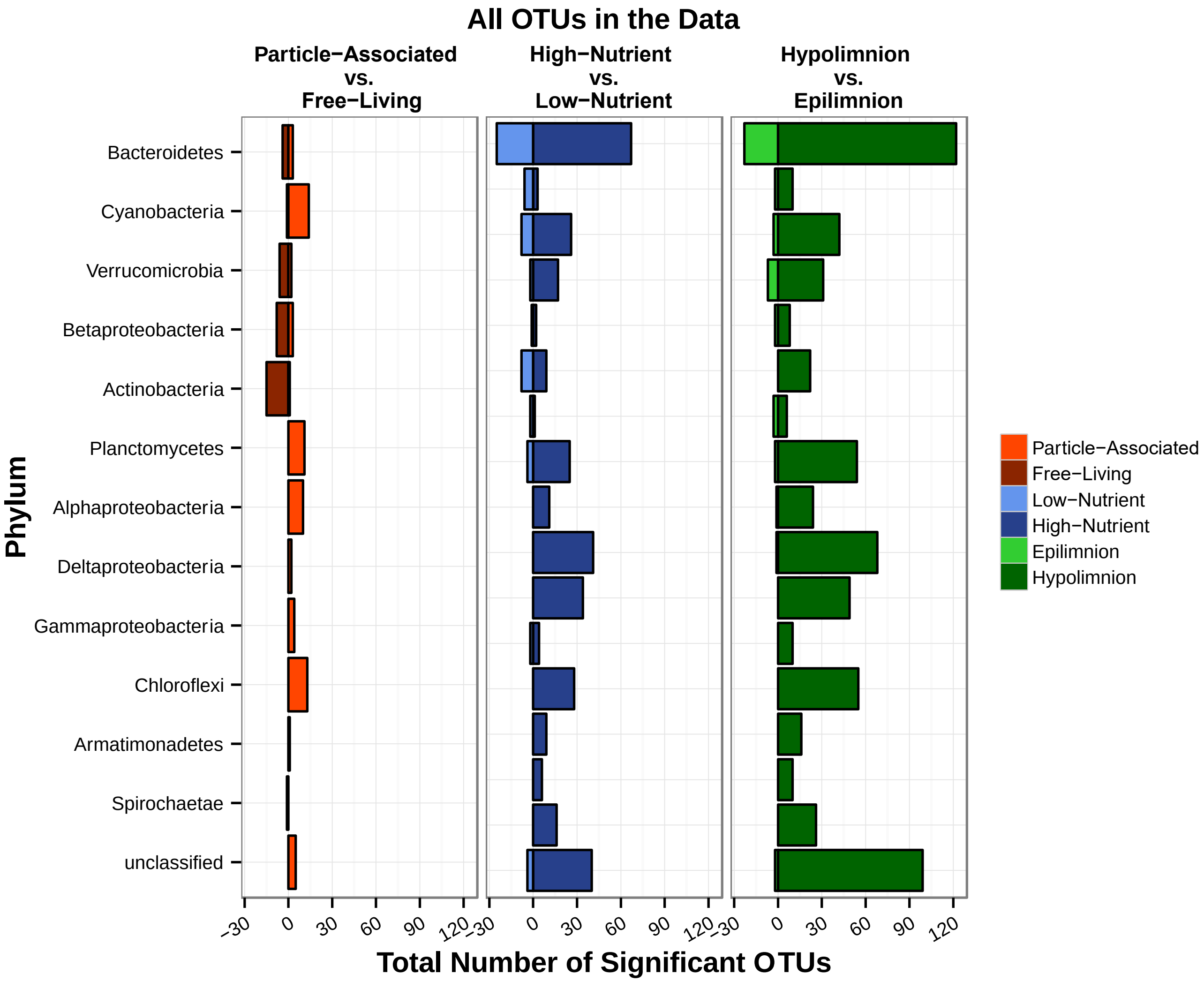

Particle−Associated vs. Free−Living
High−Nutrient vs. Low−Nutrient
Hypolimnion vs. Epilimnion
Genus or Freshwater Tribe
bacIV−B
vadinBC27_wastewater−sludge_group
bacII−A
Paludibacter
bacIV−A
bacIII−A
bacI−A
Emticicia
bacVI−B
Solitalea
bacIII−B
Synechococcus
Anabaena
Snowella
Planktothrix
Pseudanabaena
Luteolibacter
CandidatusXiphinematobacter
Pnec
betIV−A
betI−A
Dechloromonas
Iodobacter
betVII−B
Sterolibacterium
Sideroxydans
Sulfuritalea
PRD01a011B
Candidatus_Branchiomonas
betIII−A
Nitrosospira
acIV−A
acIV−D
acTH1−A
acI−A
acI−B
Myco
acI−C
Luna1−A
acSTL−A
Pirellula
Candidatus_Nostocoida
CL500−3
Planctomyces
Candidatus_Anammoximicrobium
Pir4_lineage
Phenylobacterium
Rickettsia
Azospirillum
alfI−B
alfII−A
Haematobacter
Magnetospirillum
Rubellimicrobium
Roseomonas
Rhizomicrobium
alfIV−A
Roseospirillum
OM27_clade
Desulfomonile
Peredibacter
Desulfatirhabdium
Syntrophus
Desulfocapsa
Sorangium
Desulfobulbus
Smithella
Desulfobacula
Phaselicystis
Desulfovibrio
Methylobacter
Thiocystis
Coxiella
Crenothrix
Thiodictyon
Lamprocystis
Candidatus_Competibacter
Oscillochloris
Chloronema
Anaerolinea
Leptolinea
Victivallis
Chlorobium
Armatimonas
Incertae_Sedis
Fastidiosipila
Fonticella
Clostridium_sensu_stricto_9
Anaerofustis
Geothrix
Treponema
Spirochaeta
Leptospira
Brevinema
Sulfurospirillum
Sulfurimonas
possible_genus_03
Gemmatimonas
SC103
Epilimnion High−Nutrient
Epilimnion Low−Nutrient
Hypolimnion Low−Nutrient
Free−Living Epilimnion
Free−Living Hypolimnion
Particle−Associated Hypolimnion
High−Nutrient Free−Living
High−Nutrient Particle−Associated
Low−Nutrient Free−Living
Low−Nutrient Particle−Associated
Habitat
Odds−Ratio
10
5
0
−5

Figure SI. 2.13. Significant differentially abundant OTUs**.**

OTUs within the 17 most abundant phyla and classes of *Proteobacteria*. Total number of OTUs that were significantly differentially abundant between filter fractions, nutrient levels, and lake layers. For each comparison, the number of differentially abundant OTUs in each habitat combination was summed.



Chapter III:

Microhabitats shape diversity-productivity relationships in freshwater bacterial communities[2]

Abstract

Eukaryotic communities commonly display a positive relationship between biodiversity and ecosystem function (BEF). Based on current studies, it remains uncertain to what extent these findings extend to bacterial communities. An extrapolation from eukaryotic relationships would predict there to be no BEF relationships for bacterial communities because they are generally composed of an order of magnitude more taxa than the communities in most eukaryotic BEF studies. Here, we sampled surface water of a freshwater, estuarine lake to evaluate BEF relationships in bacterial communities across a natural productivity gradient. We assessed the impact of habitat heterogeneity - an important factor influencing eukaryotic BEFs - on the relationship between species richness, evenness, phylogenetic diversity, and heterotrophic productivity by sampling co-occurring free-living (more homogenous) and particle-associated (more heterogeneous) bacterial habitats. Diversity measures, and not environmental variables, were the best predictors of particle-associated heterotrophic production. There was a strong, positive, linear relationship between particle-associated bacterial richness and heterotrophic productivity that was strengthened when considering evenness. There were no observable BEF trends in free-living bacterial communities. In contrast, per-capita but not community-wide heterotrophic productivity increased across both habitats as communities were composed of taxa that were more phylogenetically clustered. This association indicates that communities with

[2] Available on *bioRxiv*: Schmidt ML, Biddanda BA, Weinke AD, Chiang E, Januska F, Props R, Denef VJ. Microhabitats shape diversity-productivity relationships in freshwater bacterial communities. http://dx.doi.org/10.1101/231688.

more phylogenetically related taxa have higher per-capita heterotrophic production than communities of phylogenetically distantly related taxa. Our findings show that lake heterotrophic bacterial productivity can be positively affected by evenness and richness, negatively by phylogenetic diversity, and that BEF relationships are contingent on microhabitats. These results provide a stepping stone for comparison of bacterial biodiversity-productivity relationships to theory developed for Eukarya to bacterial communities.

Introduction

Our planet is currently experiencing an extreme species extinction event (Thomas et al., 2004; Wake & Vredenburg, 2008). Concern about such declines in biodiversity has resulted in hundreds of studies evaluating the relationship between biodiversity and ecosystem functions (BEF), with a large focus on terrestrial plant ecosystems. BEF relationships are generally positive and asymptotic and thus biodiversity loss causes a small change in ecosystem function at first and then, at some tipping point, a dramatic decrease in function (Cardinale et al., 2012, 2012; Tilman et al., 2014). While the focus of local and global diversity loss is typically on eukaryotic organisms, bacterial biodiversity has also been shown to be decreasing at local scales within the human gut (Blaser, 2014) and terrestrial ecosystems (Singh et al., 2014). Of particular concern is the loss of diversity of bacterial guilds responsible for key geochemical transformations, such as methane oxidation (Levine et al., 2011) that controls rates of methane emissions. Yet, the study of BEF relationships has been more limited for Bacteria and Archaea.

Based on the asymptotic BEF relationships observed for eukaryotic communities of up to 20 species, the large range of species richness observed in natural bacterial communities (hundreds to thousands) may suggest an absence of bacterial BEF relationships. While some studies have not found a BEF relationship for broad processes, such as heterotrophic respiration or biomass production, that are performed by many taxa (Levine et al., 2011), positive relationships have been shown for narrow processes performed by few taxa such as denitrification (Philippot et al., 2013), methanotrophy (Levine et al., 2011), and the degradation of triclosan and microcystin (Delgado-Baquerizo et al., 2016). Yet, studies have shown evidence of the contrary where positive BEF relationships exist even for processes performed by large numbers of taxa, such as carbon substrate oxidation (Langenheder et al., 2010) and bacterial respiration (Bell et al., 2005; Delgado-Baquerizo et al., 2016). These positive relationships with bacterial respiration have been maintained, though with a weaker slope, for up to a month-long experiment (Bell et al., 2005).

Beyond the impact of the number of species, phylogenetic relatedness is predicted to influence BEF relationships based on the phylogenetic limiting similarity hypothesis. The phylogenetic limiting similarity hypothesis posits that distantly related organisms will have more dissimilar niches and therefore reduced competition and a higher likelihood of coexistence (Violle et al., 2011). Therefore, it predicts that communities will have high phylogenetic diversity due to competitive exclusion of closely related species. Indeed, some papers show relationships across different ecosystems between phylogenetic diversity and ecosystem functions (Cadotte et al., 2008; Jiang et al., 2010; Violle et al., 2011). However, studies with freshwater green algae (Fritschie et al., 2014; Venail et al., 2014) did not find this relationship. A recent study found the opposite result by showing that closely related green algal species had weaker competition and more facilitation than distantly related species (Narwani et al., 2017). While relationships between phylogenetic relatedness among community members and ecosystem function have been assessed in bacterial systems (Tan et al., 2012; Galand et al., 2015; Roger et al., 2016), most work has focused on low-diversity, experimentally-assembled communities with bacteria that can be grown in culture. We need to expand these findings to communities with richness levels typically found in natural communities.

The nature of BEF relationships and the mechanism(s) that underpins them may depend on habitat structure or heterogeneity. Increasing habitat heterogeneity or environmental complexity has been found to enhance the strength of BEF relationships (Tylianakis et al. 2008; Replansky and Bell, 2009; Langenheder et al., 2010) presumably due to a greater role for niche complementarity effects in heterogeneous environments (Tiunov and Scheu, 2005; Cardinale 2011). While habitat heterogeneity contributes to increased diversity within bacterial populations and communities (Zhou et al., 2008; Shade et al., 2008), the influence of habitat heterogeneity on BEF relationships remains poorly studied for bacterial systems. One exception is a study by Langenheder et al. (2010) which found that in manipulated bacterial communities with up to six species environmental complexity (*i.e.*, resource richness variation with up to 3 substrates), in addition to bacterial species richness, positively impacted bacterial function but that species and resource richness did not interact with each other.

In this study, we hypothesized that bacterial diversity would be positively correlated with bacterial heterotrophic production, and that this relationship would be stronger in more heterogeneous environments. We simultaneously surveyed free-living and particle-associated surface water bacterial communities. Particulate matter comprises a variety of types and sizes of particles with each particle also harboring physicochemical gradients (Simon et al., 2002), and hence represents a more heterogeneous habitat than the surrounding water. While many previous studies have focused on the community composition (*e.g.*, Bižić-Ionescu et al., 2014, Schmidt et al., 2016) or productivity (*e.g.*, Crump et al., 1998) of these two microhabitats, here we specifically test whether there are BEF relationships within these co-occurring habitats. We tested BEF relationships using a variety of diversity metrics including observed richness, species dominance, and phylogenetic diversity. We focused on heterotrophic bacterial production as our measure of ecosystem function, as it is a key process affecting freshwater bacterial growth that in turn fuels the macroscopic food web through their recycling of nutrients bound in organic matter (Cotner & Biddanda, 2002).

Methods

Lake sampling and sample processing

Surface water samples were collected at 1 meter depth from 4 long-term sampling stations (Steinman et al., 2008) in mesotrophic Muskegon Lake (**Figure S1. 3.4**), which is a freshwater estuarine lake connecting the Muskegon River and Lake Michigan. These stations included the mouth of the Muskegon River (43.250133,-86.2557), the channel to Bear Lake (43.238717,-86.299283; a hypereutrophic lake), the channel to Lake Michigan (43.2333,-86.3229; oligotrophic lake), and the deepest basin of Muskegon Lake (43.223917,-86.2972; max depth = 24 m).

Samples were collected during the morning to early afternoon of 3 days in 2015 (May 12, July 21, & September 30) aboard the R/V W.G. *Jackson*. All water samples were collected with vertical Van Dorn samplers. Additionally, a vertical profile of temperature (T), pH, specific conductivity (SPC), oxidation-reduction potential (ORP), chlorophyll (Chl*a*), total dissolved solids (TDS), and dissolved oxygen (DO) was constructed at each station to characterize the water column using a calibrated YSI 6600 V2-4 multiparameter water quality sonde (Yellow Springs Instruments Inc.). Total Kjeldahl nitrogen (TKN), ammonia (NH_3), total phosphorus (TP), and alkalinity (Alk) were processed from whole water while nitrate (NO_3), phosphate (PO_4), and chloride (Cl^-) were hand filtered using a 60 mL syringe fitted with a Sweeny filter holder with a 13 mm diameter 0.45 µm pore size nitrocellulose filters (Millipore) and were determined by standard wet chemistry methods in the laboratory (EPA, 1993).

Bacterial abundance by epifluorescence microscopy

Lake surface water samples were processed within 2-6 hours of their collection for determination of heterotrophic bacterial abundance. Samples (5 mL) were preserved with 2% formalin and 1 mL subsamples were stained with acridine orange stain and filtered onto black 25 mm 0.2 µm pore size polycarbonate filters (Millipore) at a maximum pressure of 0.1 Bar or 1.5 PSI. Prepared slides were stored frozen until enumeration by standard epifluorescence microscopy at 1000x magnification under blue light excitation (Hobbie et al. 1977). Bacteria within the field of view (100 µm x 100 µm) that were not associated with any particles were counted as free-living bacteria, whereas bacteria that were on particles were counted as particle-associated. Sample filtration may bias counts due to free-living or particle-associated cells being hidden on the underside of particles, free-living bacteria settling on top of particles, or particle-associated cells dislodging. In the absence of any quantitative studies that have rigorously addressed this issue, we have assumed the net effect of these opposing methodological biases to be negligible in the present study**.**

Heterotrophic bacterial community production measurements

Community-wide heterotrophic bacterial production was measured using [^{3}H] leucine incorporation into bacterial protein (Kirchman et al. 1985; Simon and Azam, 1989). Quadruplicate 1m water samples were incubated in the dark under *in situ* temperatures for 1 hour (hr) with a 20 nM final concentration of [^{3}H]-leucine. One 50% trichloroacetic acid (TCA)-killed control was run for every three live incubations of the same sample. At the end of the incubation with [^{3}H]-leucine, cold TCA-extracted samples were filtered onto 3 µm filters that represented the leucine incorporation by particle-associated bacteria (>3.0 µm). Each filtrate was collected and filtered onto 0.2 µm filters and the activity therein represented incorporation of leucine by free-living bacteria (>0.2 µm-<3 µm). The rate of uptake was linear over a 2 hr incubation period and the controls accounted for 0.5-6% of the ^{3}H label found in live treatments. On the basis of such repeatable linear uptake measurements over the representative period of the incubations, we presumed there was no measurable recirculation of incorporated ^{3}H back into solution. The timeline for our incubations (1 hr) as well as the sensitivity of the ^{3}H method were insufficient to distinguish between the production rates of r- versus k-selected taxa. However, longer incubations would have likely led to problems of non-linear uptake and recirculation of the incorporated ^{3}H (Kirchman et al., 1985). Thus, we chose to run the incubations over the short time of 1 hr where bacterial community production measurements were more reliable. Measured leucine incorporation during the incubation was converted to bacterial carbon production rate using a standard theoretical conversion factor of 2.3 kg C per mole of leucine (Simon and Azam, 1989). Per-capita heterotrophic production was estimated by dividing heterotrophic production by the cell counts measured in each fraction.

Preservation of bacterial filters in the field

Microbial biomass for the particle-associated (> 3 µm) fraction and the free-living (3–0.22 µm fraction) bacterial fraction was collected by sequential in-line filtration on 3 µm isopore polycarbonate (TSTP, 47 mm diameter, Millipore, Billerica, MA, USA) and 0.22 µm Express Plus polyethersulfone membrane filters (47 mm diameter, Millipore, MA, USA). We used 47 mm polycarbonate in-line filter holders (Pall Corporation, Ann Arbor, MI, USA) and an E/S portable peristaltic pump with an easy-load L/S pump head (Masterflex®, Cole Parmer Instrument Company, Vernon Hills, IL, USA). The total volume filtered varied from 0.8–2.2 L with a maximum filtration time of 16 minutes per sample. Filters were submerged in RNAlater (Ambion) in 2 mL cryovials, frozen in liquid nitrogen, and transferred to a −80°C freezer until DNA extraction.

DNA extraction, sequencing and processing

DNA extractions were performed using an optimized method based on the AllPrep DNA/RNA/miRNA Universal kit (Qiagen; McCarthy et al., 2015; details in supplementary methods). Extracted DNA was sequenced using Illumina MiSeq V2 chemistry 2 × 250 (500 cycles) of dual index-labelled primers that targeted the V4 hypervariable region of the 16S rRNA gene (515F/806R) (Caporaso et al., 2012; Kozich et al., 2013) at the Microbial Systems Laboratories at the University of Michigan Medical School in July 2016. RTA V1.17.28 and MCS V2.2.0 software were used to generate data. Fastq files were submitted to NCBI sequence read archive under BioProject accession number PRJNA412984. We analyzed the sequence data using MOTHUR V.1.38.0 (seed = 777; Schloss et al., 2009) based on the MiSeq standard operating procedure accessed on 3 November 2015 and modified with time (see data accessibility and supplemental methods). We used a combination of the Silva Database (release 123; Quast et al., 2013) and the freshwater TaxAss 16S rRNA database and pipeline (Rohwer et al., 2017, accessed 18 August 2016) for classification of operational taxonomic units (OTUs). All non-bacterial and chloroplast sequences were pruned out of the dataset and replicate samples were merged by summing sample sequencing read counts using the *merge_samples* function (phyloseq). A batch script for our protocol can be found in this project's GitHub page in https://github.com/DenefLab/Diversity_Productivity/blob/master/data/mothur/mothur.batch.taxass.

Estimating Diversity

We focused our diversity analyses on observed richness and the inverse Simpson's index. We report the number of OTUs in our samples (*i.e.,* observed richness) to best compare with the broader BEF literature. We also describe the inverse Simpson's metric, which is a measure of species dominance, representing the proportional abundance of taxa in the community, a property that is a major difference in eukaryotic and prokaryotic communities. A higher inverse Simpson's value indicates that there are more dominant members in the community and that the community has higher evenness. More specifically, it is one over the probability that two randomly selected individuals (with replacement) will belong to the same OTU (Tuomisto, 2012). In addition, to be consistent with other literature in microbial ecology, we also report the Shannon entropy, which accounts for both abundance and evenness of species present. Finally, we describe a pure measure of evenness known as Simpson’s evenness that removes any potential effect of observed richness (Magurran, 2004; *see below*). To get the best estimate of each diversity metric, each sample was subsampled to 6,664 sequences (the smallest library size) with replacement and was averaged over 100 trials. Observed richness, Shannon entropy, and inverse Simpson’s index were calculated using the *diversity* function within the vegan (Oksanen et al., 2013) R package via the *estimate_richness* function in the phyloseq (McMurdie and Holmes, 2013) R package. Simpson’s Evenness was calculated by dividing the inverse Simpson’s index by the observed richness (Magurran, 2004). To calculate phylogenetic diversity, we first removed OTUs that had a count of 2 sequences or less throughout the entire dataset, as these are more prone to be artefacts originating from sequencing errors or the OTU clustering algorithm. Representative sequences of each of the 1,891 remaining OTUs were collected from the aligned fasta file produced within mothur, and header names in the mothur output fasta file were modified using bbmap (Bushnell, 2016) to only include the OTU name. A phylogenetic tree was created with FastTree using the GTR+CAT (general time reversible) model (Price et al., 2010). Mismatches between the species community data matrix and the phylogenetic tree were checked with the *match.phylo.comm* command (picante). Finally, both abundance-unweighted and -weighted phylogenetic diversity was estimated using specifications described in the next paragraph with the picante R package.

The most common phylogenetic diversity (PD) measure is Faith’s PD (Faith, 1992), however, this metric is very strongly correlated with species richness (**Figure SI 3.5**). Instead, the mean pairwise phylogenetic distance (or MPD) was calculated (*ses.mpd* function in the *Picante* R package (Kembell et al., 2010), null.model = “independentswap”). The MPD measures the average phylogenetic distance between all combinations of two taxa pulled from the observed community and compares it with a null community of equal richness pulled from the gamma diversity of all the samples (*see supplemental methods for more details*). Values higher than zero indicate phylogenetic evenness or overdispersion (higher phylogenetic diversity) while values less than zero indicate phylogenetic clustering (lower diversity) or that species are more closely related than expected according to the null community (Kembel, 2009). Thus, this phylogenetic metric is relative. Here we refer to the SES_{MPD} as the “phylogenetic diversity” for simplicity and clarity.

Statistical analysis

Further analysis of sequence data was performed in R version 3.4.2 (R Core Team 2017; *see supplemental methods for more details*). To test which variable(s) were the best predictors of community and per-capita heterotrophic production, we performed variable selection via a lasso regression (using the *glmnet* R package, alpha = 1, and lambda.1se as the tuning parameter (Friedman et al., 2010)) on all of the environmental, biodiversity, and principal component variables. To further validate the lasso regression results, we performed ordinary least squares (OLS) regressions on all variables, including the principal components (PCA) of the euclidean distances of the environmental data. We used the Akaike information criterion (AIC) (accessed with the *broom::glance()* command) to select the best performing OLS regression model.

Data and code availability

Original fastq files can be found on the NCBI sequence read archive under BioProject accession number PRJNA412984. Processed data and code can be found on the GitHub page for this project at https://deneflab.github.io/Diversity_Productivity/ with the main analysis at https://deneflab.github.io/Diversity_Productivity/Final_Analysis.html

Results

Free-living communities had more cells and higher community-wide heterotrophic production, but particle-associated communities had higher per-capita heterotrophic production

We observed an order of magnitude more cells per milliliter ($p = 1 \times 10^{-6}$, **Figure 3.1A**) and ~2.5 times more community-wide heterotrophic production in the free-living fraction (p = 0.024, **Figure 3.1B**). However, when calculated per-capita, particle-associated bacteria were on average an order of magnitude more productive than free-living bacteria ($p = 7 \times 10^{-5}$**, Figure 3.1C**). Particle-associated and free-living cell abundances in samples taken from the same water sample did not correlate (**Figure SI 3.6A**). Heterotrophic production between corresponding free-living and particle-associated fractions from the same water sample were positively correlated for both community (Adjusted R^2 = 0.40, p = 0.017; **Figure SI 3.6B**) and per-capita production rates (Adjusted R^2 = 0.60, p = 0.003; **Figure SI 3.6C**).

Particle-associated communities are more diverse in terms of observed richness and Shannon Entropy while free-living communities are more phylogenetically diverse

Across all samples, particle-associated bacterial communities were more diverse than free-living communities when considering richness and Shannon entropy (**Figures 3.2A & SI 3.7**), but similar in the inverse Simpson's index and Simpson's evenness (**Figure 3.2B & SI 3.7B**).

Particle-associated bacterial community richness was always higher than in free-living communities and was maintained across the four sampling stations in the lake (**Figure SI 3.8A**). Particle-associated samples at the river and Bear lake stations were on average more OTU-rich than the outlet to Lake Michigan and the Deep stations. Additionally, the river station had almost twice the inverse Simpson's value as compared with all other lake stations (Mean inverse Simpson Indices: Outlet = 23.6; Deep = 23.7; Bear = 35.3; River = 59.1; **Figure SI 3.8A**).

Particle-associated communities were more phylogenetically clustered than free-living communities based on unweighted phylogenetic diversity (p = 0.01, **Figure 3.3A**). Compared to other particle-associated samples, the outlet station that connects to oligotrophic Lake Michigan had a much larger unweighted phylogenetic diversity, indicating phylogenetic overdispersion (**Figure SI 3.8A**). Nevertheless, no sample across the entire dataset differed significantly in their unweighted phylogenetic diversity from the null model with a significance threshold p-value of 0.05. There was no difference between weighted phylogenetic diversity in particle-associated versus free-living communities (**Figure SI 3.8A**).

Diversity-Productivity relationships are only observed in particle-associated communities

There was a strong, positive, linear BEF relationship between community-wide (**Figures 3.2C-D & SI 3.7 C-D**) and per-capita (**Figures 3.2E-F & SI 3.7 E-F**) heterotrophic productivity and all richness and evenness diversity metrics in the particle-associated communities, while no BEF relationships were observed for the free-living communities. The inverse Simpson's index explained the most variation in community-wide (**Figure 3.2D;** Adjusted R^2 = 0.69, p = 5 x 10^{-4})

and per-capita (**Figure 3.2F;** Adjusted $R^2 = 0.69$, $p = 0.001$) heterotrophic production. These results are robust across a range of minimum OTU abundance filtering thresholds (see *Sensitivity Analysis of Rare Taxa* in the supplemental methods and **Figure SI 3.9**) and hold up for all threshold levels in Inverse Simpson and for richness until removal of 25 counts (community-wide heterotrophic production) and 15 counts (per-capita heterotrophic production). When the particle-associated and free-living samples were combined together into one linear model to test an overall relationship between diversity and community-wide heterotrophic production, there was no relationship (richness: $p = 0.86$; Shannon: $p = 0.99$; Inverse Simpson: $p = 0.36$), with the exception of a weak correlation for Simpson's Evenness (Adjusted $R^2 = 0.12$, $p = 0.054$). However, when particle-associated and free-living samples were combined together into one linear model to test an overall relationship between diversity and per-capita heterotrophic production, there was a strong relationship with observed richness (Adjusted $R^2 = 0.63$, $p = 3 \times 10^{-6}$), which broke down as evenness was weighted more (**Figure SI 3.11:** Shannon: Adjusted $R^2 = 0.52$, 6×10^{-5}; Inverse Simpson: Adjusted $R^2 = 0.48$, $p = 2 \times 10^{-4}$; Simpson's Evenness: $p = 0.48$).

Phylogenetic diversity correlated with per-capita heterotrophic production but not with community-wide production

Abundance-weighted phylogenetic diversity was not correlated with community or per-capita heterotrophic production (**Figure SI 3.12 C-D**) and therefore no further analyses were performed with this diversity metric.

There was a moderate, negative, linear relationship when particle-associated and free-living samples were combined together into one linear model to test an overall relationship between unweighted phylogenetic diversity and observed richness (**Figure 3.3B**; Adjusted $R^2 = 0.35$, $p = 0.001$). To further validate this trend, randomized communities were generated with an equal richness as the samples but with OTUs randomly picked across the dataset. The unweighted phylogenetic diversity was then calculated and regressed against each randomized richness and there was no relationship (**Figure SI 3.13;** Adjusted $R^2 = -0.02$, $p = 0.44$), verifying the negative relationship in the actual samples. When particle-associated and free-living samples were individually run in separate linear models to test for habitat-specific relationships between unweighted phylogenetic diversity and observed richness, no trend was found in either particle-associated or free-living models (**Figure 3.3B;** Particle: Adjusted $R^2 = 0.14$, p = 0.12; Free = Adjusted $R^2 = -0.10$, p = 0.97). In other words, particle-associated and free-living diversities did not have individual effects on community-wide or per-capita heterotrophic production but rather, all samples were necessary for a correlation between per-capita heterotrophic production and unweighted phylogenetic diversity.

There was no correlation between phylogenetic diversity and community-wide heterotrophic production (**Figure 3.3C**). However, a negative correlation was found when particle-associated and free-living samples were combined into one linear model to test an overall relationship between unweighted phylogenetic diversity and per-capita heterotrophic production (**Figure 3.3D**; $R^2 = 0.42$, $p = 5 \times 10^{-4}$). Therefore, these two results in combination indicated that communities composed of more phylogenetically similar OTUs had a higher per-capita heterotrophic production rate.

Diversity, not environmental variation, is the best predictor of particle-associated heterotrophic production

To identify variables that best predicted community-wide and per-capita heterotrophic production (*i.e.,* remove variables that were correlated with each other and/or uninformative variables), we performed lasso regression with all samples and individually with particle-associated and free-living samples. For prediction of community-wide heterotrophic production, only the inverse Simpson's index was selected for particle-associated samples whereas pH and PC5 were selected for free-living samples, and no variables were selected when all samples were included in the lasso regression. In contrast, for per-capita heterotrophic production, temperature and the inverse Simpson's index were selected for particle-associated samples whereas pH was the only predictor for free-living samples, and observed richness was the only predictor for all samples (plotted in **Figure SI 3.11A**). Therefore, the best model for particle-associated microhabitats *always* included inverse Simpson's index whereas free-living samples only included environmental variables, such as pH.

To further verify that there were no confounding impacts of seasonal and environmental variables on community-wide and per-capita heterotrophic production, we performed ordinary least square (OLS) regressions and a dimension-reduction analysis of the environmental variables through a principal component analysis (**Table SI 3.1 & 3.2; Figure SI 3.14**). Specifically, the first 2 environmental axes explained ~70% of the environmental variation in heterotrophic production among the sampling sites (**Figure SI 3.14**). Next, we predicted community-wide and per-capita heterotrophic production with all environmental variables and the first six principal components as predictor variables with individual particle-associated and free-living samples, and combined (*i.e.,* all samples) models (**Table SI 3.1 & 3.2**). The best single predictor of community-wide heterotrophic production was Inverse Simpson for particle-associated samples (AIC = 74.34; R^2 = 0.69), pH for the free-living samples (AIC =98.43; R^2 = 0.49, p = 0.006), and pH for all samples (AIC = 192.16; R^2 = 0.35) (Table S1). The best single predictor of per-capita heterotrophic production was Inverse Simpson for particle-associated samples (AIC = 8.29; R^2 = 0.69), pH for the free-living samples (AIC = -2.39; R^2 = 0.78), and observed richness for all samples (AIC = 24.72; R^2 = 0.63) (**Table SI 3.2**). Thus, the OLS regressions are in agreement with the lasso regressions.

Discussion

We examined bacterial biodiversity-ecosystem function (BEF) relationships in relation to two microhabitats within freshwater lakes: particulate matter and the surrounding water. First, we found that community-wide and per-capita heterotrophic productivity of particle-associated but not free-living bacterial communities showed a positive, linear BEF relationship with both richness and evenness contributing. Second, particle-associated heterotrophic production was better explained by diversity (*i.e.,* inverse Simpson's index) than by environmental parameters. Third, across both particle-associated and free-living communities, higher richness was associated with lower phylogenetic diversity which, in turn, was associated with higher per-capita heterotrophic bacterial production but not associated with community-wide heterotrophic production.

Microbes have a large diversity of metabolisms and the choice of which to focus on may inherently affect the BEF relationship. Indeed, “narrow” metabolic processes that are catalyzed by a small subset of taxa within bacterial communities, such as nitrogen and sulfur cycling, have been found to display BEF relationships (Levine et al., 2011; Delgado-Baquerizo et al., 2016). In contrast, for “broad” processes that are performed by the majority of taxa within a bacterial community, such as heterotrophic production (focus of the present study) and respiration, functional redundancy appears to weaken or remove the presence of BEF relationships (Griffiths et al., 2000; Wertz et al., 2006; Levine et al., 2011; Peter et al., 2011, Galand et al, 2015). These findings are in line with the absence of a BEF relationship for free-living bacterial communities in our study.

However, the above results and hypotheses surrounding narrow and broad processes are in conflict with the strong BEF relationship we observed in particle-associated bacterial communities. As such, our study shows that microhabitats or habitat heterogeneity can influence bacterial BEF relationships, in agreement with previous research in eukaryotic systems across a variety of ecosystems (Tylianakis et al., 2008; Cardinale 2011; Zeppilli et al., 2016). A study using controlled stream mesocosms by Cardinale (2011) found that niche complementarity effects are particularly important in more heterogeneous environments. In more heterogeneous streams, algal populations used different nutrients and avoided direct competition for resources, resulting in unique species occupying distinct and local microhabitats.

Our observational study could not directly test the role of niche complementarity effects. However, support for niche complementarity alone or in combination with species selection as the mechanism underlying the BEF relationship in particle-associated habitats is provided by the inverse Simpson’s index being the strongest predictor of community-wide heterotrophic production. As the inverse Simpson’s index represents a measure of species dominance, it is strongly affected by the evenness of abundant species. Communities that are more even have an increased likelihood for complementary species to be neighbors. However, it is interesting that the inverse Simpson’s metric explains more variation ($R^2 = 0.69$, **Figure 3.2D & F**) compared to Shannon entropy ($R^2 = 0.52$-0.55, **Figure SI 3.7 C & E**), which may indicate that dominant members of the community have more of an influence on heterotrophic production than does overall community evenness.

In our study, there are three main reasons why heterogeneity of particulate matter may allow for niche complementarity effects to occur and result in BEF relationships. First, particles have a two-fold layer of heterogeneity as they (A) may be composed of different substrates such as organic matter from terrestrial or aquatic environments and either heterotrophically or photosynthetically derived (Grossart, 2010), and (B) each particle may comprise physicochemical gradients as well (Simon et al., 2002). Second, particle-associated bacteria are typically more active per capita than compared to their free-living counterparts (**Figure 3.1C**; Harvey & Young, 1980; Crump et al., 1998; Riemann et al., 2000). Third, microbial interactions are more likely to occur between cells aggregated on particles as the interaction distances are usually much shorter (Cordero & Datta, 2016) compared to free-living bacterial cells. In fact, genes mediating social interactions, such as motility, adhesion, cell-to-cell transfer, antibiotic resistance, mobile element activity, and transposases, have been found to be more abundant in marine particles than compared to the surrounding water (Ganesh et al., 2014).

The importance of niche complementarity in microbial communities can also be deduced from recent findings in the field of microbiology, which have shown widespread metabolic interdependence among bacterial community members. First, a 2016 study that reconstructed 2,540 draft genomes of microbes found that most bacteria specialize in one particular step in sulfur and nitrogen pathways and "hand-off" their metabolic byproducts to nearby organisms (Anantharaman et al., 2016). It is likely that metabolic hand-offs, also known as cross-feeding, is a specific form of bacterial facilitation that will occur more in particle-associated compared to free-living communities. Indeed, Datta and Cordero's (2016) work on model marine particles found that taxa that are incapable of breaking down particles and instead rely on carbon produced by primary degraders thrive in later phases of particle degradation. Second, Lilja and Johnson (2016) demonstrated that different microbial cell types eliminate inter-enzyme competition by cross feeding, which increases substrate consumption by allowing intracellular resources to go towards a single enzyme, rather than having two enzymes that perform two separate reactions compete for nutrients within a cell. Third, some bacteria are unable to grow in laboratory cultures unless they are in co-culture with other organisms, which may be due to metabolic hand-offs or to growth factors such as siderophores or catalases (Stewart, 2012). Indeed, Bell and coauthors (2005) discussed that the positive relationship between bacterial respiration and species richness were due to "synergistic interactions among bacterial species of which complementarity is one possibility, had an important role in functioning."

Taking into account that (i) closely related taxa share more genes and metabolic pathways than distantly related bacterial taxa (Konstantinidis & Tiedje, 2005; Kim et al., 2014) and (ii) bacteria commonly have incomplete metabolic pathways, we propose that closely related bacteria may be most likely to hand-off their metabolic byproducts. This may be why we found that new taxa added to the community represented taxonomic clades similar to or already present in the community, and that these communities with lower phylogenetic diversity (relative to expected) had higher productivities. This result is in line with a recent study using freshwater algae and vascular plants that rejects predictions from the phylogenetic limiting similarity hypothesis (Narwani et al., 2017). However, recent bacteria-focused studies from Russel et al. (2017) and Venail and Vives (2013) found higher levels of antagonism (Russel et al., 2017) or more bacterial productivity (measured through colony forming units per mL;Venail and Vives, 2013) with more distantly related taxa. Both of these studies were performed in the lab with r-selected (*i.e.*, copiotrophic) species grown in stable, warm, aerobic, agar plate conditions. Thus, Venail and Vives (2013) and Russel et al. (2017) inherently break up potential interdependent relationships between bacteria either by creating artificial communities or evaluating pairwise interactions and remove the natural effect of spatial heterogeneity, environmental fluctuations, and the rest of the bacterial community. As a result, future studies on bacterial interactions and the role of phylogenetic diversity will need to maintain natural structure and complexity in bacterial communities.

Previous studies on bacterial BEF relationships have used three approaches to manipulate bacterial diversity (Krause et al., 2014): (1) dilution to extinction in which complex communities are diluted to more simple communities (Wertz et al., 2006; Peter et al., 2011; Philippot et al., 2013; see Roger et al., 2016 for a review of this approach), (2) manually assembled communities in culture (Tan et al., 2012; Salles et al., 2009), or (3) natural or manipulated environmental communities (Griffiths et al., 2000; Levine et al., 2011; Galand et al., 2015). In this study, we took the latter approach. In contrast to the other two approaches, this had the benefit of (1) maintaining high diversity with both abundant and rare taxa, (2) including both r- and k-selected organisms, (3) allowing natural environmental and ecological forcings to shape the community, and (4) evaluating BEF relationships in diversity and productivity ranges that reflect natural communities. Specifically, Langenheder et al. (2010) discussed that metabolic cross-feeding may have not been important in their experimental communities because their chosen substrates could have been completely degraded and unlikely to have produced metabolic byproducts that could be used by other taxa. In an observational system, these types of complexities should remain in the system. Admittedly, three inherent weaknesses to our approach were that (1) we cannot measure all the potential variables that influence heterotrophic productivity, (2) we only have 24 samples for a 12 versus 12 study, and (3) our analysis is correlational and we cannot manipulate the system to unequivocally separate causes and consequences of bacterial production. For example, strong correlations with heterotrophic production and pH in the free-living samples (Table S1 & S2) may point to pH being a consequence of rather than a cause of varying production levels. This is because bacterial production and bacterial respiration are positively correlated (del Giorgio & Cole, 1998) and with increased respiration and CO_2 production, pH may decrease due to CO_2 dissolution into the water.

Finally, we acknowledge that the typical sampling of bacterial communities and analysis using DNA sequencing reflects *all* bacteria present in the community and not necessarily only the *active* members of the community contributing to a given ecosystem function. In freshwater systems, up to 40% of cells from the total community have been shown to be inactive or dormant (Jones and Lennon, 2010). If one were to sample plant communities in an analogous way to bacterial systems, one would measure the diversity of all the above- and below-ground plant biomass including seeds, pollen, and detrital biomass. In this context, it is interesting to reflect on the richness in absence of function (*i.e.*, x-intercept) of the observed BEF relationship which is 295 OTUs (**Figure 3.2C**). This could be interpreted as a baseline level of 295 inactive (either dead, dormant, or not utilizing the leucine substrate used in our methods) bacterial OTUs and in the case of particulate material, environmental DNA adhered to the substrate, in the community.

The value of 295 OTUs value represents 35-85% of the total OTU richness of particle-associated communities and may obscure the *actual* diversity (and BEF relationship) of the bacterial community (Carini et al., 2016). In other words, the BEF relationships are driven by the fraction of cells that are active relative to dormant. While heterotrophic productivity (**Figure SI 3.6 B & C**) of particle-associated and free-living communities are correlated, the fact that a BEF relationship was only observed for particle-associated communities suggests that differences may exist in the relationship between total diversity and the fraction of dormant cells in free-living and particle associated microhabitats. However, we did not measure the extent of dormancy in each sample and can only make the above inferences based on the model in **Figure 3.2C**.

In conclusion, we show that increased bacterial diversity, especially when measured by the inverse Simpson's index, leads to increased total and per-capita bacterial heterotrophic production in particle-associated but not in free-living communities. As such, we extend the validity of principles of the impact of microhabitat on BEF relationships from Eukarya to Bacteria, contributing to current efforts to integrate ecological theories into the field of microbiology (Barberán et al., 2014). Additionally, we show that communities with low phylogenetic diversity have higher per-capita heterotrophic production rates, which we hypothesize to be related to genome evolutionary patterns specific to bacteria that result in the dependence on metabolic hand-offs. Differences between Bacteria and Eukarya in patterns of genome evolution and its ecological consequences, as well as in how active and dormant fractions of the community are measured, need to be considered when trying to integrate BEF studies across all domains of life.

Acknowledgements

This work was supported by the National Science Foundation Graduate Research Fellowship Grant No. DGE 1256260 (MLS), the University of Michigan Office for Research MCubed program (VJD), the American Society of Microbiology-Undergraduate Research Fellowship, the University of Michigan Honors Summer Fellowship, and the Beckman Scholars Program (EC). RP was supported by Ghent University (BOFDOC2015000601) and a Sofina Gustave-Boël grant from the Belgian American Educational Foundation. We are grateful to the crew of the R/V

W.G. Jackson and the Grand Valley State University Robert B. Annis Water Resources Institute science staff, and the generous help we received in the field from Amelia Waters and Daniel S.W. Katz. Thank you to Kyle Buffin and Amadeus Twu for help with DNA extractions. Finally, we thank Deborah Goldberg, George Kling, and members of the Denef, Dick, and Duhaime laboratories for their comments on the manuscript.

References

Anantharaman, K. et al. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Communications 7:13219.

Barberán, A., E. O. Casamayor, and N. Fierer. 2014. The microbial contribution to macroecology. Frontiers in Microbiology 5:1–8.

Bell, T., Newman, J.A., Silverman, B.W., Turner, S.L. and A.K. Lilley. 2005. The contribution of species richness and composition to bacterial services. Nature 436:1157–1160.

Bižić-Ionescu, M., Zeder, M., Ionescu. D, Orlic, S., Fuchs, B.M., Grossart, H.P., and R. Amann. 2014. Comparison of bacterial communities on limnic versus coastal marine particles reveals profound differences in colonization. Environmental Microbiology doi:10.1111/1462–2920.12466.

Blaser, M. J. 2014. Missing Microbes: How the Overuse of Antibiotics Is Fueling Our Modern Plagues. Henry Holt and Company LLC, New York 35:261.

Bushnell B. 2016. BBMap short read aligner. https://sourceforge.net/projects /bbmap/.

Cadotte, M. W., B. J. Cardinale, and T. H. Oakley. 2008. Evolutionary history and the effect of biodiversity on plant productivity. Proceedings of the National Academy of Sciences 105:17012–17017.

Caporaso, J. G. et al. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6:1621–1624.

Cardinale, B. J. 2011. Biodiversity improves water quality through niche partitioning. Nature 472:86–89.

Cardinale, B. J. et al. 2012. Biodiversity loss and its impact on humanity. Nature 489:326–326.

Carini, P., P. J. Marsden, J. W. Leff, E. E. Morgan, M. S. Strickland, and N. Fierer. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nature Microbiology 2:16242.

Cordero, O. X., and M. S. Datta. 2016. ScienceDirect Microbial interactions and community assembly at microscales. Current Opinion in Microbiology 31:227–234.

Cotner, J. B., and B. A. Biddanda. 2002. Small players, large role: Microbial influence on biogeochemical processes in pelagic aquatic ecosystems. Ecosystems 5:105–121.

Crump, B., Baross, J.A., and C.A. Simenstad. 1998. Dominance of particle-attached bacteria in the Columbia River estuary, USA. Aquatic Microbial Ecology 14:7–18.

Datta, M. S., E. Sliwerska, J. Gore, M. F. Polz, and O. X. Cordero. 2016. Microbial interactions lead to rapid micro-scale successions on model marine particles. Nature Communications 7:11965.

Delgado-Baquerizo, M., L. et al. 2016. Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. Journal of Ecology 104:936–946.

del Giorgio, P. A., and J. J. Cole. 1998. Bacterial Growth Efficiency in Natural Aquatic Systems. Annual Review of Ecology and Systematics 29:503–541.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33:1-22.

Fritschie, K. J., B. J. Cardinale, M. A. Alexandrou, and T. H. Oakley. 2014. Evolutionary history and the strength of species interactions: Testing the phylogenetic limiting similarity hypothesis. Ecology 95:1407–1417.

Galand, P. E., I. Salter, and D. Kalenitchenko. 2015. Ecosystem productivity is associated with bacterial phylogenetic distance in surface marine waters. Molecular Ecology 24:5785–5795.

Ganesh, S., D. J. Parris, E. F. DeLong, and F. J. Stewart. 2014. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. The ISME journal 8:187–211.

Griffiths, B. S. et al. 2000. Ecosystem response of pasture soil communities to fumigation-induced microbial diversity reductions: an examination of the biodiversity-ecosystem function relationship. Oikos 90:279–294.

Grossart, H. P. 2010. Ecological consequences of bacterioplankton lifestyles: Changes in concepts are needed. Environmental Microbiology Reports 2:706–714.

Harvey, R.W. and L.Y. Young. 1980. Enumeration of particle-bound and unattached respiring bacteria in the salt marsh environment. Applied and Environmental Microbiology. 40(1):156–160

Hobbie, J. E., R. J. Daley, and S. Jasper. 1977. Use of nuclepore filter counting bacteria by fluorescence microscopy. Applied and Environmental Microbiology 33:1225–1228.

Jiang, L., J. Tan, and Z. Pu. 2010. An Experimental Test of Darwin’s Naturalization Hypothesis. The American Naturalist 175:415–423.

Jones, S. E., and J. T. Lennon. 2010. Dormancy contributes to the maintenance of microbial diversity. Proceedings of the National Academy of Sciences 107:5881–5886.

Kembel, S. W. 2009. Disentangling niche and neutral influences on community assembly: Assessing the performance of community phylogenetic structure tests. Ecology Letters 12:949–960.

Kembel, S.W. et al.. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26:1463-1464.

Kim, M., H. S. Oh, S. C. Park, and J. Chun. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International Journal of Systematic and Evolutionary Microbiology 64:346–351.

Kirchman, D., E. K’nees, and R. Hodson. 1985. Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. Applied and Environmental Microbiology 49:599–607.

Konstantinidis, K. T., and J. M. Tiedje. 2005. Genomic insights that advance the species definition for prokaryotes. Proceedings of the National Academy of Sciences 102:2567–2572.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the miseq illumina sequencing platform. Applied and Environmental Microbiology 79:5112–5120.

Krause, S. et al. 2014. Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. Frontiers in Microbiology 5:1–10.

Langenheder, S., E. Bulling, M.T., Solan, M., and J.I. Prosser. 2010. Bacterial biodiversity-ecosystem functioning relations are modified by environmental complexity. PLoS ONE 5(5):e10834.

Levine, U. Y., T. K. Teal, G. P. Robertson, and T. M. Schmidt. 2011. Agriculture's impact on microbial diversity and associated fluxes of carbon dioxide and methane. The ISME Journal 5:1683–1691.

Lilja, E. E., and D. R. Johnson. 2016. Segregating metabolic processes into different microbial cells accelerates the consumption of inhibitory substrates. The ISME Journal 10:1–11.

Magurran, A. E. 2004. Chapter four: An index of diversity... in Measuring Biological Diversity, Wiley-Blackwell, Hoboken, NJ.

McMurdie, P. J., and S. Holmes. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217.

Narwani, A. et al. 2017. Ecological interactions and coexistence are predicted by gene expression similarity in freshwater green algae. Journal of Ecology 105:580–591.

Oksanen, A. J. et al. 2015. vegan: Community Ecology Package. R package version 2.3-0.

Peter, H., S. Beier, S. Bertilsson, E. S. Lindström, S. Langenheder, and L. J. Tranvik. 2011. Function-specific response to depletion of microbial diversity. The ISME Journal 5:351–361.

Philippot, L. et al. 2013. Loss in microbial diversity affects nitrogen cycling in soil. The ISME Journal 7:1609–1619.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE 5.

Quast, C. et al. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41:590–596.

R Core Team (2017) R: A Language and Environment for Statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https:// www.R-project.org/.

Replansky T. and G. Bell. 2009. The relationship between environmental complexity, species diversity and productivity in a natural reconstructed yeast community. Oikos 118:233–239.

Riemann, L., Steward, G.F., and F. Azam. 2000. Dynamics of bacterial community composition and activity during a mesocosm diatom bloom. Applied and Environmental Microbiology 66(2):578–587

Roger, F., S. Bertilsson, S. Langenheder, O. A. Osman, and L. Gamfeldt. 2016. Effects of multiple dimensions of bacterial diversity on functioning, stability and multifunctionality. Ecology 97:2716–2728.

Rohwer, R. R., J. J. Hamilton, R. J. Newton, and K. D. McMahon. 2017. TaxAss: Leveraging Custom Databases Achieves Fine-Scale Taxonomic Resolution. bioRxiv:214288.

Russel, J., H. L. Røder, J. S. Madsen, M. Burmølle, and S. J. Sørensen. 2017. Antagonism correlates with metabolic similarity in diverse bacteria. Proceedings of the National Academy of Sciences:201706016.

Salles, J. F. et al. 2009. Community niche predicts the functioning of denitrifying bacterial assemblages. Ecology, 90:3324–3332.

Schloss, P. D. et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75:7537–7541.

Schmidt, M.L., J.D. White, and V.J. Denef. 2016. Phylogenetic conservation of freshwater lake habitat preference varies between abundant bacterioplankton phyla. Environmental Microbiology 18(4):1212–1226.

Shade, A., S. E. Jones, and K. D. McMahon. 2008. The influence of habitat heterogeneity on freshwater bacterial community composition and dynamics. Environmental Microbiology 10:1057–1067.

Simon, M., and F. Azam. 1989. Protein content and protein synthesis rates of planktonic marine bacteria. Marine Ecology Progress Series 51:201–213.

Simon, M., H. P. Grossart, B. Schweitzer, and H. Ploug. 2002. Microbial ecology of organic aggregates in aquatic ecosystems. Aquatic Microbial Ecology 28:175–211.

Singh, B. K. et al. 2014. Loss of microbial diversity in soils is coincident with reductions in some specialized functions. Environmental Microbiology 16:2408–2420.

Steinman, A. D., M. Ogdahl, R. Rediske, C. R. Ruetz III, B. a Biddanda, and L. Nemeth. 2008. Current status and trends in Muskegon Lake, Michigan. Journal of Great Lakes Research 34:169–188.

Stewart, E. J. 2012. Growing unculturable bacteria. Journal of Bacteriology 194:4151–4160.

Tan, J., Z. Pu, W. A. Ryberg, and L. Jiang. 2012. Species phylogenetic relatedness, priority effects, and ecosystem functioning. Ecology 93:1164–1172.

Thomas, C. D. et al. 2004. Extinction risk from climate change. Nature 427:145–8.

Tilman, D., F. Isbell, and J. M. Cowles. 2014. Biodiversity and Ecosystem Functioning. Annual Review of Ecology, Evolution, and Systematics 45:471–493.

Tuomisto, H. 2012. An updated consumer’s guide to evenness and related indices. Oikos. 121:1203–1218.

Tylianakis, J. M. et al. 2008. Resource heterogeneity moderates the biodiversity-function relationship in real world ecosystems. PLoS Biology 6:0947–0956.

Venail, P. A., A. Narwani, K. Fritschie, M. A. Alexandrou, T. H. Oakley, and B. J. Cardinale. 2014. The influence of phylogenetic relatedness on species interactions among freshwater green algae in a mesocosm experiment. Journal of Ecology 102:1288–1299.

Venail, P. A., and M. J. Vives. 2013. Phylogenetic distance and species richness interactively affect the productivity of bacterial communities. Ecology 94:2529–2536.

Violle, C., D. R. Nemergut, Z. Pu, and L. Jiang. 2011. Phylogenetic limiting similarity and competitive exclusion. Ecology Letters 14:782–787.

Wake, D. B., and V. T. Vredenburg. 2008. Are we in the midst of the sixth mass extinction? A view from the world of amphibians. Proceedings of the National Academy of Sciences 105:11466–11473.

Wertz, S., V. et al. 2006. Maintenance of soil functioning following erosion of microbial diversity. Environmental Microbiology 8:2162–2169.

Zeppilli, D., A. Pusceddu, F. Trincardi, and R. Danovaro. 2016. Seafloor heterogeneity influences the biodiversity–ecosystem functioning relationships in the deep sea. Scientific Reports 6:26352.

Zhou, J., S. et al. 2008. Spatial scaling of functional gene diversity across various microbial taxa. Proceedings of the National Academy of Sciences 105:7768–7773.

Figure 3.1. Bacterial counts, community-wide and per-capita heterotrophic production differ between microhabitats.

Particle-associated and free-living samples were taken from four stations within Muskegon Lake during 2015 in May, July, and September. **(A)** Free-living bacteria were an order of magnitude (10^6 cells/mL) more abundant compared to particle-associated bacteria. **(B)** Free-living bacteria were more heterotrophically productive compared to particle-associated bacteria. **(C)** Particle-associated bacteria were disproportionately heterotrophically productive per cell compared to free-living bacteria.

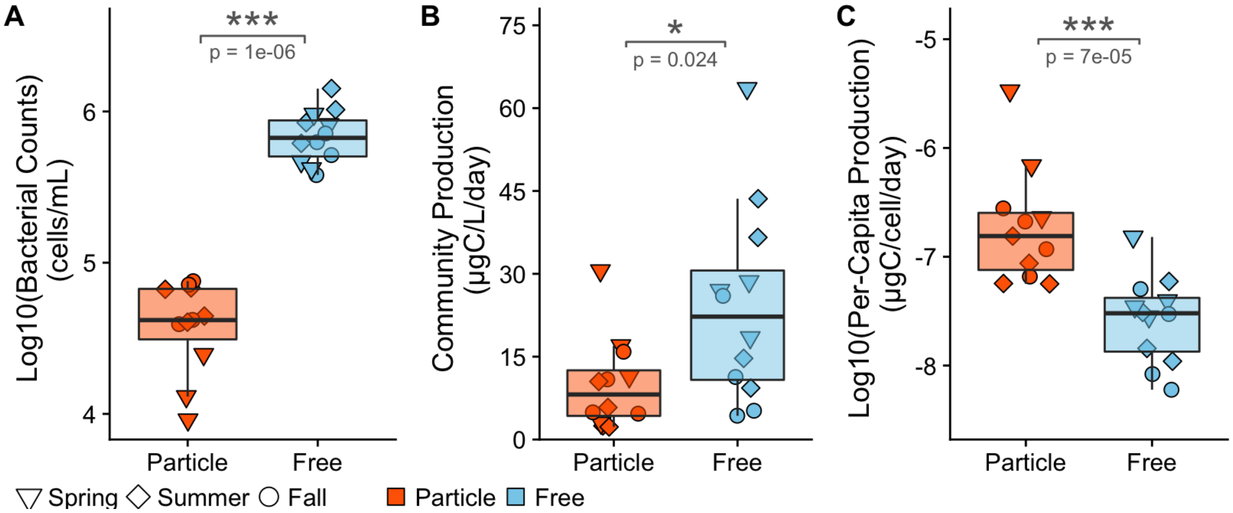


Figure 3.2. Richness and Inverse Simpson correlate with heterotrophic productivity.

Top panel: Differences in **(A)** the observed richness and **(B)** the inverse Simpson diversity metrics between particle-associated (orange) and free-living (blue) habitats. **Middle panel:** Biodiversity and community-wide heterotrophic production (µgC/L/day) relationships. The y-axis between **(C)** and **(D)** is the same, however, the x-axis represents **(C)** richness and **(D)** Inverse Simpson. **Bottom panel:** Biodiversity and $\log_{10}$(per-capita heterotrophic production) (µgC/cell/day) relationships. The y-axis between **(E)** and **(F)** is the same, however, the x-axis represents **(E)** richness and **(F)** Inverse Simpson's index. Solid lines represent ordinary least squares models for the free-living (blue) and particle associated (orange) communities. All R^2 values represent the adjusted R^2 from an ordinary least squares model.

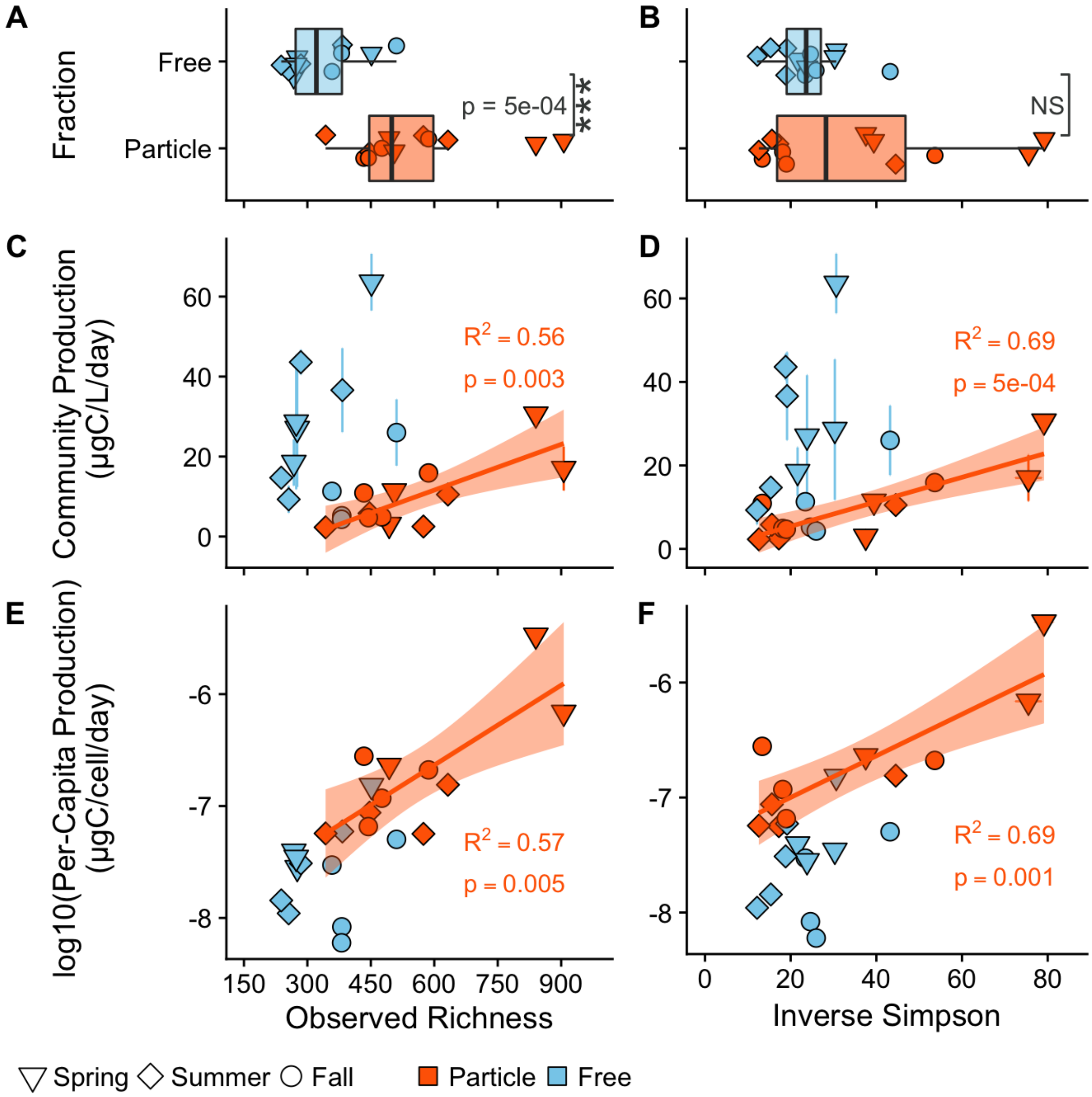
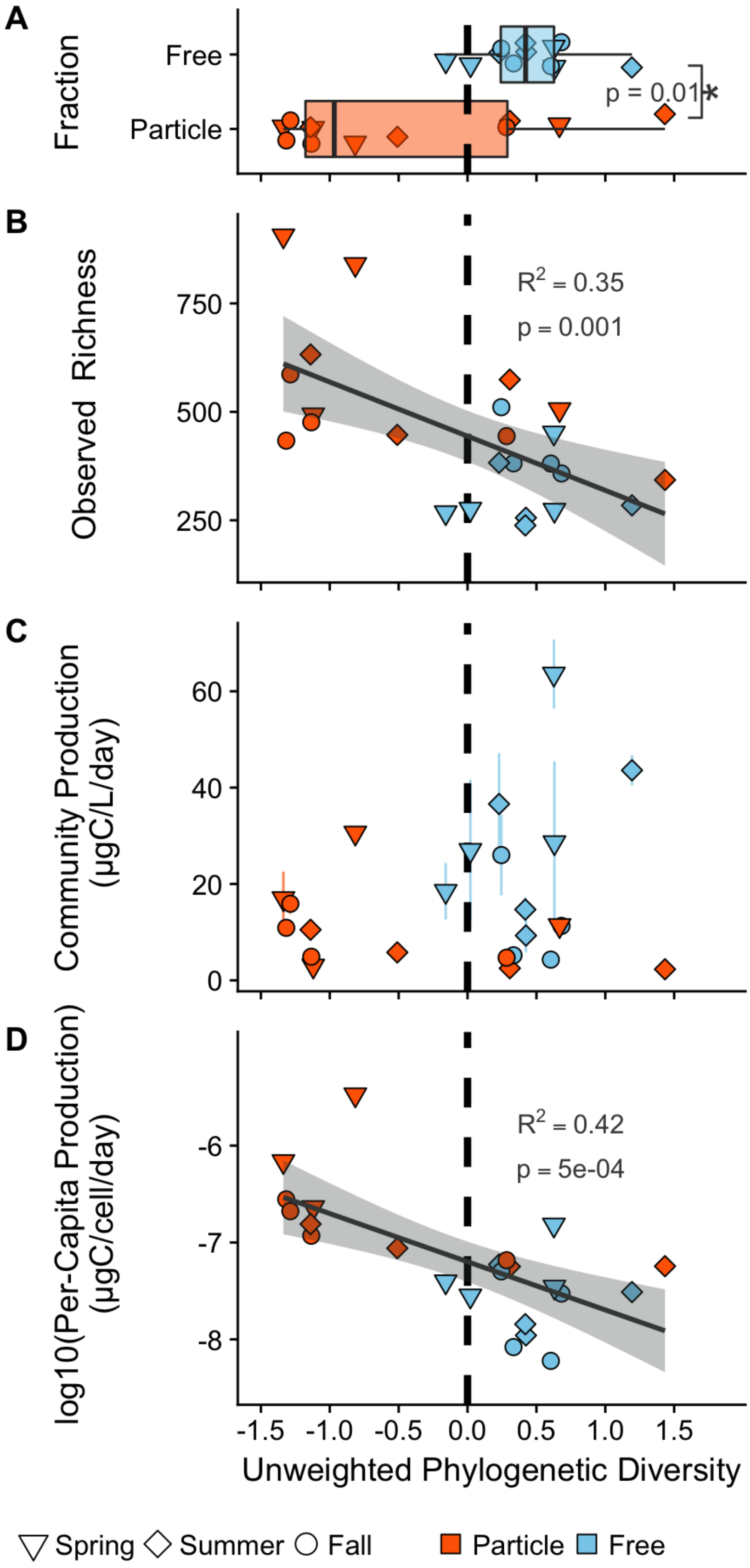


Figure 3.3. The relationship between heterotrophic productivity and unweighted phylogenetic diversity.

(SES$_{MPD}$; ses.mpd function in *picante* with null.model = "independentswap"). Positive phylogenetic diversity values represent communities that are phylogenetically diverse (*i.e.*, overdispersed) while negative phylogenetic diversity values represent communities that are phylogenetically less diverse (*i.e.*, clustered) compared to a null community with equal species richness. **(A)** Phylogenetic diversity was higher in free-living communities compared to particle-associated communities. **(B)** Negative relationship between observed richness and phylogenetic diversity. **(C)** Absence of phylogenetic diversity and community bulk heterotrophic production (µgC/L/day) relationships. **(D)** Negative phylogenetic diversity and per-capita heterotrophic production (µgC/cell/day) relationship. Linear models in figure **B** and **D** represent trends over all samples.



Supporting Information 3.A. Supplemental Methods for Chapter III

Supplemental Methods

Map of Muskegon Lake

The Muskegon Lake map (Figure S1) was created by using the pre-existing 2006 National Hydrography Dataset (NHD) GIS feature data for the Muskegon Lake shoreline was adjusted to an updated 2008 format using ESRI™ ArcGIS software and high resolution (6” pixel) leaf-off aerial orthophotography. Historic mean water level for Lake Michigan was then used to estimate Muskegon Lake water level for April 2008 at 176.4 meters. This water level was used as the base elevation and adjustment point for correcting all relevant lake bathymetric data, taken from the recently, published February 2008 NOAA electronic bathymetric chart for Muskegon Lake. The corrected GIS shoreline boundary and the supplemental NOAA bathymetric point data were used to generate a new bathymetric grid (raster) feature for Muskegon Lake from which, contours were created at 2m depth intervals. This bathymetric map was then laid under a Google Earth image of Muskegon Lake and the immediately surrounding area.

DNA Extraction

In summary, the filters were first washed with phosphate-buffered saline (pH 7.4) and folded (cell-side in) to minimize cell loss and to remove RNAlater, which inhibits DNA yields. Then the filters were placed in a 2 mL tube with 600 µL of buffer RLT plus (Qiagen) and incubated for 90 min at room temperature using a vortex on medium setting (5 of 10). After incubation, tubes were vortexed on high for 10 min. The lysate was transferred to a QiaShredder column (Qiagen), 300 µL of 100% ethanol was added to the lysate and then transferred to a DNA column (DNeasy Blood and Tissue Kit, Qiagen) and washed with 350 µL of buffer AW1 (Qiagen). Next, 80 µL of proteinase K solution was added to the DNA column and incubated at room temperature for 5 min. The DNA column was washed with buffer AW1 and buffer AW2. DNA was eluted using 2 × 30 µL elution buffer (buffer EB, Qiagen) into two separate fresh 1.5 mL centrifuge tubes for temporary storage at 4°C until processed for sequencing or in −80°C freezer for sample archiving.

Sequence Processing

We analyzed the sequence data using MOTHUR V.1.38.0 (seed = 777; Schloss et al., 2009) based on the MiSeq standard operating procedure accessed on 3 November 2015 and modified with time (see data accessibility and supplemental methods). Briefly, paired-end reads were merged into contigs based on the Phred quality score heuristic with *make.contigs* (Kozich et al., 2013). Contigs were filtered based on ambiguous bases, more than 8 homopolymers, a length outside of 240-275. Next, sequences were de-replicated and representative sequences were aligned with the Silva database and those not corresponding to the V4 region were removed. After the sequences were de-replicated and filtered, sequencing errors were removed using *pre.cluster* and chimeras were removed with UCHIME (Edgar et al., 2011). We then clustered representative sequences into OTUs at 97% similarity using the average neighbor algorithm (*cluster.split* command) and assigned the taxonomy of the OTUs using the Wang method implemented in the Ribosomal Database Project classifier (*classify.seqs* command).

Sensitivity Analysis of Rare Taxa

Due to the long tails within microbial rank-abundance curves (*i.e.*, many rare taxa), we performed a sensitivity analysis to see the impact of rare taxa on the BEF relationships we observed (Figure S6). We removed OTUs that had a count of 1, 5, 10, 20, 30, 60, 90, 150, 225, and 300 sequences throughout the entire dataset and then checked the relationship with diversity versus community-wide and per-capita heterotrophic production (Figure S6). For example, removing 10-tons will be removing any OTUs that have a count of less than 10 sequences throughout the entire dataset. All code is for this supplementary analysis is available at: https://deneflab.github.io/Diversity_Productivity/analysis/OTU_Removal_Analysis.html

Calculating Phylogenetic Diversity

Instead of using Faith's phylogenetic diversity (Faith, 1992), we used the mean pairwise phylogenetic distance (or MPD), which measures the average phylogenetic distance between all combinations of two taxa pulled from the observed community. The MPD of the observed community was compared to the MPD of a null community with the same OTU richness and abundances randomized across OTUs pulled from all of the samples in the dataset. The difference between the observed MPD metric and randomized MPD metric were compared to each other while dividing by the standard deviation of the null community, known as the standardized effect size or SES (Gurevitch et al., 1993) of the MPD, or SES_{MPD}, (*ses.mpd* function in picante using null.model = “independentswap”; equation 1). We calculated both the abundance-unweighted and -weighted metrics of SES_{MPD} (Webb et al., 2002). The SES_{MPD} of each local community was calculated as follows:

$$SES_{MPD} = \frac{MPD_{Observed} - Mean(MPD_{Randomized})}{SD(MPD_{Randomized})} \qquad (1)\ \textit{(Kembel, 2009)}$$

Specifically, this model tests whether the mean SES_{MPD} across samples differs from the null community (randomly generated from all samples to generate a randomized regional species pool) with an SES value of zero. Therefore, values higher than zero indicate phylogenetic evenness or overdispersion (higher phylogenetic diversity) while values less than zero indicate phylogenetic clustering (lower diversity) or that species are more closely related than expected according to the null community (Kembel, 2009).

Standard Statistical Testing

Data analysis was performed using R version 3.4.2 (R Core Team 2017), specifically with the phyloseq (McMurdie and Holmes, 2013), *stats* (R Core Team 2017), and *broom* (Robinson, 2017) R packages. All main figures were made using the ggplot2 R package (Wickham, 2009).

To assess a statistical difference in particle-associated and free-living cell abundances, community production rates, per-capita production rates, and biodiversity metrics, a Wilcoxon rank sum test (*wilcox.test* function) was performed. We evaluated whether diversity metrics or environmental variables predicted heterotrophic production rates using ordinary least squares linear regression (*lm* function) and accessed specific variables with *broom::glance()* (*i.e.,* AIC, adjusted R^2). P-values were corrected using the false discovery rate method using the *p.adjust(method = “fdr”)* function in the stats package.

Supplemental References

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince, and R. Knight. 2011. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194–2200.

Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10.

Fox, J., and S. Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gurevitch, J., L. L. Morrow, A. Wallace, and J. S. Walsh. 1992. A Meta-Analysis of Competition in Field Experiments. The American Naturalist 140:539–572.

Kembel, S. W. 2009. Disentangling niche and neutral influences on community assembly: Assessing the performance of community phylogenetic structure tests. Ecology Letters 12:949–960.

Kozich, J. J., S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and Environmental Microbiology 79:5112–5120.

McMurdie, P. J., and S. Holmes. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217.

R Core Team (2017) R: A Language and Environment for Statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https:// www.R-project.org/.

Robinson, R. 2017. broom: Convert Statistical Analysis Objects into Tidy Data Frames. R package version 0.4.2. https://CRAN.R-project.org/package=broom

Schloss, P. D. et al. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75:7537–7541.

Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and Community Ecology. Annual Review of Ecology and Systematics 33:475–505.

Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.

Table SI 3.1. Results from linear regression for predicting community-wide heterotrophic production.
All diversity and environmental variables with a FDR-corrected p-value of less than 0.05, sorted by AIC.

FRACTION	INDEPENDENT VARIABLE	AIC	ADJUSTED R SQUARED	FDR-CORRECTED P-VALUE
PARTICLE	Inverse Simpson	74.34	0.69	0.0019
PARTICLE	Richness	78.68	0.56	0.0063
PARTICLE	Shannon Entropy	79.61	0.52	0.0063
PARTICLE	Simpsons Evenness	80.85	0.47	0.0082
ALL SAMPLES	pH	192.16	0.35	0.0329

Table SI 3.2. Results from regressions of per-capita heterotrophic production.
Ordinary least squares linear regressions predicting **per-capita heterotrophic production** from all diversity and environmental variables with a FDR-corrected p-value of less than 0.05, sorted by AIC.

FRACTION	INDEPENDENT VARIABLE	AIC	ADJUSTED R SQUARED	FDR-CORRECTED P-VALUE
FREE	pH	-2.39	0.78	0.0025
PARTICLE	Inverse Simpson	8.29	0.69	0.0038
PARTICLE	Richness	11.95	0.57	0.0075
PARTICLE	Shannon Entropy	12.43	0.55	0.0075
PARTICLE	Simpsons Evenness	13.65	0.49	0.0096
ALL SAMPLES	Richness	24.72	0.63	0
ALL SAMPLES	Shannon Entropy	30.87	0.52	0.0001
ALL SAMPLES	Inverse Simpson	33.42	0.46	0.0003
ALL SAMPLES	Unweighted PD	35.21	0.42	0.0124

Figure SI 3.4. Bathymetric map of Muskegon Lake.

Bathymetric map of Muskegon Lake with locations of the Muskegon Lake Observatory Buoy (MLO) and the four sampling locations used in this study. Bathymetric iso-lines represent approximately 2m changes in water depth.

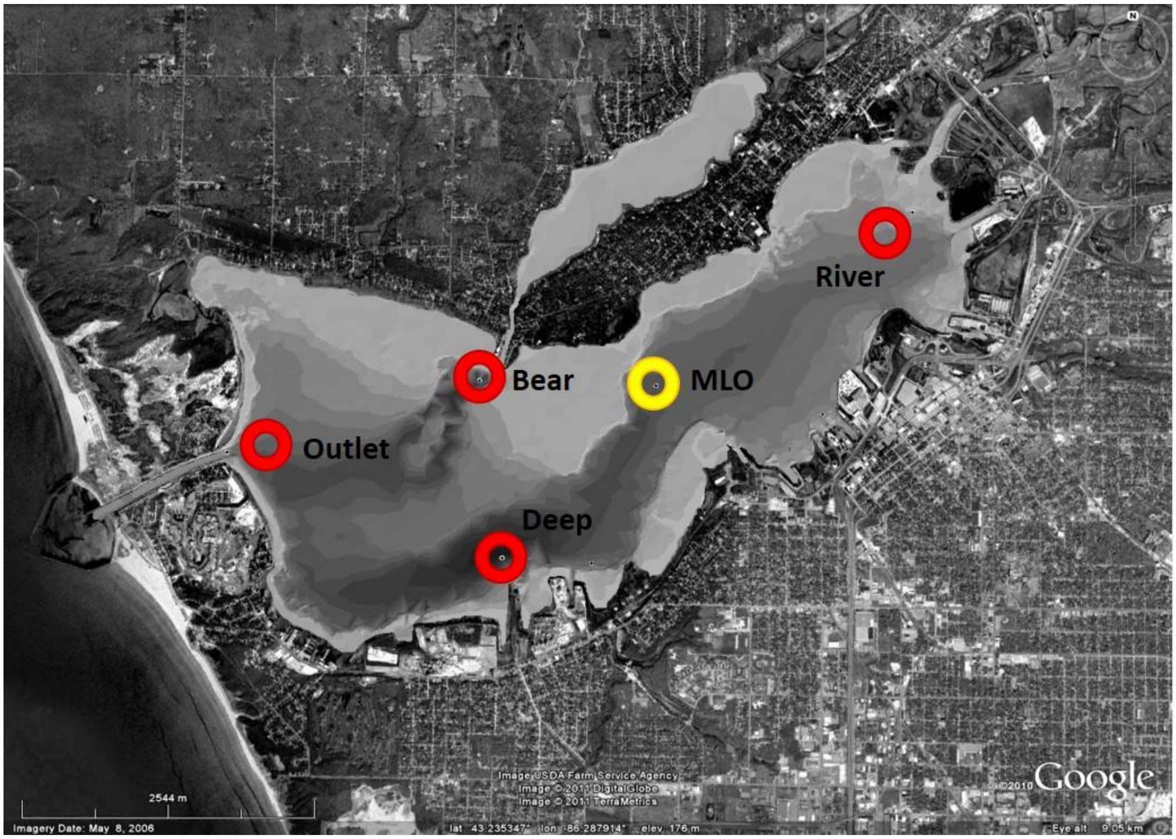

N
River
Bear
MLO
Outlet
Deep
2544 m
Imagery Date: May 8, 2006
Image USDA Farm Service Agency
Image © 2011 DigitalGlobe
Image © 2011 TerraMetrics
lat 43.235347° lon -86.287914° elev 176 m
©2010 Google
Eye alt 9.05 km

Figure SI 3.5. Faith's phylogenetic correlates with species richness.

Faith's phylogenetic diversity is highly correlated with species richness and thus, it is important to compare standardized effect sizes that are measured when actual samples are compared to a randomized null model.

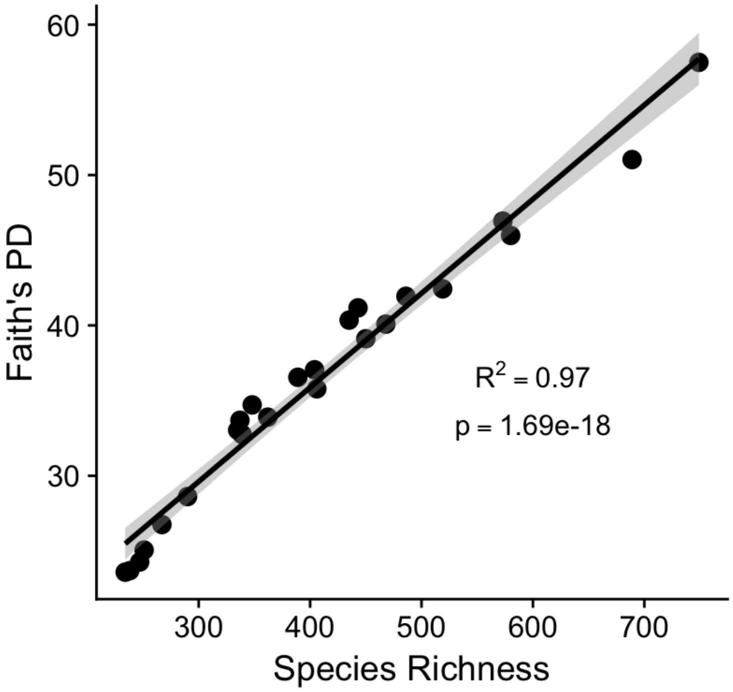


Figure SI 3.6. Correlation between corresponding particle-associated and free-living samples.

(A) cell abundances in $\log_{10}$(cells/mL), **(B)** community-wide heterotrophic production (µg C/L/day), **(C)** $\log_{10}$**(**per-capita production) in µgC/cell/day.

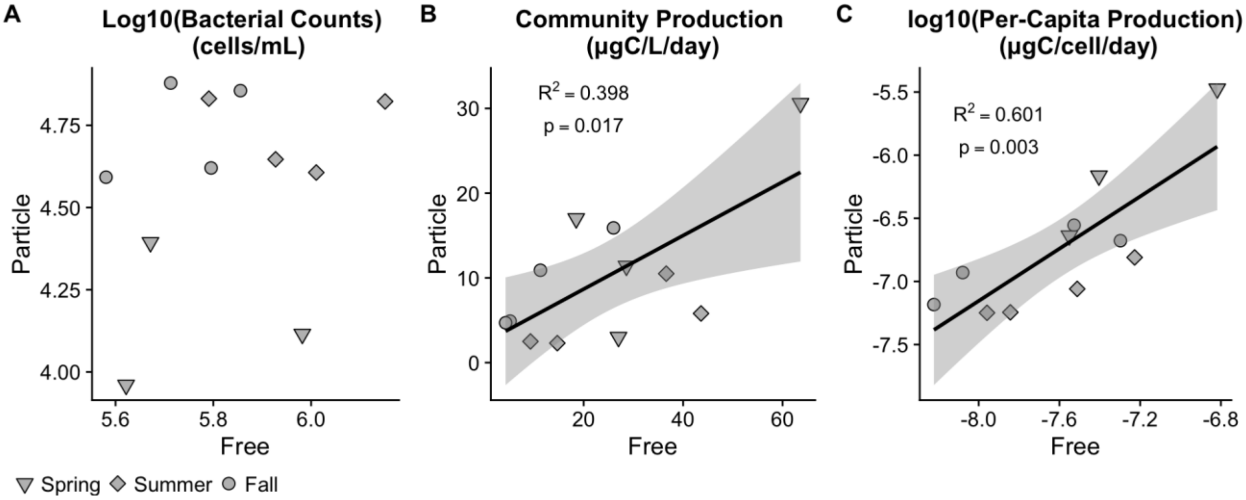


Figure SI 3.7. Correlation between corresponding particle-associated and free-living samples.

(A) cell abundances in $\log_{10}$(cells/mL), **(B)** community-wide heterotrophic production (μg C/L/day), **(C)** $\log_{10}$**(**per-capita production) in μgC/cell/day.

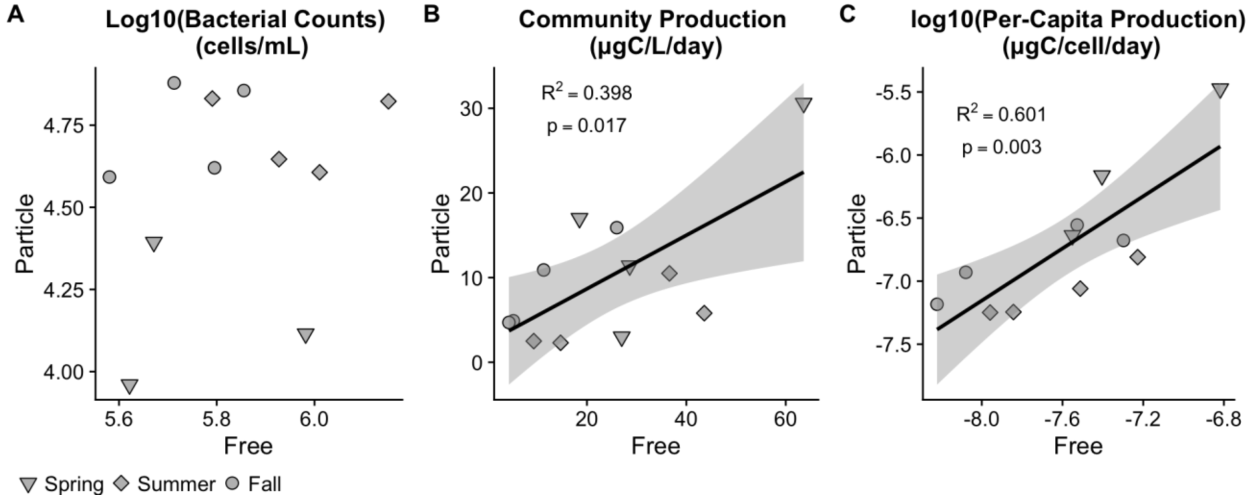


Figure SI 3.8. Diversity analysis with Shannon entropy and Simpson's Evenness.

Top Panel: (A) Shannon entropy and **(B)** Simpson's evenness diversity metric between free-living (blue) and particle-associated (orange) habitats. **Middle panel:** The relationship between bacterial diversity and community-wide heterotrophic production (µgC/L/day). The y-axis is the same however, the x-axis represents **(C)** Shannon entropy and **(D)** Simpson's evenness. **Bottom panel:** The relationship between bacterial diversity and $\log_{10}$(heterotrophic production/cell) (µgC/cell/day). The y-axis is the same however, the x-axis represents **(E)** Shannon entropy and **(F)** Simpson's evenness. R^2 and p-values represented in the figures are outcomes of an ordinary least squares regression for the particle associated (orange) samples.

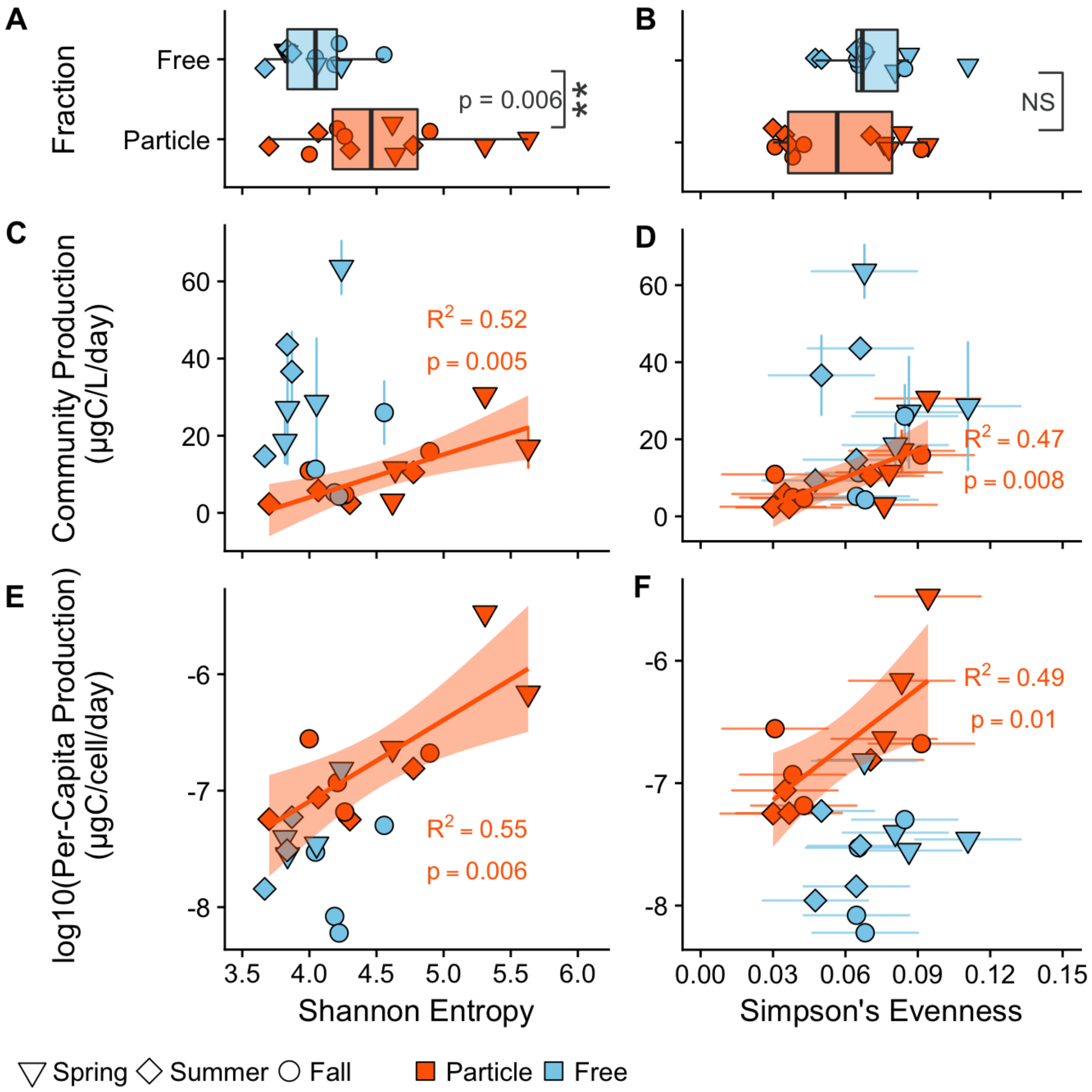


Figure SI 3.9. The six diversity metrics assessed in this study across station and seasons.

Top: Particle-associated (orange) and free-living (blue) diversity values across the estuarine gradient in Muskegon Lake by station (from west to east). All diversity metrics calculated in this study are included. **Bottom:** Diversity values by season.

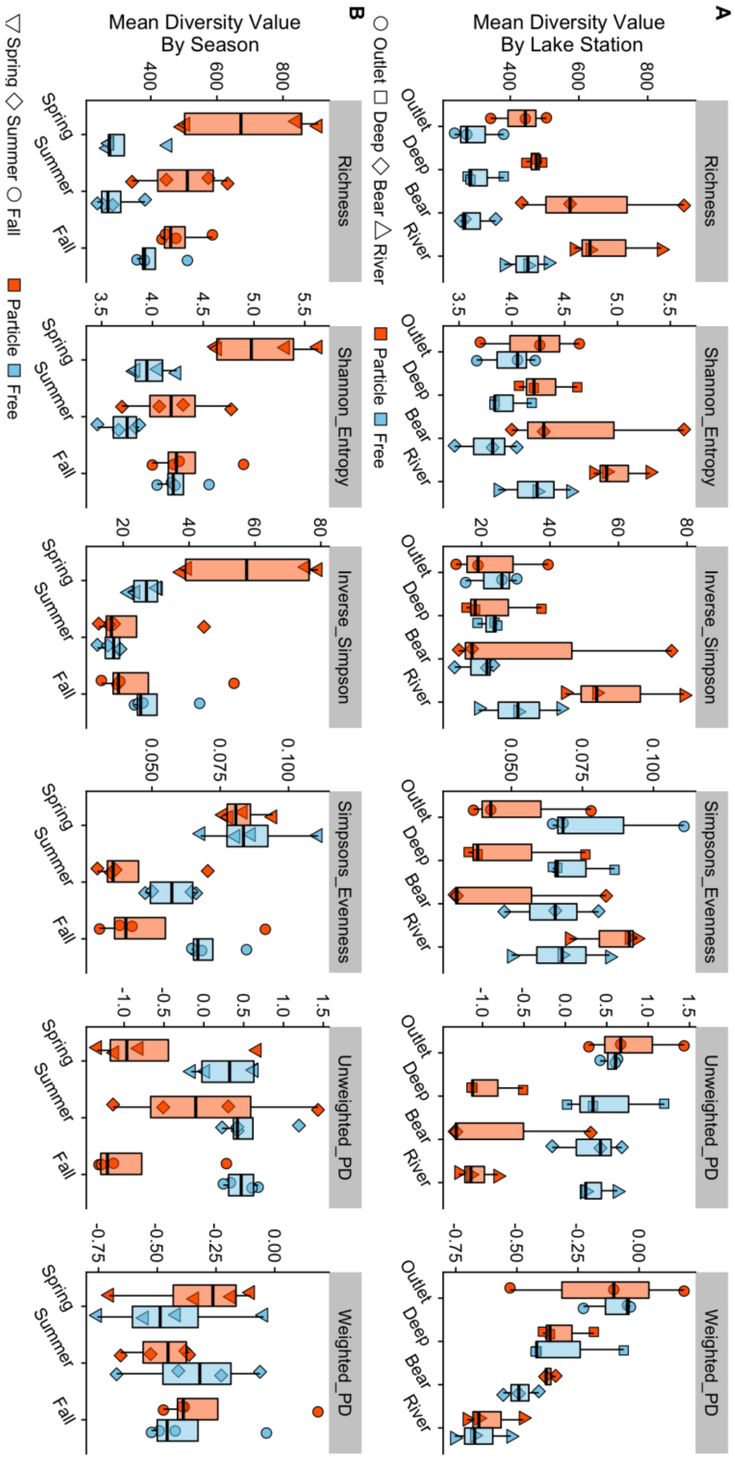


Figure SI 3.10. Sensitivity analysis of rare taxa on biodiversity-ecosystem function relationships.

OTU removal analysis of particle-associated communities of singletons, doubletons, up to 300-tons. **(A)** and **(B)** represent the diversity-productivity patterns of observed richness while **(C)** and **(D)** represent the diversity-productivity relationship with the inverse Simpson's index. Plots **(A)** and **(C)** are community-wide heterotrophic production, while plots **(B)** and **(D)** are the $\log_{10}$(Per-capita heterotrophic production).

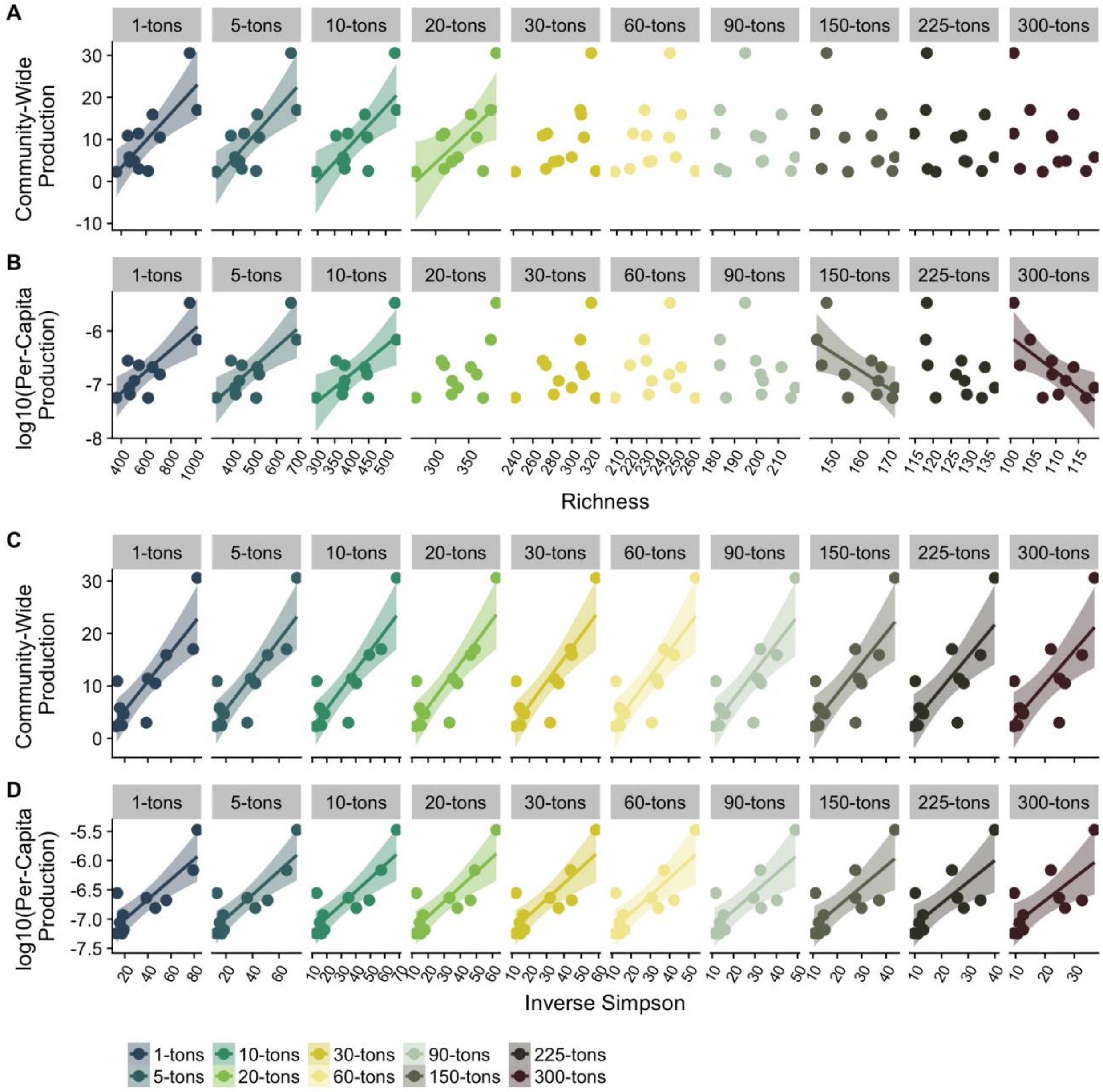


Figure SI 3.11. The relationship between bacterial diversity and $\log_{10}$(per-capita heterotrophic production) (µgC/cell/day) across all samples in the dataset.

Diversity metrics on the x-axis are: **(A)** Observed richness, **(B)** Shannon entropy, **(C)** inverse Simpson's and **(D)** Simpson's evenness. Adjusted R^2 and p-values represent were calculated from an ordinary least squares linear regression model.

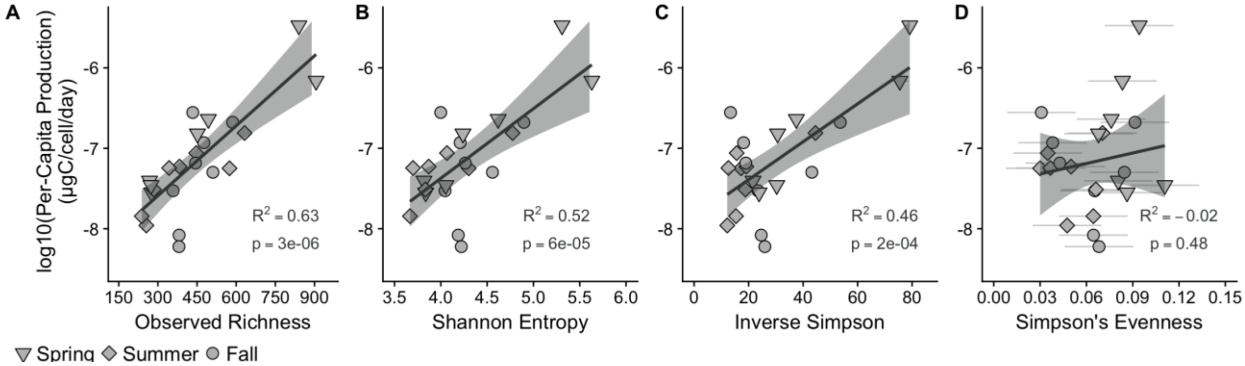


Figure SI 3.12. Abundance-weighted phylogenetic diversity analysis.

(A) between particle-associated and free-living communities. **(B)** Absence of a correlation between abundance-weighted phylogenetic diversity and inverse Simpson's index, **(C)** community-wide heterotrophic production, and **(D)** per-capita production.

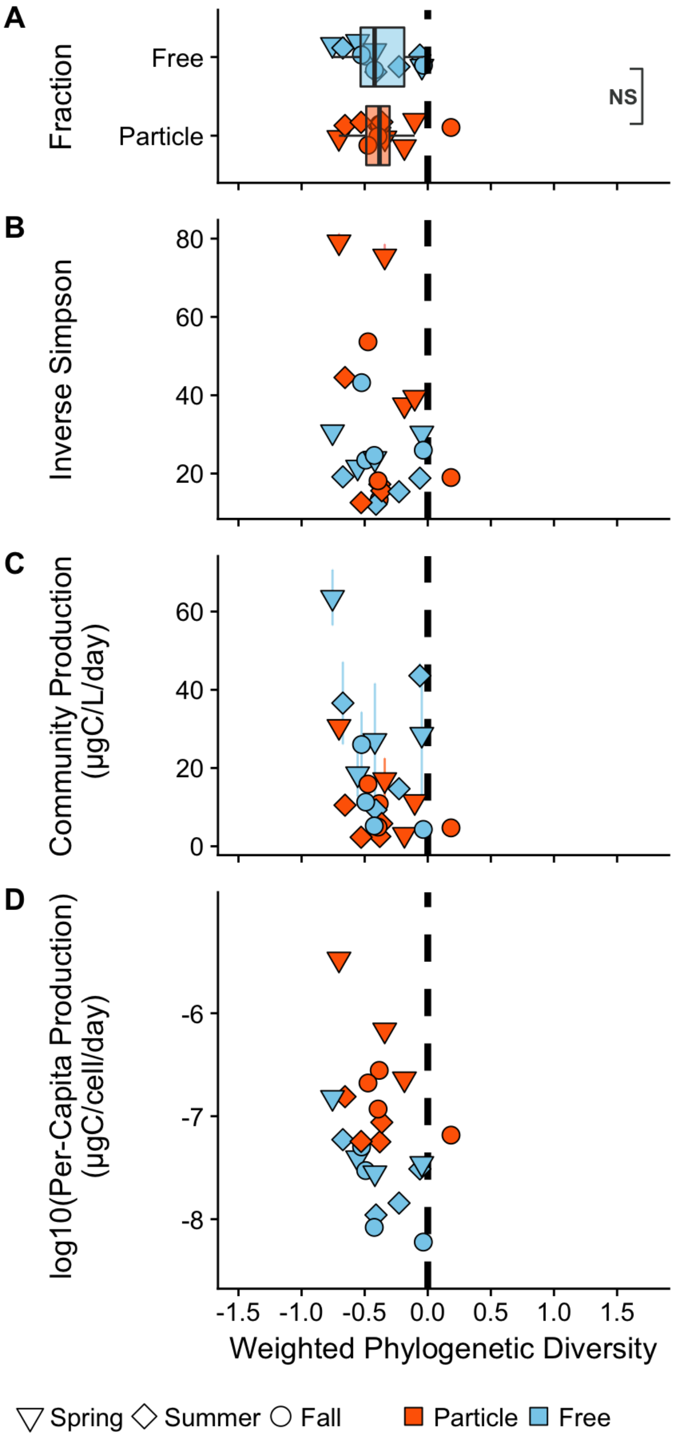


Figure SI 3.13. The relationship between randomized richness and standardized effect size.

The richness values were the same value as actual samples, however, the OTUs across the samples were randomized across the gamma diversity of the entire dataset.

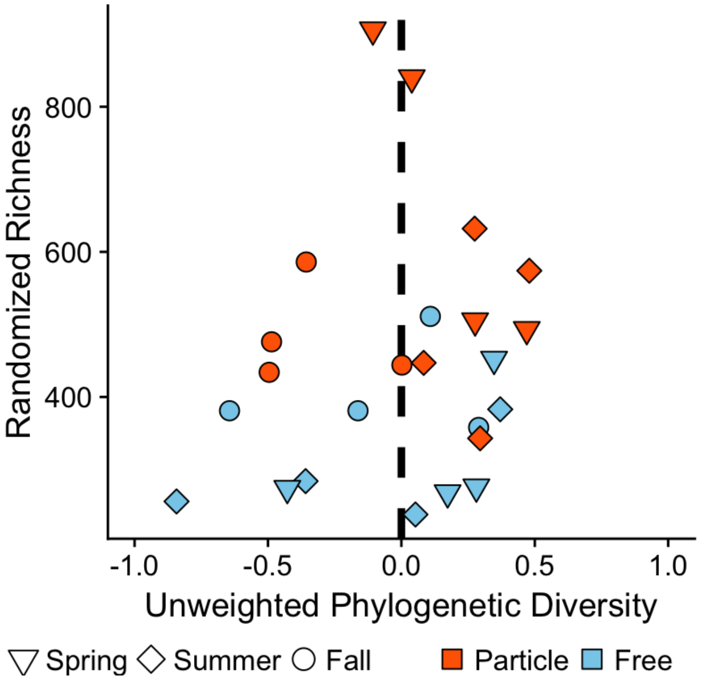
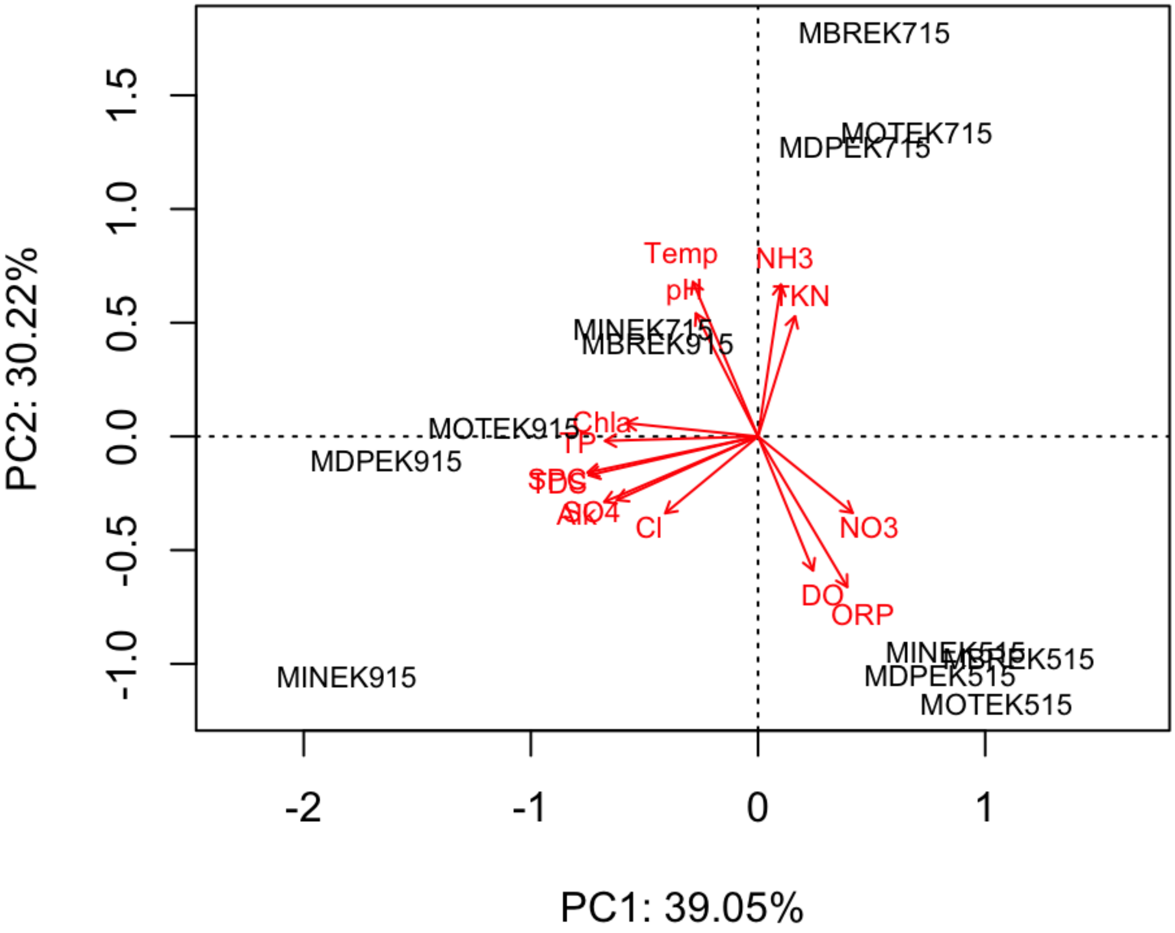


Figure SI 3.14. Principal components analysis of the Euclidean distances of the environmental variables with a biplot (vectors) of the environmental drivers of stations in ordination space.



Chapter IV:

The genomic basis of three aquatic bacterial lifestyles

Introduction

Particles play an important role in shaping aquatic bacterial community composition. First, aquatic particle-associated bacterial communities are generally taxonomically distinct from free-living bacteria (Smith et al. 2013b, Mohit et al. 2014, Simon et al. 2014, Bižić-Ionescu et al. 2015, Satinsky et al. 2016, Schmidt et al. 2016, 2017). Second, particles can shape aquatic microbial communities by acting as connective conduits between surface and deep-water microbial communities (Mestre et al. 2018). Finally, particles also provide microheterogeneity in the water column that helps sustain high levels of taxonomic and metabolic diversity in aquatic microbial communities (Hunt et al. 2008a, Grossart 2010, Stocker 2012, Salcher 2014).

These differences in composition appear to have functional consequences. Cells attached to particulate matter have been shown to be disproportionately active, especially during the nighttime (Ghiglione et al. 2007), showing higher rates of proteolytic activities (Lemarchand et al. 2006) and a larger spectrum of polysaccharide hydrolase activities (Balmonte et al. 2018). Also, we have recently show that particles, and not their free-living counterparts, display positive biodiversity-ecosystem function relationships (Schmidt et al. 2017). Metagenomic analysis of particle-attached bacterial communities has also indicated a higher potential for gene exchange than in free-living bacteria, thus highlighting the potential role of particles in mediating bacterial genome evolution. For example, a study of bacteria living on the surface of green macroalgae possess more genes for DNA transfer than planktonic organisms (Burke et al. 2011), and the particle-associated bacterial communities in an oxygen minimum zone have shown a similar trend (Ganesh et al. 2014b). This latter study, as well as a study conducted on the California coast (Zeigler Allen et al. 2012), also found that genes originating from viruses were more abundant in the particle-associated fractions. Therefore, it is likely that bacteria attached to particulate matter have unique genomic signatures compared with those living sympatrically as planktonic cells in the water.

Recent research, particularly in marine systems, has found that the predominant free-living surface water bacterioplankton have streamlined genomes (Lauro et al. 2009, Swan et al. 2013). These “streamlined” organisms are categorized as living an oligotrophic lifestyle with small genomes, fewer gene duplications, lower GC content, and fewer non-coding regions (García-Fernandez et al. 2004, Swan et al. 2013, Giovannoni et al. 2014). In contrast, particle-associated bacteria are likely to be copiotrophic (Lauro et al. 2009, Zeigler Allen et al. 2012, Allen et al. 2013, Polz and Cordero 2016), although empirical data supporting this remains scarce (Zeigler Allen et al. 2012). Copiotrophic bacteria grow optimally at high nutrient concentrations, have larger cell and genome sizes, and higher maximum growth rates (Lauro et al. 2009). In addition, the increased prevalence of transposable elements and phage integration genes in particle-associated bacterial communities likely has important implications for the genomes of particle-associated bacteria (Shapiro et al. 2012). Therefore, I hypothesize that particle-associated bacteria do not carry signatures of genome streamlining, such as reduced genome size, low GC percentage, and fewer non-coding regions (Swan et al. 2013, Giovannoni et al. 2014), while genomes of free-living bacteria do.

The higher cell density and the difficulty in accessing nutrients in particles may lead to particle-associated bacteria relying on each other to obtain nutrients. For example, particle-associated bacteria mutually benefit when they excrete extracellular enzymes, such as chitinases and collagenases, a known trait of copiotrophs (Lauro et al. 2009). Since the newly dissolved nutrients liberated from the enzymes randomly dissolve into the environment and become a public good, such dependencies among community members could allow for the evolution of different nutrient acquisition strategies (Cordero et al. 2012). On the other hand, free-living bacteria are often non-motile (Newton et al. 2011b) and are less likely to be physically in contact with each other or their metabolites. I hypothesize that free-living bacteria will have less potential for niche complementarity as they may not have as many unique functions that could allow for facilitation between nearby organisms.

Beyond examining the overall differences in genomic architecture, there is a need to identify the genetic traits that determine differences in bacterial composition between particle-associated and free-living habitats and whether these (or linked) genetic traits also determine the potential for distinct functional capacities between habitats. Assessing the genetic traits that allow bacteria to live in specific environments and catalyze certain ecosystem functions also relates to the broader question of how much of an organism's ecology is reflected in, and thus can be predicted from, its genome? Here, particles and the surrounding free-water were used to test two hypotheses regarding the genomic traits that are signatures of specialist and generalist bacteria. First, that particle-associated bacteria do not carry signatures of genome streamlining, such as reduced genome size, low GC percentage, and fewer non-coding regions (Swan et al. 2013, Giovannoni et al. 2014), while genomes of free-living bacteria do. Second, that particle-associated bacteria have more unique metabolic functions per genome compared to free-living bacteria, which leads to a higher likelihood of particle-associated bacteria having complementary niches. This would help explain the higher diversity in particle-associated relative to free-living communities and provide a mechanistic explanation for the presence of a biodiversity-ecosystem function relationship in particle-associated bacterial communities (Schmidt et al. 2017). Specifically, we reconstructed genomes from 16 paired metagenomic surveys taken from two consecutive years and two nearby stations in a freshwater estuarine lake, Muskegon Lake, in Michigan USA. This lake is a hotspot for geochemical transformations and characterized by a stark gradient and broad range in both primary and secondary productivity across space and time (Weinke et al. 2014, Defore et al. 2016). I previously used this system to show BEF relationships particle-associated but not in in free-living communities in **Chapter III**.

Methods

DNA Extraction and metagenome sequencing

The 16 samples used in this study were taken from 2 meters depth from two locations in the western part of Muskegon Lake, the deepest site of the lake and adjacent to the Lake Michigan channel during July and September of 2014 and 2015 as described in Chiang et al. (2018). DNA was extracted following the protocol in McCarthy et al. (2015) as previously described in Chiang et al. (2018). Sequencing of the 16 samples was multiplexed across 5 lanes and was performed at the University of Michigan DNA sequencing core using Illumina HiSeq 2500. Libraries were prepared with a mean insert size of 530 bp.

Summary of quality control of sequencing reads

Original quality evaluation of the raw sequences was performed with FastQC (Andrew, 2010) and summarized per sample with MultiQC (Ewels et al. 2016). Next, sequencing files were deduplicated using FastUniq (Xu et al. 2012) and bbmerge (Bushnell et al. 2017) was used to find adapter contamination in the deduplicated reads. Adapters and all bases to the right were removed based on pair overlap detection and both of the reads were trimmed to the same length with bbduk (also from bbmap). Trimmomatic (version 0.36; Bolger et al. 2014) was used for quality trimming of poor bases at the beginning and ends of reads with a quality score threshold of 20, a sliding window of 4:20, and minimum length requirement of 40 bp. Finally, FastQC (Andrew, 2010) and MultiQC (Ewels et al. 2016) were used as a sanity check to make sure quality control worked.

Assembly and non-competitive coverage estimation

Quality controlled sequences from each sample were assembled separately with MEGAHIT v1.0.6 (Li et al. 2015) with the meta-sensitive parameters of a kmer size ranging from 21 to 99 with a k-step of 10. The coverage of resulting contigs across all quality-controlled sequences from each sample (via non-competitive mapping) was performed by indexing and mapping with BWA (*index* and *mem* functions; Li and Durbin 2009). Coverage files to inform binning were created using the bbmap function pileup.sh (Bushnell et al. 2017).

Binning, and bin refinement

Multiple binners were used to create bins, also known as metagenome-assembled genomes (MAGs), with contigs larger than 2,000 bp from the assembly. Binners that were used to create MAGs include MetaBAT 0.32.4 (Kang et al. 2015), MaxBin 2.2.4 (Wu et al. 2014; with idba 1.1.3; bowtie2 2.1.0; hmmer 3.1b2), CONCOCT (SpeedUP_mp version; https://github.com/BinPro/CONCOCT/tree/SpeedUp_Mp; Alneberg et al. 2014), and VizBin (kmer length of 5; in which 2,500 bp and larger contigs were used; Laczny et al. 2015). A custom python script (that will be available on this project's GitHub repository) was used to parse VizBin's points.txt output with differing levels of data classified as noise (*min_samples* argument; with values of 10, 15, 25, 50, 75, 100) by using a hierarchical density-based spatial clustering of applications with noise (HDBScan; McInnes et al. 2017). Next, bin refinement was performed with DAS Tool (Sieber et al. 2018) with a score threshold of 0.0 in two steps. First, DAS Tool was run with MAGs that varied in levels of noise from HDBScan to select the optimal MAGs from VizBin. Second, the MAGs from the first DAS Tool step with VizBin were used in addition to the output from CONCOCT, MaxBin, and MetaBAT.

Dereplication and quality control of Metagenome Assembled Genomes (MAGs)

The MAGs from the second DAS Tool run (total of 2,774 MAGs) were renamed based on their original sample and then de-replicated with dRep (Olm et al. 2017) with a primary cluster average nucleotide identity (ANI) threshold of 0.9, a secondary ANI threshold of 0.98, and a minimum genome completeness of 50%, which used completeness values estimated from CheckM (Parks et al. 2015). Only genomes that had a genome completeness higher than 50% were maintained (544 MAGs; **Figure 1A**). Next, genomes were removed using a genome quality score cutoff proposed in Parks et al. (2017) where quality was defined as *completeness - 5 x contamination* and only genomes with a quality equal to or greater than 50 were retained (325 MAGs, **Figure 1B**). The quality was divided into three groups including "Near Complete" (completeness greater than or equal to 90% with contamination equal to or less than 5%), "Medium Quality" (completeness greater than or equal to 70% with contamination equal to or less than 10%), and "Partial Quality" (completeness greater than or equal to 50% with contamination equal to or less than 4%; Parks et al. 2017). Finally, because the chosen threshold value in dRep was not stringent enough, pyani (average_nucleotide_identity.py; Pritchard et al. 2016) was used to calculate the ANI using blast (*i.e.,* ANIb) of each of the pairwise comparison. All genomes that had another genome that was 99% similar were removed and only the genome with the highest quality and lowest contamination value was kept (175 MAGs; **Figure 1C**).

MAG abundance estimation

Contigs within recovered and de-replicated MAGs were compiled into one fasta file and used to index and perform competitive mapping of quality controlled sequencing reads to the compiled fasta file with BWA (Li and Durbin 2009). Sam files from BWA were used to calculate the coverage estimates of each bin across all samples with pileup.sh from BBMap (Bushnell et al. 2017). The abundance of each MAG per sample was calculated by summing the Plus_reads and Minus_reads output from pileup.sh per bin and per lane.

Classification

Classification of MAGs was performed with GraftM (Boyd et al. 2018), which uses hidden Markov models to search for genes within each MAG and then places each MAG into pre-constructed gene trees. Here, 15 different ribosomal gene families were used to classify each MAG. Only classifications that were present at least 80% of the time were used for final classification. With the data from the 15 ribosomal datasets, a taxonomy table with a corresponding certainty table was constructed by checking all 15 classifications at each level, starting with the domain. If more than one classification was selected at the domain level, then the most frequent was selected. However, if there was a tie the domain was set to “unknown” along with all other lower taxonomic levels. Next, this process was performed at each taxonomic level. The certainty table included the percentage of the 15 datasets where the entire classification (*i.e.*, with all levels) appeared.

Annotation

Genes and their associated predicted proteins were called within each MAG using Prodigal (Hyatt et al. 2010). Predicted protein sequences were used to search for orthologous protein clusters using the reciprocal best alignment heuristic within proteinortho (Lechner et al. 2011). Next, the longest sequence within each orthologous cluster was chosen as a representative of each group for annotation. Finally, representative sequences were compared against the Pfam database release (version 31) for annotation using HMMER3 (Eddy 2011).

Statistical analysis

Further analysis of sequence data was performed in R version 3.5.1 (R Core Team 2018) using the phyloseq (McMurdie and Holmes 2013) and vegan (Oksanen et al. 2013) R packages. All figures were made using the ggplot2 R package (Wickham 2009). All code for the analyses will be available on GitHub when the study is complete.

Calculating microhabitat specialists and generalists

Particle-associated and free-living specialist bacteria were identified by using significantly differentially abundant MAGs between microhabitats by calculating the $\log_2$-ratio using the negative binomial generalized linear model framework of the DESeq function within the DESeq2 R package (McMurdie and Holmes 2013, Love et al. 2014). P-values were adjusted for multiple tests with a Benjamini-Hochberg false discovery rate correction and MAGs with a threshold P-value of less than 0.05 were retained in the dataset.

We also assess the genome coverage, estimated genome size, percent coding region, GC content, and the carbon to nitrogen ratio per MAG. The genome coverage (from the original sample) was calculated by multiplying the average sequence read length (117-119 bps) and the total number of sequences that mapped divided by the bin length. The estimated genome size of each MAG was measured by dividing the bin length by the completeness value from CheckM (Parks et al. 2015). The GC content was determined by concatenating all of the contigs from a MAG and inputting it into the gc content calculator for sequence objects from biopython (Cock et al. 2009). The average element (*i.e.*, carbon and nitrogen) composition of each gene was calculated by calling genes with Prodigal (Hyatt et al. 2010), measuring the number of elements for each amino acid in each gene, summing the count of elements over all amino acids in each gene, and dividing the sum by the length of the gene. The nutrient requirement for protein synthesis for each genome was expressed as the C:N ratio per amino acid.

Differences between free-living specialists, particle-associated specialists, and generalists were tested using a Kruskal-Wallis test (*kruskal.test()* function, *stats* R package, R Core Team 2018) along with a multiple comparison post-hoc test using the *kruskalmc* function (*pgirmess* R Package; Giraudoux 2018). To obtain more specific p-values for plotting, pairwise Wilcoxon tests were performed using the *wilcox.test* function (*stats* R package, R Core Team 2018). Correlational tests were performed using the *lm* function (*stats* R package, , R Core Team 2018). Rarefaction curves of gene-level (*i.e.*, pfam) diversity were plotted using the *ggrare* function, a phyloseq-extension wrapper (https://github.com/mahendra-mariadassou/phyloseq-extended).

Results

Of the original 2,774 metagenome assembled genomes (MAGs), only MAGs that had a genome completeness higher than 50% were maintained (544 MAGs; **Figure 4.1A**). Next, MAGs with a quality threshold (*as defined in the methods*) of 50 were retained (325 MAGs, **Figure 4.1B**). Finally, all MAGs that had another MAG that was at least 99% similar (based on the ANIb) was removed to obtain a final dataset of 175 MAGs (**Figure 4.1C**).

The number of total reads was similar between free-living (147.3 million sequencing reads) and particle-associated samples (144.4 million sequencing reads; 2-way ANOVA p = 0.82; **Figure SI 4.5A**). However, a higher percentage of reads mapped to free-living samples (46.6%) compared to particle-associated samples (28.1%; 2-way ANOVA p = 1 x 10^{-5}; **Figure SI 4.5B**). In the free-living, a higher percentage of reads mapped in 2014 (an average of 50%) compared to 2015 (an average of 43.2%; T-test: p = 0.048).

Specialization of free-living and particle-associated habitats

Of the 175 total MAGs, there were 88 MAGs (50.3%) were differentially abundant in either the free-living (*i.e.,* "free-living specialists", 50 MAGs) or particle-associated (*i.e.,* "particle-associated specialists", 38 MAGs) habitats, whereas 87 MAGs did not show a significant habitat preference and were deemed "generalist". Particle-associated specialists were dominated by phyla such as the Planctomycetes, Alphaproteobacteria, Gammaproteobacteria, and Armatimonadetes whereas the free-living specialists were dominated the Actinobacteria and Betaproteobacteria (**Figure SI 4.6**).

Genome coverage estimates were higher in particle-associated specialists than free-living specialists and generalists (KW: p = 4 x 10^{-6}; **Figure SI 4.7A**). However, free-living specialists had lower genome completeness estimates than generalists (Wilcoxon: p = 0.04, **Figure SI 4.7B**) but not particle-associated specialists (Kruskal-Wallis (KW): p = 0.11).

Particle-Associated specialists have larger genomes with less coding DNA

Particle-associated specialists had larger estimated genome sizes than both generalists and free-living specialists (KW: $p = 2 \times 10^{-16}$; **Figure 4.2A**) and more variance. On average, particle-associated specialist genomes were 4.75 Mbp, which were about two-times the size of generalists (2.69 Mbp average) and free-living specialists (2.12 Mbp average). This pattern was also reflected within members of the same Phylum (**Figure SI 4.8A**). In addition, particle-associated specialists had lower percent coding regions in their genomes compared to both free-living specialists and generalists (KW: $p = 3 \times 10^{-12}$; **Figure 4.2B**). Therefore, even though particle-associated bacteria have larger estimated genomes, the organisms represented by these MAGs have a lower percent of their genome dedicated to protein coding genes. To support this point, there was a moderate negative relationship between the estimated genome size and percent coding region of the genome across all MAGs ($R^2 = 0.42$, $p = 2 \times 10^{-22}$). However, the relationships for each of the individual three groups had less explanatory power, with the free-living having the strongest relationship ($R^2 = 0.38$, $p = 1 \times 10^{-6}$), followed by generalists ($R^2 = 0.19$, $p = 2 \times 10^{-5}$), and the weakest relationship in the particle-associated specialists ($R^2 = 0.16$, $p = 0.008$; **Figure 4.2C**).

Particle-Associated specialists have higher %GC and a larger nitrogen demand

Particle-associated specialists had a higher GC content compared with free-living specialists and generalists (KW: $p = 0.005$; **Figure 4.3A**). Particle-associated specialists had an average GC content of 58.3% whereas generalists had an average of 51.9% and free-living specialists had an average of 51.7%. In addition, the predicted protein sequences of particle-associated specialists had significantly lower carbon to nitrogen ratios per amino acid (KW: $p = 0.001$; **Figure 4.3BA**), indicating that they have a higher nitrogen content per amino acid than both generalist (Wilcoxon: $p = 0.003$) and free-living specialists (Wilcoxon: $p = 1 \times 10^{-4}$; **Figure 4.3B**). The C:N ratio was negatively associated with estimated genome size with a more severe slope in free-living specialists and generalists ($R^2 = 0.28$, $p = 5 \times 10^{-5}$ and $R^2 = 0.28$, $p = 7 \times 10^{-8}$, respectively) compared with particle-associated specialists ($R^2 = 0.11$, $p = 0.026$; **Figure 4.3C**).

Particle-Associated specialists have fewer total and unique genes per region of the genome

Particle-associated specialists had more total (KW: $p = 8.9 \times 10^{-9}$) and unique (KW: $p = 6.8 \times 10^{-8}$) genes in their genomes compared to generalists and free-living specialists (**Figure SI 4.9A and Figure SI 4.9B**). However, larger genomes tended to have more total and unique genes within each group as there was a strong to moderate, positive and linear relationship between estimated genome size and the number of genes and unique genes in free-living specialists ($R^2 = 0.63$, $p = 6 \times 10^{-12}$), generalists ($R^2 = 0.43$, $p = 3 \times 10^{-12}$), and particle-associated specialists ($R^2 = 0.36$, $p = 4 \times 10^{-5}$; **Figure SI 4.9B**). Therefore, to make meaningful ecological conclusions regarding the number of total and unique genes, the data must be normalized by genome size.

After normalizing by genome size, particle-associated specialists had fewer total genes per region of the genome compared to generalists (Wilcoxon: $p = 1.2 \times 10^{-11}$) and free-living specialists (Wilcoxon: $p = 6.5 \times 10^{-20}$; **Figure 4.4, x-axis and top**). Additionally, particle-associated specialists had fewer unique genes per region of their genome than generalists (Wilcoxon: $p = 2.2 \times 10^{-10}$) and free-living specialists (Wilcoxon: $p = 3.3 \times 10^{-16}$; **Figure 4.4, y-axis and right**). There was a very strong positive, linear relationship between the normalized total and unique genes per region of the genome ($R^2 = 0.83$; $p = 5 \times 10^{-66}$; **Figure 4.4**). When normalized using the coding length of each MAG, the relationship was comparable ($R^2 = 0.80$; $p = 5.8 \times 10^{-62}$). These results indicate that free-living specialists, and to a lesser extent generalists, pack more genes and more unique genes into their genomes per region than do particle-associated bacteria.

Discussion

I present data testing two hypotheses regarding the genomic characteristics of free-living and particle-associated bacterial specialist and generalist taxa. First, I found that particle-associated bacteria have larger genomes (**Figure 4.2A**), a high proportion of non-coding regions (**Figure 4.2B**), and higher GC content (**Figure 4.3A**). In contrast, free-living specialists, and to a lesser extent generalists, had signatures of genome streamlining in all tested characteristics. The average ratio of carbon to nitrogen for predicted protein sequences of particle-associated specialist genomes was lower (**Figure 4.3B**), suggesting that particle-associated bacteria have a higher nitrogen content than generalists and free-living bacteria, which depending on the turnover time may lead to higher or lower nitrogen demand for the cell. Second, I showed that while particle-associated bacteria have more unique predicted gene functions in their genome, they have fewer genes and less diverse genes per region in the genome compared to generalists and particle-associated bacteria.

I showed that particle-associated bacteria have larger genomes (**Figure 4.2A**). Large genomes may give organisms a competitive advantage during times of higher nutrient concentrations (Giovannoni 2005). One correlate of prokaryotic genome size is temporal environmental variability because large fluctuations in conditions select for larger genomes that have more genes (Bentkowski et al. 2015). Consistent with this control is the indication that prokaryotes that have evolved in more stable environmental conditions have a smaller genome and are more constrained in their ability to adjust to new environmental conditions (Bentkowski et al. 2015). Aquatic particles are known to contain environmental gradients, especially of dissolved organic matter and oxygen (Alldredge and Cohen 1987, Stocker 2012). Due to this, previous studies have proposed (Allen et al. 2013) and reported (Zeigler Allen et al. 2012) that particle-associated bacteria have on average larger genomes, however, Zeigler Allen et al. (2012) used gene-centric metagenomics which estimated the genome size across the entire sample instead of creating metagenome assembled genomes (MAGs). Therefore, I extend these initial findings from Zeigler Allen et al. (2012) by performing genome-centric analyses in our samples.

On the other hand, predominate free-living bacterial communities are often dominated by abundant oligotrophic taxa with streamlined genomes. Some examples include the marine alphaproteobacterial SAR11 clade (Giovannoni 2005), the marine cyanobacterium *Prochlorococcus* in oligotrophic surface oceans (García-Fernandez et al. 2004), and freshwater Actinobacteria (Neuenschwander et al. 2018). While some particle-associated bacteria are better suited to quickly take advantage of transient bursts in nutrients, free-living bacteria with streamlined genomes have been thought to make efficient use of ambient environmental conditions (Giovannoni 2005, Lauro et al. 2009). For example, the ubiquitous and abundant marine Alphaproteobacterial SAR11 clade (Morris et al. 2002) is one of the most abundant marine organisms, however, out of the organisms that replicate on their own in nature it also has one of the smallest genomes (1.3 Mbps; Giovannoni 2005, Luo 2015). Recently, an abundant freshwater SAR11 clade with an even smaller genome of 1.16 Mbps was found that also had an extremely low GC content of 29% (Henson et al. 2018). In freshwater lakes, Actinobacteria also have small genomes (1.16-1.47 Mbps) with low GC content (40-48%; Ghai et al. 2013, 2014, Neuenschwander et al. 2018) and are numerically dominant members of freshwater bacterial communities (Allgaier and Grossart 2006).

Selection for small cell size and efficient use of nutrients in oligotrophic environments is typically viewed as the major ecological reason for genome streamlining in free-living aquatic bacteria (Giovannoni et al. 2014). This is highlighted in the fact that marine bacteria have a bias for A and T usage, which decreases the GC content, and reduces the nitrogen budget of a cell by 3-10% (Grzymski and Dussaq 2012). Free-living bacteria in the current study had a lower GC content compared to particle-associated bacteria; however, there appeared to be a bimodal distribution in the generalists (**Figure 4.3A**). The C:N ratio was higher for free-living bacteria indicating that they have a lower nitrogen amino acid content than particle-associated bacteria (but not compared to generalists; **Figure 4.3B**). Recently, a genomic transition zone in GC content was found in Hawaii at station ALOHA where bacteria living in the ultra-oligotrophic surface waters had much lower GC content compared to all organisms living beneath the deep chlorophyll maximum (Mende et al. 2017). While a genomic transition zone cannot be explored in this dataset (as all samples were taken at 1 m), it does posit the question of whether there are similar ecological drivers for changes in GC and C:N ratios for particle-associated and free-living genomes in this dataset as there is for marine surface and deep-water bacterial genomes from Mende et al. (2017).

While it is known that the number of genes scale with genome size (Mira et al. 2001, Giovannoni 2005), we show that: 1) particle-associated specialists have higher absolute diversity in their gene capabilities (**Figure SI 4.9**), and 2) after normalizing for genome size, generalists, and to an even larger extent free-living specialists, pack more genes and more unique genes into their genomes per region than particle-associated bacteria. This may be partially explained by fewer intergenic spacers in streamlined genomes (Giovannoni et al. 2014). It may also be consistent with ecological interdependencies in bacterial communities, which recent findings in metagenomics have revealed is commonplace (Zelezniak et al. 2015, Anantharaman et al. 2016). Interdependence and metabolic "handouts" occur when microbes produce beneficial compounds that stimulate the growth of other organisms in the population and community and are public goods (West et al. 2007, Hug and Co 2018, Zengler and Zaramela 2018). In a metagenomic assessment of particle-associated and free-living habitats, Ganesh et al. (2014) found that genes for different steps of denitrification were more abundant in particles compared to the open water, indicating the potential of metabolic handoffs for nitrogen cycling, which has been reported in other systems (Anantharaman et al. 2016). Another example of handouts in particles is the production of siderophores, which has been shown to stimulate the growth and promote the evolution of diverse iron acquisition methods, such as production or cheating, which increases with association with particles (Cordero et al. 2012).

In oligotrophic aquatic conditions, free-living bacterial communities have vast population-level diversity (Kashtan et al. 2014, Garcia et al. 2018b) and, with use of network analysis, have been shown to be highly connected and correlated (Milici et al. 2016). These aspects of free-living communities hint at the potential for metabolic interdependence. For example, a recent study using environmental enrichments of abundant and streamlined freshwater free-living Actinobacteria showed that cells depended on "handouts" from other organisms in the poly-culture (*i.e.*, vitamins, amino acids, reduced sulfur; (Garcia et al. 2018a)). The handouts were from a variety of cells that they randomly encountered, rather than from specific cells or taxa within the community (Garcia et al. 2018a). This is in contrast to genome streamlining patterns of obligate symbionts where the symbiont relies so heavily on their host species that large sections of the genome are removed (McCutcheon and Moran 2012). However, because interdependence requires cooperation in microbial communities, free-living organisms in the environment may become dependent on complementary species (OTUs) that may not be reliably encountered and therefore selection often acts to limit cooperation because it is inefficient (Oliveira et al. 2014). Finally, it is currently undetermined whether the connectedness and correlations between free-living bacteria are due to the fact that free-living bacteria experience similar bulk environmental forces or whether these correlations are driven by actual community interactions (*e.g.*, interdependence and metabolic handouts) between free-living bacteria. Future work on metabolic connectedness could focus on testing the potential limitations of selection limiting cooperation and the environmental co-variation on free-living communities.

In this study, we confirm that freshwater free-living bacteria have streamlined genomes whereas particle-associated bacteria have larger genomes with higher GC content, more non-coding regions, higher nitrogen content for protein synthesis, and lower per-capita diversity in genes. A deeper look into the genes that are differentially abundant between specialist and generalist taxa will provide a more thorough view of their ecological adaptations.

References

Alldredge, A., and Y. Cohen. 1987. Can Microscale Chemical Patches Persist in the Sea? Microelectrode Study of Marine Snow, Fecal Pellets. Science 235:689–691.

Allen, A. E., L. Z. Allen, and J. P. McCrow. 2013. Lineage specific gene family enrichment at the microscale in marine systems. Current Opinion in Microbiology 16:605–617.

Allgaier, M., and H. P. Grossart. 2006. Seasonal dynamics and phylogenetic diversity of free-living and particle-associated bacterial communities in four lakes in northeastern Germany. Aquatic Microbial Ecology 45:115–128.

Alneberg, J., B. S. Bjarnason, I. De Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. 2014. Binning metagenomic contigs by coverage and composition. Nature Methods 11:1144–1146.

Anantharaman, K., C. T. Brown, L. A. Hug, I. Sharon, C. J. Castelle, A. J. Probst, B. C. Thomas, A. Singh, M. J. Wilkins, U. Karaoz, E. L. Brodie, K. H. Williams, S. S. Hubbard, and J. F. Banfield. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nature Communications 7:13219.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Balmonte, J. P., A. P. Teske, and C. Arnosti. 2018. Structure and function of high Arctic pelagic, particle- associated, and benthic bacterial communities. Environmental Microbiology.

Bentkowski, P., C. Van Oosterhout, and T. Mock. 2015. A model of genome size evolution for prokaryotes in stable and fluctuating environments. Genome Biology and Evolution 7:2344–2351.

Bižić-Ionescu, M., M. Zeder, D. Ionescu, S. Orlić, B. M. Fuchs, H. P. Grossart, and R. Amann. 2015. Comparison of bacterial communities on limnic versus coastal marine particles reveals profound differences in colonization. Environmental microbiology 17:3500–3514.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Boyd, J. A., B. J. Woodcroft, and G. W. Tyson. 2018. GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. Nucleic Acids Research 46.

Burke, C., P. Steinberg, D. B. Rusch, S. Kjelleberg, and T. Thomas. 2011. Bacterial community assembly based on functional genes rather than species. Proceedings of the National Academy of Sciences of the USA 108:14288–14293.

Bushnell, B., J. Rood, and E. Singer. 2017. BBMerge – Accurate paired shotgun read merging via overlap. PLoS ONE 12:1–15.

Chiang, E., M. L. Schmidt, M. A. Berry, B. A. Biddanda, A. Burtner, T. H. Johengen, D. Palladino, and V. J. Denef. 2018. Verrucomicrobia are prevalent in north- temperate freshwater lakes and display class- level preferences between lake habitats. PLoS ONE 13:1–20.

Cock, P. J. A., T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. De Hoon. 2009. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Cordero, O. X., L. Ventouras, E. F. DeLong, and M. F. Polz. 2012. Public good dynamics drive evolution of iron acquisition strategies in natural bacterioplankton populations. Proceedings of the National Academy of Sciences 109:20059–20064.

Defore, A. L., A. D. Weinke, M. M. Lindback, and B. A. Biddanda. 2016. Year-round measures of planktonic metabolism reveal net autotrophy in surface waters of a Great Lakes estuary. Aquatic Microbial Ecology 77:139–153.

Eddy, S. R. 2011. Accelerated profile HMM searches. PLoS Computational Biology 7.

Ewels, P., M. Magnusson, S. Lundin, and M. Käller. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32:3047–3048.

Ganesh, S., D. J. Parris, E. F. DeLong, and F. J. Stewart. 2014. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. The ISME journal 8:187–211.

García-Fernandez, J. M., N. de Marsac, and J. Diez. 2004. Streamlined Regulation and Gene Loss as Adaptive Mechanisms in. Microbiology and molecular biology reviews 68:630–638.

Garcia, S. L., M. Buck, J. J. Hamilton, C. Wurzbacher, H.-P. Grossart, K. D. McMahon, and A. Eiler. 2018a. Model Communities Hint at Promiscuous Metabolic Linkages between Ubiquitous Free-Living Freshwater Bacteria. mSphere 3:1–8.

Garcia, S. L., S. L. R. Stevens, B. Crary, M. Martinez-Garcia, R. Stepanauskas, T. Woyke, S. G. Tringe, S. G. E. Andersson, S. Bertilsson, R. R. Malmstrom, and K. D. McMahon. 2018b. Contrasting patterns of genome-level diversity across distinct co-occurring bacterial populations. ISME Journal 12:742–755.

Ghai, R., C. M. Mizuno, A. Picazo, A. Camacho, and F. Rodriguez-Valera. 2013. Metagenomics uncovers a new group of low GC and ultra-small marine Actinobacteria. Scientific Reports 3:1–8.

Ghai, R., C. M. Mizuno, A. Picazo, A. Camacho, and F. Rodriguez-Valera. 2014. Key roles for freshwater Actinobacteria revealed by deep metagenomic sequencing. Molecular Ecology 23:6073–6090.

Ghiglione, J. F., G. Mevel, M. Pujo-Pay, L. Mousseau, P. Lebaron, and M. Goutx. 2007. Diel and seasonal variations in abundance, activity, and community structure of particle-attached and free-living bacteria in NW Mediterranean Sea. Microbial Ecology 54:217–231.

Giovannoni, S. J. 2005. Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science 309:1242–1245.

Giovannoni, S. J., J. Cameron Thrash, and B. Temperton. 2014. Implications of streamlining theory for microbial ecology. The ISME journal 8:1–13.

Giraudoux, P. 2018. pgirmess: Spatial Analysis and Data Mining for Field Ecologists.

Grossart, H. P. 2010. Ecological consequences of bacterioplankton lifestyles: Changes in concepts are needed. Environmental Microbiology Reports 2:706–714.

Grzymski, J. J., and A. M. Dussaq. 2012. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME Journal 6:71–80.

Henson, M. W., V. C. Lanclos, B. C. Faircloth, and J. C. Thrash. 2018. Cultivation and genomics of the first freshwater SAR11 (LD12) isolate. ISME Journal 12:1846–1860.

Hug, L. A., and R. Co. 2018. It Takes a Village: Microbial Communities Thrive through Interactions and Metabolic Handoffs. mSystems 3:e00152-17.

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320:1081–1085.

Hyatt, D., G. L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, and L. J. Hauser. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics 11:119.

Kang, D. D., J. Froula, R. Egan, and Z. Wang. 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3:e1165.

Kashtan, N., S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, P. Marttinen, R. R. Malmstrom, R. Stocker, M. J. Follows, R. Stepanauskas, and S. W. Chisholm. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science 344:416–420.

Konstantinidis, K. T., and J. M. Tiedje. 2004. Trends between gene content and genome size in prokaryotic species with larger genomes. Proceedings of the National Academy of Sciences 101:3160–3165.

Laczny, C. C., T. Sternal, V. Plugaru, P. Gawron, A. Atashpendar, H. H. Margossian, S. Coronado, L. V. der Maaten, N. Vlassis, and P. Wilmes. 2015. VizBin - An application for reference-independent visualization and human-augmented binning of metagenomic data. Microbiome 3:1–7.

Lauro, F. M., D. McDougald, T. Thomas, T. J. Williams, S. Egan, S. Rice, M. Z. DeMaere, L. Ting, H. Ertan, J. Johnson, S. Ferriera, A. Lapidus, I. Anderson, N. Kyrpides, A. C. Munk, C. Detter, C. S. Han, M. V Brown, F. T. Robb, S. Kjelleberg, and R. Cavicchioli. 2009. The genomic basis of trophic strategy in marine bacteria. Proceedings of the National Academy of Sciences of the United States of America 106:15527–15533.

Lechner, M., S. Findeiß, L. Steiner, M. Marz, P. F. Stadler, and S. J. Prohaska. 2011. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. BMC Bioinformatics 12:124.

Lemarchand, C., L. Jardillier, J. F. Carrias, M. Richardot, D. Debroas, T. Sime-Ngando, and C. Amblard. 2006. Community composition and activity of prokaryotes associated to detrital particles in two contrasting lake ecosystems. FEMS Microbiology Ecology 57:442–451.

Li, D., C. M. Liu, R. Luo, K. Sadakane, and T. W. Lam. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15:1–21.

Luo, H. 2015. Evolutionary origin of a streamlined marine bacterioplankton lineage. The ISME Journal 9:1423–1433.

McCarthy, A., E. Chiang, M. L. Schmidt, and V. J. Denef. 2015. RNA Preservation Agents and Nucleic Acid Extraction Method Bias Perceived Bacterial Community Composition. Plos One 10:e0121659.

McCutcheon, J. P., and N. A. Moran. 2012. Extreme genome reduction in symbiotic bacteria. Nature Reviews Microbiology 10:13–26.

McInnes, L., J. Healy, and S. Astels. 2017. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software 2:11–12.

McMurdie, P. J., and S. Holmes. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217.

Mende, D. R., J. A. Bryant, F. O. Aylward, J. M. Eppley, T. Nielsen, D. M. Karl, and E. F. Delong. 2017. Environmental drivers of a microbial genomic transition zone in the ocean's interior. Nature Microbiology 2:1367–1373.

Mestre, M., C. Ruiz-González, R. Logares, C. M. Duarte, J. M. Gasol, and M. M. Sala. 2018. Sinking particles promote vertical connectivity in the ocean microbiome. Proceedings of the National Academy of Sciences 115:E6799–E6807.

Milici, M., Z. L. Deng, J. Tomasch, J. Decelle, M. L. Wos-Oxley, H. Wang, R. Jï¿½uregui, I. Plumeier, H. A. Giebel, T. H. Badewien, M. Wurst, D. H. Pieper, M. Simon, and I. Wagner-Dï¿½bler. 2016. Co-occurrence analysis of microbial taxa in the Atlantic ocean reveals high connectivity in the free-living bacterioplankton. Frontiers in Microbiology 7:1–20.

Mira, A., H. Ochman, and N. Moran. 2001. Deletional bias and the evolution of bacterial genomes Cited by me. Trends in Genetics 17:589–596.

Mohit, V., P. Archambault, N. Toupoint, and C. Lovejoy. 2014. Phylogenetic differences in attached and free-living bacterial communities in a temperate coastal lagoon during summer, revealed via high-throughput 16S rRNA gene sequencing. Applied and Environmental Microbiology 80:2071–2083.

Morris, R. M., M. S. Rappé, S. A. Connon, K. L. Vergin, W. A. Siebold, C. A. Carlson, and S. J. Giovannoni. 2002. SAR11 clade dominates ocean surface bacterioplankton communities. Nature 420:806–810.

Neuenschwander, S. M., R. Ghai, J. Pernthaler, and M. M. Salcher. 2018. Microdiversification in genome-streamlined ubiquitous freshwater Actinobacteria. ISME Journal 12:185–198.

Newton, R. J., S. E. Jones, A. Eiler, K. D. McMahon, and S. Bertilsson. 2011. A guide to the natural history of freshwater lake bacteria. Page Microbiology and molecular biology reviews : MMBR.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner. 2013. vegan: Community Ecology Package.

Oliveira, N. M., R. Niehus, and K. R. Foster. 2014. Evolutionary limits to cooperation in microbial communities. Proceedings of the National Academy of Sciences 111:17941–17946.

Olm, M. R., C. T. Brown, B. Brooks, and J. F. Banfield. 2017. DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME Journal 11:2864–2868.

Parks, D. H., M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research 25:1043–1055.

Parks, D. H., C. Rinke, M. Chuvochina, P. A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson. 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature Microbiology 2:1533–1542.

Polz, M. F., and O. X. Cordero. 2016. Bacterial evolution: Genomics of metabolic trade-offs. Nature Microbiology 1.

Pritchard, L., R. H. Glover, S. Humphris, J. G. Elphinstone, and I. K. Toth. 2016. Genomics and taxonomy in diagnostics for food security: Soft-rotting enterobacterial plant pathogens. Analytical Methods 8:12–24.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Salcher, M. M. 2014. Same same but different: Ecological niche partitioning of planktonic freshwater prokaryotes. Journal of Limnology 73:74–87.

Satinsky, B. M., C. B. Smith, S. Sharma, P. M. Medeiros, V. J. Coles, P. L. Yager, B. C. Crump, and M. A. Moran. 2016. Variation in Expression Levels of Elemental Cycling Genes within the Amazon River Plume. Isme J:submitted.

Schmidt, M. L., B. A. Biddanda, A. D. Weinke, E. Chiang, F. Januska, R. Props, and V. J. Denef. 2017. Microhabitats shape diversity-productivity relationships in freshwater bacterial communities. bioRxiv:231688.

Schmidt, M. L., J. D. White, and V. J. Denef. 2016. Phylogenetic conservation of freshwater lake habitat preference varies between abundant bacterioplankton phyla. Environmental Microbiology 18:1212–1226.

Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, G. Szabo, M. F. Polz, and E. J. Alm. 2012. Population Genomics of Early Events in the Ecological Differentiation of Bacteria. Science 336:48–51.

Sieber, C. M. K., A. J. Probst, A. Sharrar, B. C. Thomas, M. Hess, S. G. Tringe, and J. F. Banfield. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 3:836–843.

Simon, H. M., M. W. Smith, and L. Herfort. 2014. Metagenomic insights into particles and their associated microbiota in a coastal margin ecosystem. Frontiers in Microbiology 5:466.

Smith, M. W., L. Zeigler Allen, A. E. Allen, L. Herfort, and H. M. Simon. 2013. Contrasting genomic properties of free-living and particle-attached microbial assemblages within a coastal ecosystem. Frontiers in microbiology 4:120.

Stocker, R. 2012. Marine microbes see a sea of gradients. Science 338:628–33.

Swan, B. K., B. Tupper, A. Sczyrba, F. M. Lauro, M. Martinez-Garcia, J. M. González, H. Luo, J. J. Wright, Z. C. Landry, N. W. Hanson, B. P. Thompson, N. J. Poulton, P. Schwientek, S. G. Acinas, S. J. Giovannoni, M. A. Moran, S. J. Hallam, R. Cavicchioli, T. Woyke, and R. Stepanauskas. 2013. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proceedings of the National Academy of Sciences of the United States of America 110:11463–8.

Weinke, A. D., S. T. Kendall, D. J. Kroll, E. A. Strickler, M. E. Weinert, T. M. Holcomb, A. A. Defore, D. K. Dila, M. J. Snider, L. C. Gereaux, and B. A. Biddanda. 2014. Systematically variable planktonic carbon metabolism along a land-to-lake gradient in a Great Lakes coastal zone. Journal of Plankton Research 36:1528–1542.

West, S. a., S. P. Diggle, A. Buckling, A. Gardner, and A. S. Griffin. 2007. The Social Lives of Microbes. Annual Review of Ecology, Evolution, and Systematics 38:53–77.

Wickham, H. 2009. ggplot2: elegant graphics for data analysis. Springer New York.

Wu, Y. W., Y. H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer. 2014. MaxBin: An automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome 2:1–18.

Xu, H., X. Luo, J. Qian, X. Pang, J. Song, G. Qian, J. Chen, and S. Chen. 2012. FastUniq: A Fast De Novo Duplicates Removal Tool for Paired Short Reads. PLoS ONE 7:1–6.

Zeigler Allen, L., E. E. Allen, J. H. Badger, J. P. McCrow, I. T. Paulsen, L. D. Elbourne, M. Thiagarajan, D. B. Rusch, K. H. Nealson, S. J. Williamson, J. C. Venter, and A. E. Allen. 2012. Influence of nutrients and currents on the genomic composition of microbes across an upwelling mosaic. ISME Journal 6:1403–1414.

Zelezniak, A., S. Andrejev, O. Ponomarova, D. R. Mende, P. Bork, and K. R. Patil. 2015. Metabolic dependencies drive species co-occurrence in diverse microbial communities. Proceedings of the National Academy of Sciences 112:6449–6454.

Zengler, K., and L. S. Zaramela. 2018. The social network of microorganisms - How auxotrophies shape complex communities. Nature Reviews Microbiology 16:383–390.

Figure 4.1. Quality assessment of metagenome assembled genomes used in this study.

Assessment of genome quality using three levels of quality control thresholds of “Near Complete” (dark teal), “Medium Quality” (green), and “Partial Quality” (tan) based on values from (Parks et al., 2017) as described in the methods. **A:** 544 genomes were constructed with completeness values of 50% and higher. **B:** 325 genomes were left after using a quality cutoff of 50. **C:** 175 genomes were maintained after removal of genomes with 99% ANIb values.

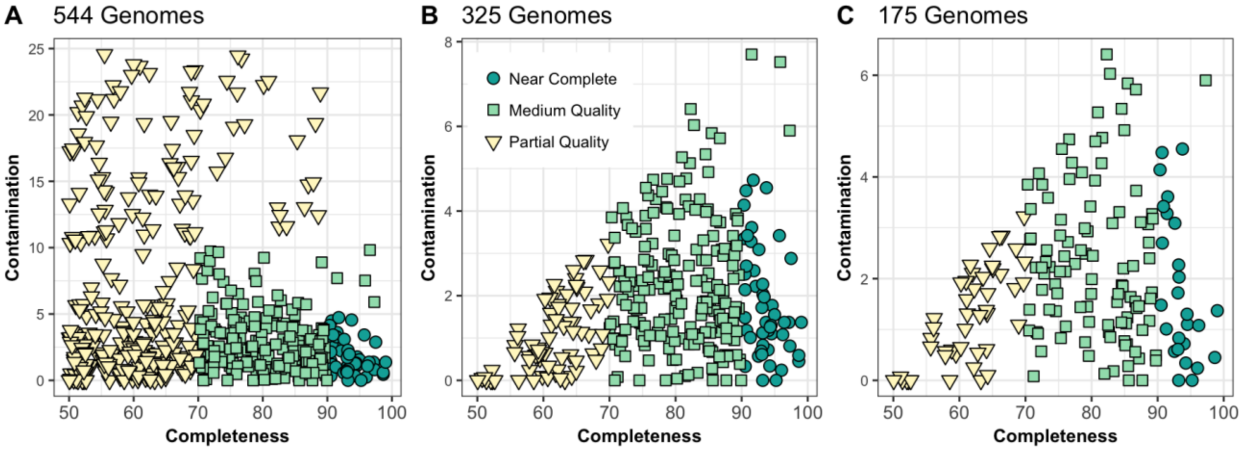


Figure 4.2. Estimated genome size, percent coding region, and their relationship.

Estimated genome size **(A)**, percent coding region of genome **(B)** and the relationship between the two **(C)** for free-living specialists (blue), generalists (grey), and particle-associated specialists (orange). Significance values in A and B represent p-values from pairwise Wilcoxon tests. R^2 and p-values in **(C)** represent the individual linear regressions between percent coding region and estimated genome size per each of the three groups.

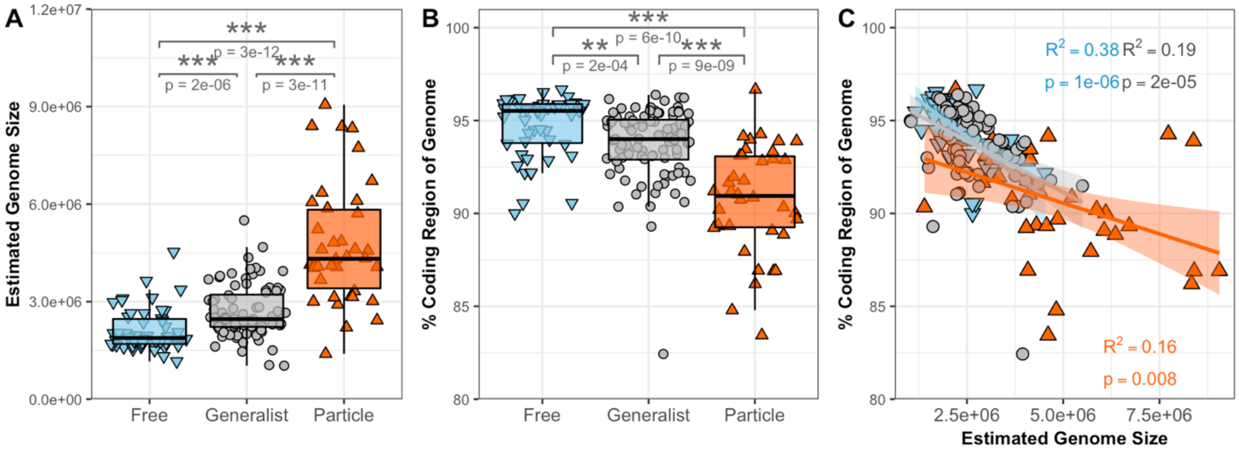


Figure 4.3. Percent GC content, average carbon to nitrogen (C:N) ratio per amino acid, and the relationship of C:N ratio to genome size.

Percent GC content of MAGs **(A),** average carbon to nitrogen ratio per amino acid **(B)** and the relationship between the average carbon to nitrogen ratio per amino acid and estimated genome size **(C)** for generalists (grey) and free-living (blue) and particle-associated (orange) specialists.

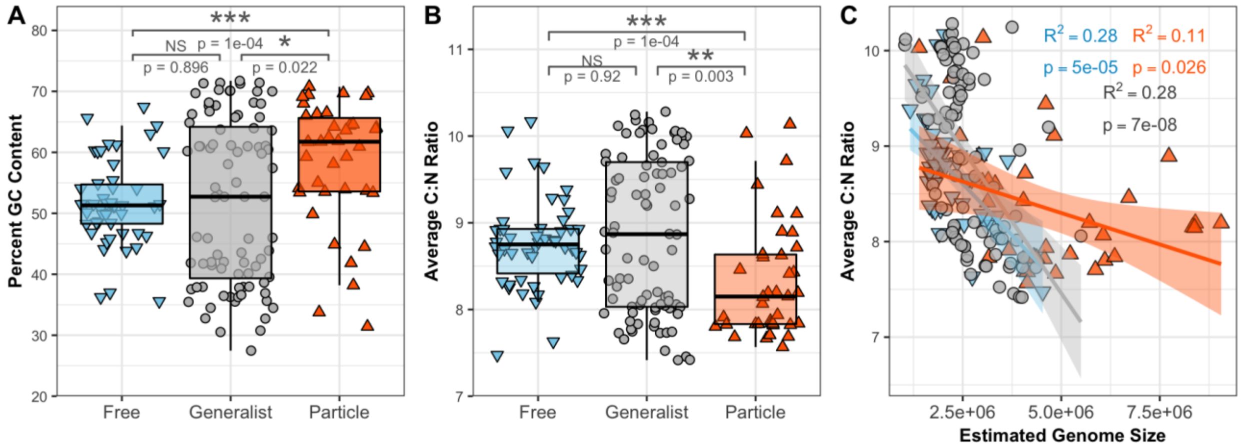


Figure 4.4. Normalized unique genes per Mbps versus normalized total genes per Mbps.

Normalized total gene abundance (by MAG length in bps; x-axis) versus normalized unique gene count (by MAG length in bps; y-axis). **Top:** Distribution of points along the x-axis (KW: p = 1.2 x 10^{-22}). **Right:** Distribution of points along the y-axis (KW: p = 5.2 x 10^{-17}).

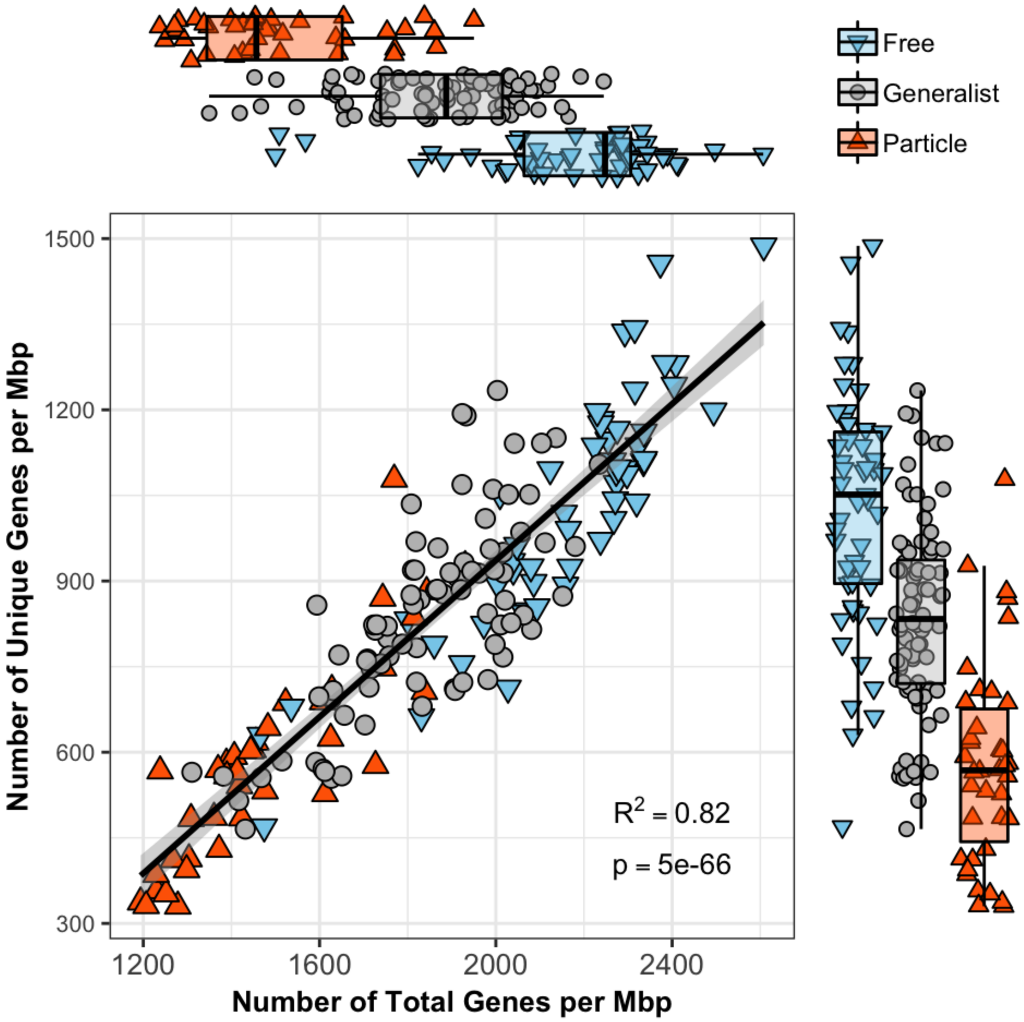


Figure SI 4.5. Total number of sequencing reads and percent of reads mapped.

The total number of sequencing reads **(A)** and the percentage of competitively mapped to total reads **(B)** across the 8 free-living (blue) and 8 particle-associated (orange) samples.

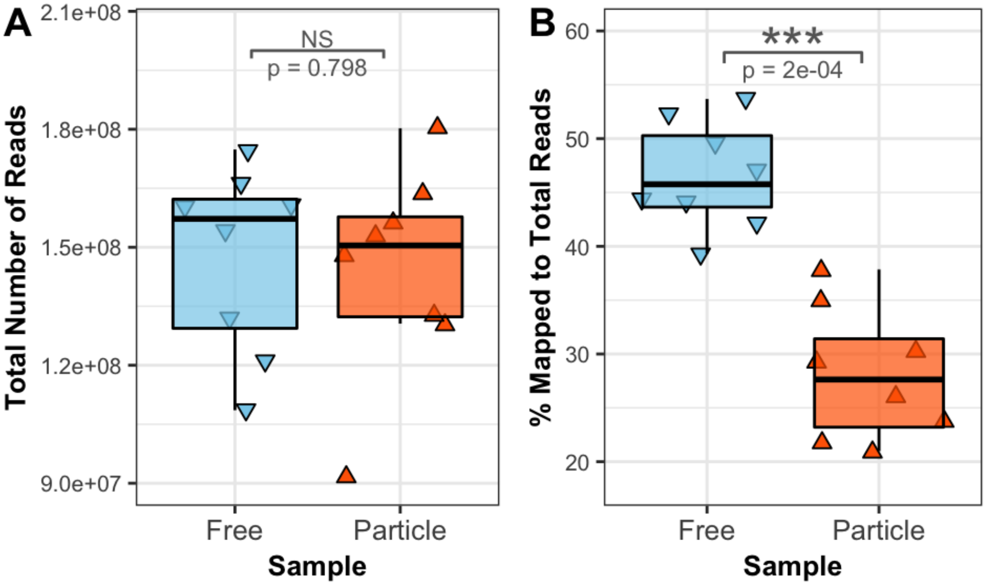


Figure SI 4.6. Phylogeny of the 175 genomes with their phylum (inner circle) and microhabitat specialization (outer circle).

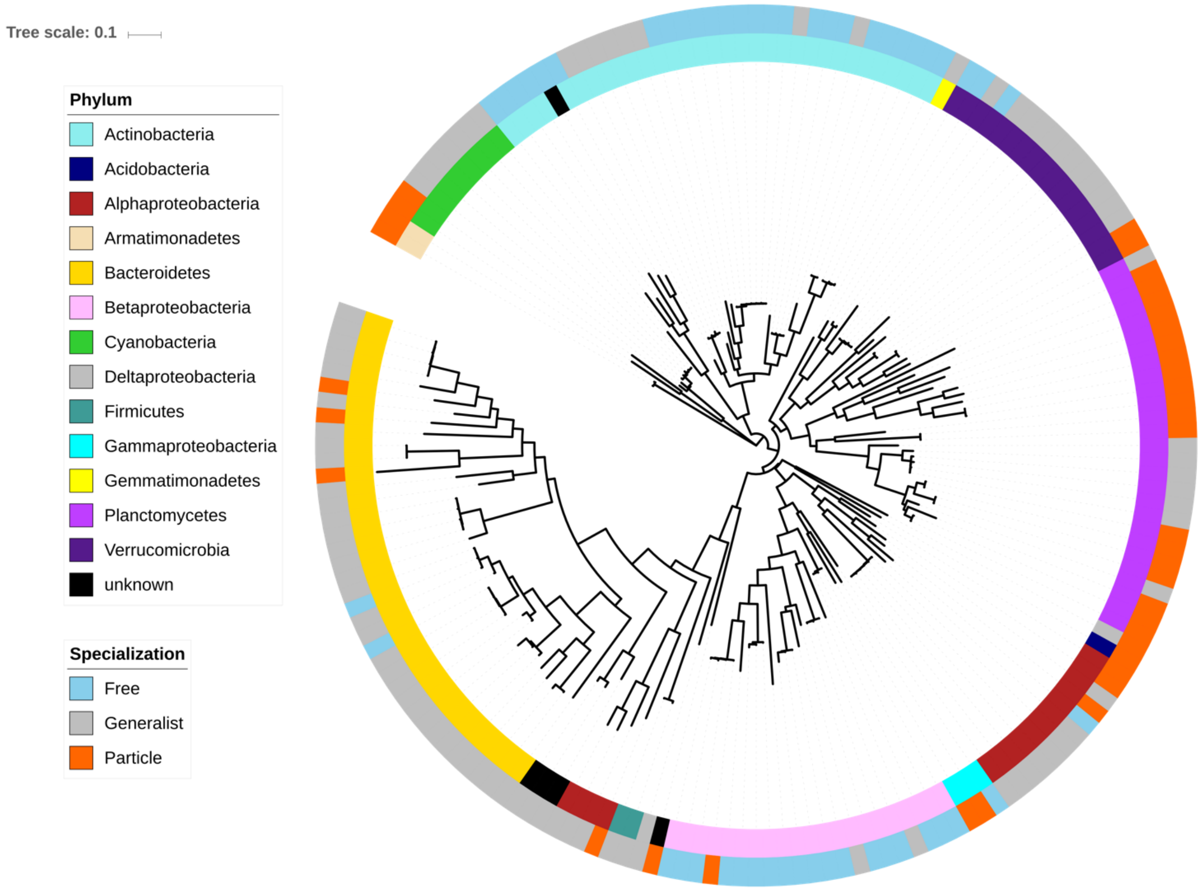


Figure SI 4.7. Genome coverage and completeness of MAGs in this study.

Genome coverage in percent **(A)** and completeness in percent **(B)** of free-living (blue), generalist (grey), and particle-associated (orange).

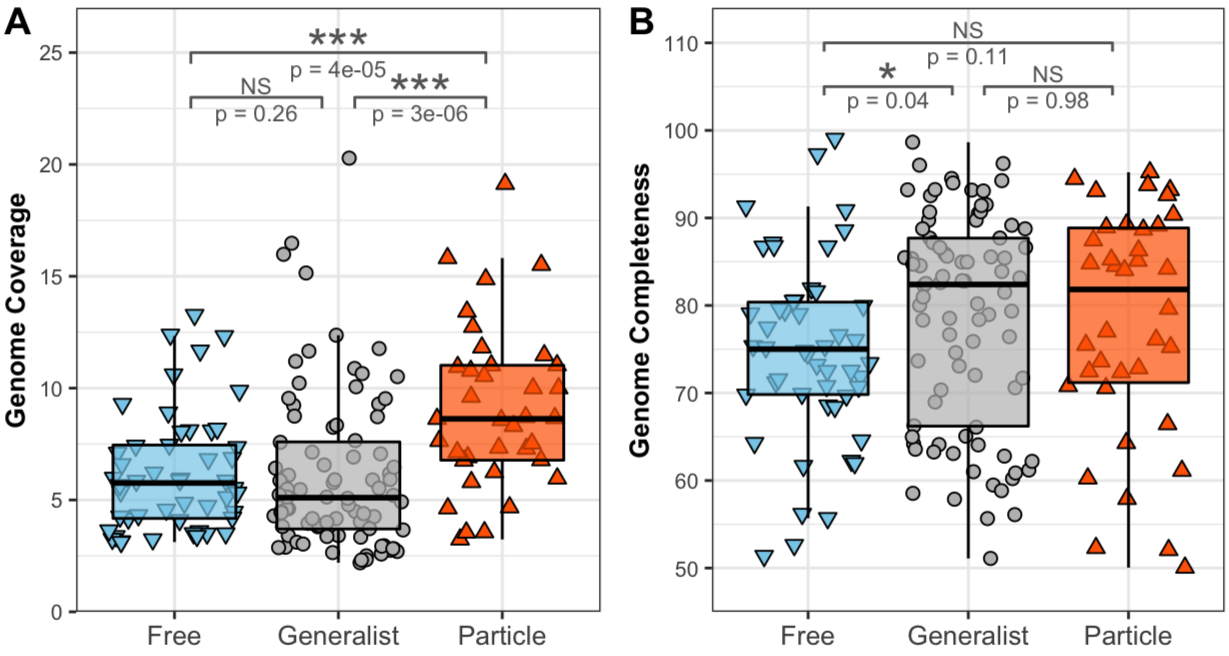


Figure SI 4.8. The estimated genome sizes of bacterial phyla that have at least 20 MAGs separated by microhabitat specialization.

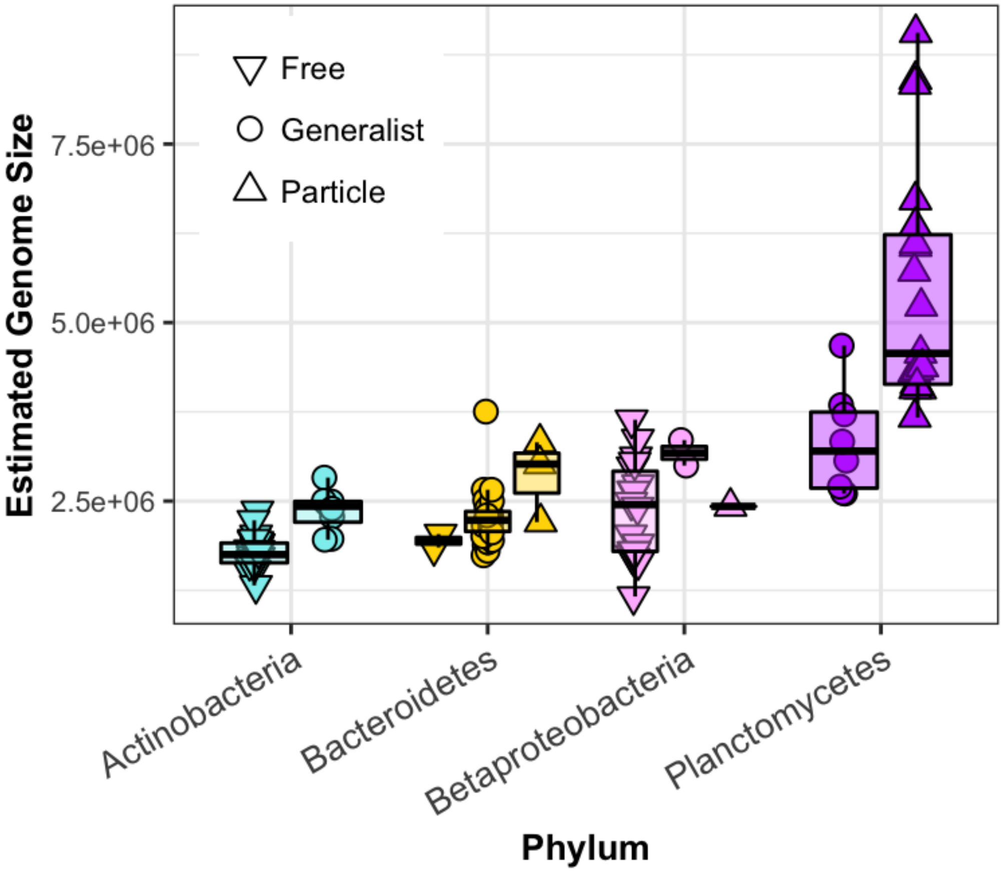
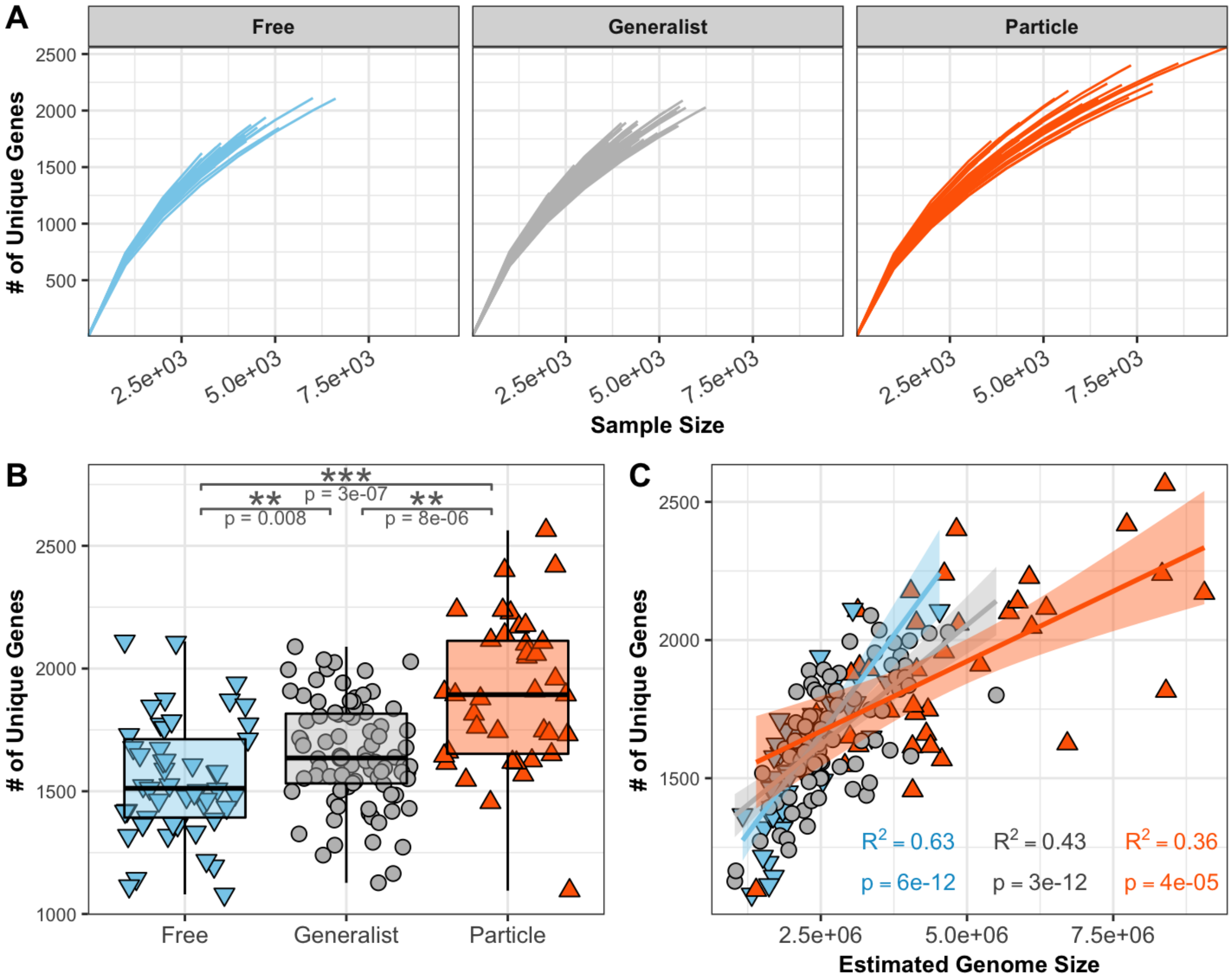


Figure SI 4.9. The number of unique genes per MAG and its relationship with genome size.

The number of unique genes within each MAG by habitat specialization (*i.e.*, free-living specialists (blue), generalists (grey), and particle-associated specialists (orange)) visualized with a rarefaction curve **(A)**, boxplot **(B)**, and the association between the number of unique genes and estimated genome size **(C)**.



Chapter V:

Using machine learning to associate bacterial taxa with functional groups through flow cytometry, 16S rRNA gene sequencing, and productivity data[3]

ABSTRACT

High- (HNA) and low-nucleic acid (LNA) bacteria are two separated flow cytometry (FCM) groups that are ubiquitous across aquatic systems. HNA cell density often correlates strongly with heterotrophic production. However, the taxonomic composition of bacterial taxa within HNA and LNA groups remains mostly unresolved. Here, we associated freshwater bacterial taxa with HNA and LNA groups by integrating FCM and 16S rRNA gene sequencing using a machine learning-based variable selection approach. There was a strong association between bacterial heterotrophic production and HNA cell abundances ($R^2 = 0.65$), but not with more abundant LNA cells, suggesting that the smaller pool of HNA bacteria may play a disproportionately large role in the freshwater carbon flux. Variables selected by the models were able to predict HNA and LNA cell abundances at all taxonomic levels, with highest accuracy at the OTU level. There was high system specificity as the selected OTUs were mostly unique to each lake ecosystem and some OTUs were selected for both groups or were rare. Our approach allows for the association of OTUs with FCM functional groups and thus the

[3] Available on *bioRxiv* and currently in review:
Rubbens P*, Schmidt ML*, Props R, Biddanda BA, Boon N, Waegeman W, Denef VJ. Using machine learning to associate bacterial taxa with functional groups through flow cytometry, 16S rRNA gene sequencing, and productivity data. https://doi.org/10.1101/392852

* *Indicates equal contribution of authors*

identification of putative indicators of heterotrophic activity in aquatic systems, an approach that can be generalized to other ecosystems and functions of interest.

Introduction

A key goal in the field of microbial ecology is to understand the relationship between microbial diversity and ecosystem functions. However, it is challenging to associate bacterial taxa to specific ecosystem processes. Marker gene surveys have shown that natural bacterial communities are extremely diverse; however, the presence of a taxon does not imply its activity. Taxa present in these surveys may have low metabolic potential, be dormant, or have recently died (Lennon and Jones 2011, Carini et al. 2016). Therefore, new methodologies that integrate different data types are needed to associate bacterial taxa with ecosystem functions in order to ultimately model and predict them (Widder et al. 2016).

One such advance is the use of flow cytometry (FCM), which has been used extensively to study aquatic microbial communities (Gasol and Del Giorgio 2000, Vives-Rego et al. 2000, Wang et al. 2010). This single-cell technology partitions individual microbial cells into phenotypic groups based on their observable optical characteristics. Most commonly, cells are stained with a nucleic acid stain (*e.g.*, SYBR Green I) and upon analysis assigned to either a low nucleic acid (LNA) or a high nucleic acid (HNA) group (Gasol et al. 1999, Lebaron et al. 2001, Bouvier et al. 2007, Wang et al. 2009). HNA cells differ from LNA cells in both a considerable increase in fluorescence due to cellular nucleic acid content and scatter intensity due to cell morphology. The HNA group is thought to correspond to the ‘active’ fraction, whereas the LNA population has been considered as the ‘dormant’ or ‘inactive’ group of a microbial community (Gasol and Del Giorgio 2000, Lebaron et al. 2002, Servais et al. 2003, Morán et al. 2007). This is based on positive linear relationships between HNA abundance and (a) bacterial heterotrophic production (BP) (Servais et al. 1999, 2003, Lebaron et al. 2001, Morán et al. 2007, Bowman et al. 2017), (b) bacterial activity measured using the dye 5-cyano-2,3-ditolyl tetrazolium chloride (Morán et al. 2011, Read et al. 2015), and (c) phytoplankton abundance (Sherr et al. 2006). Additionally, growth rates are higher for HNA than LNA cells (Servais et al. 1999, Lebaron et al. 2002, Jochem et al. 2004) and HNA cells accrue cell damage significantly faster than the LNA cells under stress from temperature (Arnoldini et al. 2013) and chemical oxidants (Ramseier et al. 2011).

However, it is still unclear whether HNA and LNA groups are composed of unique taxa or if they are different physiological states of the same taxa. Bouvier et al. (2007) proposed four possible scenarios: (1) bacteria start their life cycle in the HNA group and move to the LNA group upon death or inactivity; (2) cells in the HNA group originate from LNA cells undergoing cell division; (3) HNA and LNA consist of different non-overlapping taxa; (4) bacteria switch between groups from time to time in addition to having part of the community that is unique to each fraction. The view that HNA cells are more active is in line with scenario 1 and 2. On the other hand, several studies have found distinct groups with little taxonomic overlap and proposed scenario 3 (Schattenhofer et al. 2011, Proctor et al. 2018) or 3 and 4 (Vila-Costa et al. 2012). In this case, HNA and LNA groups have been associated with different life strategies in bacterioplankton communities, such as large cell size (HNA) versus small cell size (LNA) (Morán et al. 2007, Proctor et al. 2018), genome size (Bowman et al. 2017) and ploidy (Schattenhofer et al. 2011). By combining FCM with taxonomic identification of bacterial communities, one can associate individual taxa with population dynamics and functioning.

In this study, we developed a novel approach to associate the dynamics of individual taxa with those of the LNA and HNA groups in freshwater lakes by using a machine learning variable selection strategy. We applied two variable selection methods, the Randomized Lasso (Meinshausen and Bühlmann 2010) and the Boruta algorithm (Kursa and Rudnicki 2010) to associate individual taxa with HNA and LNA cell abundances. This approach allowed us to associate specific taxa to FCM functional groups, and via the observed HNA-productivity relationship, to functioning. In addition, this approach enabled us to test the influence of rare taxa on these two groups as recent research has found that rare taxa may have a strong impact on community structure and functioning (Shade et al. 2014, Herren and McMahon 2018). To validate the RL-based association with the HNA or LNA group, we correlated taxon abundances with specific regions in the FCM fingerprint without prior knowledge of the HNA/LNA group. Furthermore, we tested for phylogenetic conservation of HNA and LNA functional groups and for the association between the selected taxa and productivity. The combination of FCM and 16S rRNA gene sequencing allows for the inference and assessment of the taxonomic structure of HNA and LNA groups, therefore advancing our ability to link bacterial taxa to their functionality in nature. This knowledge will help identify the taxa that drive carbon fluxes in freshwater ecosystems, which are disproportionately large relative to the global freshwater surface area (Biddanda 2017).

Results

In this study, we developed a machine learning variable selection strategy to integrate FCM and 16S rRNA gene sequencing with the aim of inferring the bacterial drivers of functional groups in freshwater lake systems. We studied a set of oligo- to eutrophic small inland lakes, a short residence time mesotrophic freshwater estuary lake (Muskegon Lake), and a large oligotrophic Great Lake (Lake Michigan), all located in Michigan, USA. We showed that abundance variation of these FCM functional groups is predicted by a small subset of all taxa that are present in the environment. Selected taxa were mostly FCM groups and lake system specific, and across systems association with HNA or LNA was not phylogenetically conserved. The relationship between selected taxa and productivity measurements was assessed for one of the lake systems (Muskegon Lake), thereby showing that HNA cells (and their putative bacterial taxa) likely turn over faster and disproportionately contribute to the freshwater carbon flux.

Study lakes are dominated by LNA cells

The inland lakes (6.3 x 10^6 cells/mL) and Muskegon Lake (6.0 x 10^6 cells/mL) had significantly higher total cell abundances than Lake Michigan (1.7 x 10^6 cells/mL; p = 2.7 x 10^{-14}). Across all lakes, the mean proportion of HNA cell counts (HNAcc) to total cell counts was much lower (29-33%) compared to the mean proportion of LNA cell counts (LNAcc; 67-71%). Ordinary least squares regression showed a strong correlation between HNAcc and LNAcc across all data (R^2 = 0.45, P = 2 x 10^{-24}; **Figure 5.1A**); however, only Lake Michigan (R^2 = 0.59, p = 5 x 10^{-11}) and Muskegon Lake (R^2 = 0.44, p = 2 x 10^{-9}) had significant correlations when the three ecosystems were considered separately.

HNA cell counts and heterotrophic bacterial production are strongly correlated

For mesotrophic Muskegon Lake, there was a strong correlation between total bacterial heterotrophic production and HNAcc ($R^2 = 0.65$, $p = 1 \times 10^{-5}$; **Figure 5.1B**), no correlation between BP and LNAcc ($R^2 = 0.005$, $p = 0.31$; **Figure 5.1C**), and a weak correlation between heterotrophic production and total cell counts ($R^2 = 0.18$, $p = 0.03$; **Figure 5.1D**). There was a positive (HNA) and negative (LNA) correlation between the fraction of HNA or LNA to total cells and productivity, but the relationship was weak and not significant ($R^2 = 0.14$, $p = 0.057$).

Association of OTUs to functional groups by Randomized Lasso regression

The relevance of specific OTUs for predicting freshwater FCM functional group abundance was assessed using the Randomized Lasso (RL) approach, which assigns a score between 0 (unimportant) to 1 (highly important) to each taxon in function of the target variable: HNAcc or LNAcc. This score can be interpreted as the probability that an OTU will be included in the Lasso model to predict HNA or LNA cell abundances. Variations of HNAcc and LNAcc were modelled in function of relative changes of OTUs. To address the negative correlation bias intrinsic to compositional data, compositions were first transformed using a centered log-ratio (CLR) transformation.

The RL score was used to implement a recursive variable elimination scheme. Specifically, we iteratively removed the lowest-ranked OTUs based on the RL score (*i.e.,* OTUs were ranked according to the score from high to low) and the Lasso was fitted to the data to predict HNAcc and LNAcc based on the corresponding subset of OTUs. The performance was expressed in terms of the R^2_{CV}, the R^2 between predicted and true values of HNAcc and LNAcc of samples that were held-out using a leave-one-group-out cross-validation scheme, in which samples were grouped according to year and location of measurement. If R^2_{CV} equals 1, predictions were equal to the true values, a value of 0 is equivalent to random guessing.

There was taxonomic dependency for both HNAcc and LNAcc across lake systems (**Figure 5.2**). R^2_{CV} increased when lower-ranked OTUs were removed (moving from right to left on **Figure 5.2**), which was gradual for the inland lakes (**Figure 5.2A**) and Muskegon Lake (**Figure 5.2C**) but was abrupt for Lake Michigan (**Figure 5.2B**). The number of taxa that resulted in the highest

R^2_{CV} contained less than a quarter of the total amount of taxa that were present (*see solid (HNA) and dotted (LNA) lines in* **Figure 5.2**), being 10.2% HNA and 15.3% LNA for the inland lakes, 4.0% HNA and 3.0% LNA for Lake Michigan, and 25.0% for both HNA and LNA in Muskegon Lake. This behavior was consistent for each lake system and FCM population. The Lake Michigan results differed the most from other lake systems, having the lowest R^2_{CV}, a sharp instead of gradual increase in R^2_{CV}, and a considerably lower minimal amount of OTUs (13 for HNAcc, 10 for LNAcc). No relationship could be established between rankings of variable selection methods and the relative abundance of individual OTUs (**Figure SI 5.7**). Multiple taxa with low average abundance were included in the minimal set of predictive variables, whereas few highly abundant OTUs were included. HNAcc and LNAcc could be predicted with equivalent performance to relative HNA and LNA proportions, yet the increase between initial and optimal performance was bigger (**Figure SI 5.8**). The final predictive performance was lower when compositional data was not transformed using the CLR-transformation (**Figure SI 5.9**).

Identification on different taxonomic levels: OTUs outperform all other taxonomic levels

To assess whether HNA and LNA groups were taxonomically conserved, compositional data was analyzed on all possible taxonomic levels for Muskegon Lake (**Figure 5.3**), using the same strategy as outlined in previous paragraph. The resulting R^2_{CV} values were considerably higher than zero on all taxonomic levels, meaning that at all levels individual taxonomic changes can be related to changes in HNAcc and LNAcc. Even though the OTU level resulted in the best prediction of HNAcc and LNAcc (**Figure 5.3**), each individual OTU has a lower RL score compared to other taxonomic levels, which on average became lower as the taxonomic level decreased (**Figure SI 5.10**). The fraction of variables (taxa) that could be removed to reach the maximum R^2_{CV} decreased as the taxonomic level became less resolved.

Validation of OTU selection results with the Boruta algorithm

The OTU results were validated with an additional variable selection strategy, called the Boruta algorithm. This approach allowed the further generalization of the findings presented above. In addition, it connects with Random Forest results from other studies, which have been described recently in microbiome studies of other systems *(see* (Ma et al. 2014) *and* (Chen et al. 2016)*).*

The Boruta algorithm selects relevant variables based on statistical hypothesis testing between the importance of an original variable and the importance of the most important permuted variable (*see materials and methods*), as retrieved from multiple Random Forest models. Selected variables are ranked as ‘1’, tentative variables as ‘2’, and all other variables get lower ranks, depending on the stage in which they were eliminated. The Boruta algorithm was applied for all three lake systems at the OTU-level, selected OTUs are visualized in **Figure SI 5.11**. The fraction of selected OTUs was always smaller than 1% across lake systems and functional groups (**Figure SI 5.12)**. The top scored OTU according to the RL was also selected according to the Boruta algorithm for HNAcc for all lake systems; for LNAcc both methods only agreed for Lake Michigan (**Table 5.1**). OTU060 (Proteobacteria; Sphingomonadales; alfIV_unclassified) was the only OTU selected in function of LNAcc across all lake systems, whereas no OTUs were selected across lake systems for HNAcc. As Random Forest regressions are the base method of the Boruta algorithm, we compared the predictive power of Boruta selected OTUs to those of all OTUs using Random Forest regression. For all lake systems and functional groups, performance increased when only selected OTUs were included in the model (**Table SI 5.2)**. Lasso predictions, in which OTUs were selected according to the RL, were better as opposed to Random Forest predictions in which OTUs were selected according to the Boruta algorithm (**Figure SI 13**). The fraction of selected OTUs according to the Boruta algorithm was lower than the optimal amount of OTUs according to the RL.

In this way, a number of findings could be generalized independent of a specific method: 1) Selected OTUs were mostly lake system specific, 2) a small fraction of OTUs was needed to predict changes in community composition, 3) selected OTUs are often rare and do not show a relationship with abundance, and 4) top RL-ranked HNA OTUs were also selected according to the Boruta algorithm, suggesting closer inspection of more closely the phylogeny of these taxa.

HNA- and LNA-associated OTUs differed across lake systems

Selected OTUs were mostly assigned to either the HNA or LNA groups and there was limited correspondence across lake systems between the selected OTUs (**Figure 5.4**). In Muskegon Lake, OTU173 (Bacteroidetes;Flavobacteriales;bacII-A) was selected as the major HNA-associated taxon while OTU29 (Bacteroidetes; Cytophagales; bacIII-B) had the highest RL score for LNA OTUs. In Lake Michigan, OTU25 (Bacteroidetes*;* Cytophagales; bacIII-A), was selected as the major HNA-associated taxon while OTU168 (Alphaproteobacteria: Rhizobiales: alfVII) was selected as a major LNA-associated taxon. For the inland lakes, OTU369 (Alphaproteobacterial; Rhodospirillales; alfVIII) was the major HNA-associated OTU while the OTU555 (Deltaproteobacteria;Bdellovibrionaceae;OM27) was the major LNA-associated taxon. Many more OTUs were selected in Muskegon Lake (197 OTUs; compared to 134 OTUs from the Inland Lakes and 21 OTUs from Lake Michigan) and these OTUs were often associated with both HNA and LNA groups.

RL scores were correlated for HNAcc and LNA within each lake system (Inland r = 0.25, P < 0.001; Michigan r = 0.59, P < 0.001, Muskegon r = 0.59, P < 0.001). Only OTUs that were present in all three freshwater environments were considered to calculate correlations between lake systems (190 in total, **Figure SI 5.14**). RL scores were lake ecosystem specific, with only a significant similarity between the Inland lakes and Muskegon lake using the RL for HNAcc (r = 0.21, P = 0.0042). Note that the correlation within a lake system therefore differs from previously reported values (as not all OTUs were considered), yet differences were small and results were comparable. The Boruta algorithm selected mostly OTUs which were unique both for the lake system and functional population (**Figure SI 5.11**).

Selected HNA and LNA OTUs do not have a phylogenetic signal

While many of the 258 OTUs selected by the RL were one of a few members of their phylum (*e.g.*, Firmicutes; Epsilonproteobacteria; OTU717 in Lentisphaerae; OTU267 in Omnitrophica; etc), the Bacteroidetes (60 OTUs), Betaproteobacteria (36 OTUs), Alphaproteobacteria (22 OTUs), and Verrucomicrobia (21 OTUs) were a total of 54% of the selected OTUs (**Figure 5.1**). Of these top four phyla, the majority of their membership were within the LNA group (41-52% of selected OTUs), with the minority of OTUs within the HNA group (14-30% of selected OTUs), and a quarter to a third of the OTUs were selected as members of both the LNA and HNA groups (23-36% of selected OTUs).

To evaluate how much phylogenetic history explains whether a selected taxon was associated with the HNA or LNA group(s), we calculated the phylogenetic signal, which is a measure of the dependence among species' trait values on their phylogenetic history (Revell et al. 2008). If the phylogenetic signal is very strong, taxa belonging to similar phylogenetic groups (*e.g.,* a Phylum) will share the same trait (*i.e.,* association with HNAcc or LNAcc). Alternatively, if the phylogenetic signal is weak, taxa within a similar phylogenetic group will have different traits. Pagel's lambda was used (Pagel 1999) to test for phylogenetic signal where lambda varies between 0 and 1. A lambda value of 1 indicates complete phylogenetic patterning whereas a lambda of 0 indicates no phylogenetic patterning and leads to a tree collapsing into a single polytomy. There was no phylogenetic signal with FCM functional group used as a discrete character (*i.e.* HNA, LNA, or Both) or as a continuous character using the RL scores for HNA (**Figure SI 5.15**; lambda = 0.16; P = 1). There was a significant LNA signal (p = 0.003) but the lambda value was 0.66, suggesting weak phylogenetic structuring in the LNA group. However, this significant result in the LNA was not replicated with other measures of phylogenetic signal (Blomberg's K (HNA: p = 0.63; LNA: p = 0.54), and Moran's I (HNA: p = 0.88; LNA: p = 0.12)) indicating that there is likely no phylogenetic signal in the taxa that drive the dynamics in either the HNA or the LNA group.

Flow cytometry fingerprints confirm associated taxa and reveal complex relationships between taxonomy and flow cytometric fingerprints

To confirm the association of the final selected OTUs with the HNA and LNA groups, we calculated the correlation between the density of individual regions (i.e., "bins") in the flow cytometry data with the relative abundances of the OTUs. The Kendall rank correlation coefficient between OTU abundances and counts in the flow cytometry fingerprint was calculated for each of the top HNA OTUs selected by the RL within each of the three systems. The correlation coefficient was visualized for each bin in the flow cytometry fingerprint (**Figure 5.6**). As these values denote correlations, they do not indicate actual presence. OTU25 correlated with almost the entire HNA region, whereas OTU173 was limited to the lower part of the HNA region. In contrast, OTU369 was positively correlated to both the LNA and HNA regions of the cytometric fingerprint, highlighting results from **Figure 5.4** where OTU369 was selected in function of both HNA and an LNA. The threshold that was used to define HNAcc and LNAcc lies very close to the actual corresponding regions.

Proteobacteria and rare taxa correlate with productivity measurements

The Kendall rank correlation coefficient was calculated between CLR-transformed abundances of individual OTUs and productivity measurements. OTU481 was significantly correlated after correction for multiple hypothesis testing using the Benjamini-Hochberg procedure (P < 0.001, P_adj = 0.016). This OTU had however a low RL score (0.022) and was not selected according to the Boruta algorithm. Of the top 10 OTUs according to the RL, three still had significant P-values (OTU614: P = 0.0064; OTU412, P = 0.044; OTU487, P = 0.014). Some OTUs that had a high RL score also had a positive correlation with productivity measurements (**Figure SI 5.16)**. At the phylum level, only Proteobacteria were significantly correlated to productivity measurements after Benjamini-Hochberg correction (P < 0.001, P_adj = 0.010).

Discussion

Our study introduces a novel computational workflow to investigate relationships between microbial diversity and ecosystem functioning. Specifically, studied the ecology of flow cytometric functional groups (i.e., HNA and LNA) by associating their dynamics with those of bacterial taxa (i.e., OTUs). We simultaneously collected flow cytometry and 16S rRNA gene sequencing data from three types of freshwater lake systems in the Great Lakes region, and bacterial heterotrophic productivity from one lake ecosystem, and used a machine learning based variable selection strategy, known as the Randomized Lasso, to associate one with the other. Our results showed that (1) there was a strong correlation between bacterial heterotrophic productivity and HNA cell abundances, (2) HNA and LNA cell abundances were best predicted by a small subset of OTUs that were unique to each lake type, (3) some OTUs were included in the best model for both HNA and LNA abundance, (4) there was no phylogenetic conservation of HNA and LNA group association, and (5) freshwater FCM fingerprints display more complex patterns related to OTUs and productivity compared to the traditional dichotomy of HNA and LNA. While HNA and LNA groups are universal across aquatic ecosystems, our data suggest that some bacterial taxa contribute to both HNA and LNA groups and that the taxa driving HNA and LNA abundance are unique to each lake system, supporting the fourth scenario in Bouvier et al. (2007).

Although high-nucleic acid cell counts (HNAcc) and low-nucleic acid cell counts (LNAcc) were correlated with each other, only the association between bacterial heterotrophic production (BP) and HNAcc was strong and significant. This correlation between BP and HNA is higher than previously reported values, although previous reports have focused on the proportion of HNA rather than absolute cell abundances with the majority of data collected from marine systems. For example, Bouvier et al. (2007) found a correlation between the fraction of HNA cells and BP within a large dataset of 640 samples across various freshwater to marine samples (r = 0.49), whereas a study off the coast of the Antarctic Peninsula found a moderate correlation (R^2 = 0.36; (Bowman et al. 2017)). Another study in the Bay of Biscay also found this association (R^2 = 0.16; (Morán et al. 2007)); however, the authors attributed this difference to be related to cell size and not due to the activity of HNA. Notably, these studies were predominantly testing the association of marine HNA and the reason for the stronger correlation in our study may be due to the nature of the freshwater samples. As such, future studies in freshwater environments should test the hypothesis of HNA taxa as driving forces of productivity, which is especially important for understanding the broader influence that HNA bacteria may have in the context of the disproportionately large role that freshwater systems play as hotspots in the global carbon cycle (Biddanda 2017). Finally, as our correlations with proportional HNA abundance also indicated less strong correlations than with absolute HNAcc, we suggest absolute HNAcc should be used to best predict heterotrophic bacterial production with FCM data.

The use of machine learning methods, such as the Lasso and Random Forest, are becoming more common in microbiome literature as these approaches are able to deal with multi-dimensional data and test the predictive power of a combined set of variables ((Baxter et al. 2014, Lin et al. 2014, Schubert et al. 2014). Although the Lasso already uses an intrinsic variable selection strategy, it has been noted that the Lasso method is not suited for compositional data because the regression coefficients have an unclear interpretation, and single variables may be selected when correlated to other variables (Li 2015). When performing variable selection with Random Forests, traditional variable importance measures such as the mean decrease in accuracy can be biased towards correlated variables (Strobl et al. 2008). Our approach included algorithms that extended these traditional machine learning algorithms, *i.e.*, the Randomized Lasso or Boruta algorithm (Kursa and Rudnicki 2010, Meinshausen and Bühlmann 2010). These methods make use of resampling and randomization that allow either assigning a probability of selection (RL) or statistically deciding which OTU to select (Boruta). Both the RL and Boruta algorithm have been applied to microbiome studies before. Examples for RL include the selection of genera in the gut microbiome inrelation to BMI (Lin et al. 2014) or the selection of OTUs from the oral microbiome in function of salivary pH and lysozyme activity (Zaura et al. 2017). The Boruta algorithm has been applied to select relevant genera, for example in the gut microbiome in relation to multiple sclerosis (Chen et al. 2016) or in the function of different diets during pregnancy of primates (Ma et al. 2014). Moreover, the Boruta algorithm has been recently proposed as one of the top-performing variable selection methods that make use of Random Forests (Degenhardt et al. 2017). The ability of our approach to identify unique sets of OTUs predictive of HNAcc and LNAcc despite the correlation between HNAcc and LNAcc (**Figure 5.1A**) illustrates the power of the machine learning based-variable selection methods. However, there is still room for improvement when attempting to integrate taxonomic and flow cytometry data. For example, 16S rRNA gene sequencing still faces the hurdles of DNA extraction (McCarthy et al. 2015) and 16S copy number bias (Louca et al. 2018). Moreover, detection limits are different for FCM (expressed in the number of cells) and 16S rRNA gene sequencing (expressed in the number of gene counts or relative abundance), which create data that may be different in resolution. Future work may focus on developing ways around these shortcomings to further improve the integration of FCM with 16S rRNA gene sequencing.

In our study, only a minority of OTUs was needed to predict specific flow cytometric group abundances. While each OTU individually had low predictive power, the selected group of OTUs was generally a strong predictor of HNAcc and LNAcc. In addition, the selected OTUs were often rare and thus no relationship could be established between the RL score and the abundance of an OTU (**Figure SI 5.9**). These results are in line with recent findings of Herren & McMahon (2018), who reported that a minority of low abundance taxa explained temporal compositional changes of microbial communities. The selection of different sets of HNA and LNA OTUs across the three freshwater systems indicates that different taxa underlie the universally observed HNA and LNA functional groups across aquatic systems. This is in line with strong species sorting in lake systems (Van der Gucht et al. 2007, Adams et al. 2014), shaping community composition through diverging environmental conditions between the lake systems presented here (Chiang et al. 2018). This high system specificity also explains the low RL scores for individual OTUs, as the spatial dynamics of an OTU diverged strongly across systems. (For example, an OTU that has an RL score of 0.5 implies that on average it will only be chosen one out of two times in a Lasso model).

Based on the high correlation of BP with HNAcc and low correlation of BP and LNAcc, the high proportion of LNA cells across all lake systems might indicate that the majority of cells in the bacterial community are dormant or have very low activity. This agrees with previous research showing that up to 40% (Jones and Lennon 2010) or even 64-95% (Zimmerman et al. 1978) of cells in freshwater systems are inactive or dormant. In fact, up to 60-80% of the OTUs in freshwater lakes have been reported to be dormant (Aanderud et al. 2016). Based on variable environmental conditions sampled across our dataset, some of the taxa that are predominantly dormant in one sample may contribute to activity in another sample. If this differing contribution to activity also covaries with a taxon’s abundance, these taxa may be considered to be ‘conditionally rare taxa’ (Jia et al. 2018) and previously 1-2% of freshwater lake OTUs have been reported to be conditionally rare (Shade et al. 2014). It has also been shown that marine heterotrophic bacteria can survive for at least 8 months (maximum tested length) in a starved state (Amy and Morita 1983). These factors may explain why some OTUs were included in both the HNAcc and LNAcc models and is in line with scenario 1 from Bouvier et al (2007) (*i.e.*, the transitioning of cells from active growth to death or inactivity). Alternatively, the same OTU may occur in both HNA and LNA groups due to phenotypic plasticity. Phenotypic plasticity has been shown for bacterial morphology and size, for example during predation and carbon starvation (Corno and Jürgens 2006). The fact that HNA and LNA groups have been suggested to correspond to cells of differing size, with HNA harboring larger cell sizes (Wang et al. 2009, Proctor et al. 2018), is in line with this hypothesis. Finally, the OTU level grouping of bacterial taxa can disguise genomic and phenotypic heterogeneity (Coleman et al. 2006, Hunt et al. 2008b, Denef et al. 2010a, Shapiro and Polz 2014), which may be an explanation for inconsistent associations between OTUs and FCM functional groups.

While all taxonomic levels resulted in a model with predictive power, the best model was at the most resolved taxonomy (*i.e.*, OTU) indicating that it is unlikely that OTUs within the HNA and LNA groups are phylogenetically conserved. Indeed, when analyzing the data at an OTU level, very little phylogenetic conservation was found between selected OTUs for HNA and LNA groups. This is in contrast to a recent study that found a clear phylogenetic signal at the phylum level (Proctor et al. 2018). Proctor et al. (Proctor et al. 2018) showed separate bacterial clusters between HNA and LNA groups across different aquatic systems. However, this was not the case for lake water samples. It is notable that Proctor et al. (Proctor et al. 2018) separated HNA and LNA cells based on cell size (where HNA cells were >0.4 um and LNA cells were 0.2-0.4 um, based on 50-90% removal of HNA cells after filtering), while our study separated these FCM functional groups on the basis of fluorescence intensity alone. Moreover, our study assessed associations between OTUs and population dynamics, while Proctor et al. (Proctor et al. 2018) assessed actual presence.

The Boruta algorithm and RL scores agreed on the top-ranked HNA OTU for all lake systems, which motivates further investigation of the ecology of these taxonomic groups. While little information on the identities of HNA and LNA freshwater lake bacterial taxa exists, several studies identified Bacteroidetes among the most prominent HNA taxa and this finding is in line with our results. Vila-Costa et al. (Vila-Costa et al. 2012) found that the HNA group was dominated by Bacteroidetes in summer samples from the Mediterranean Sea, Read et al. (Read et al. 2015) showed that HNA abundances correlated with Bacteroidetes, and Schattenhofer et al. (Schattenhofer et al. 2011) reported that the Bacteroidetes accounted for the majority of HNA cells in the North Atlantic Ocean. In Muskegon Lake, OTU173 was the dominant HNA taxon and is a member of the Order *Flavobacteriales* (bacII-A). The bacII group is a very abundant freshwater bacterial group and has been associated with senescence and decline of an intense algal bloom (Newton et al. 2011c). BacII-A has also made up ~10% of the total microbial community during cyanobacterial blooms, reaching its maximum density immediately following the bloom (Woodhouse et al. 2016). In Lake Michigan, OTU25, a member of the Bacteroidetes Order *Cytophagales* known as bacIII-A, was the top HNA OTU. However, much less is known about this specific group of Bacteroidetes. Though, the bacII-A/bacIII-A group has been strongly associated with more heterotrophically productive headwater sites (compared to higher order streams) from the River Thames, showing a negative correlation in rivers with dendritic distance from the headwaters, indicating that these taxa may contribute more to productivity (Read et al. 2015). In the inland lakes, OTU369 was the major HNA taxon and is associated with the Alphaproteobacteria Order Rhodospirillales (alfVIII), which to our knowledge is a group with very little information available in the literature. In contrast to our findings of Bacteroidetes and Alphaproteobacterial HNA selected OTUs, Tada & Suzuki (Tada and Suzuki 2016) found that the major HNA taxon from an oceanic algal culture was from the Betaproteobacteria whereas LNA OTUs were within the Actinobacteria phylum.

Conclusions

Our results indicate that there are taxonomic differences between HNA and LNA groups in freshwater lake systems, although these differences are lake system specific. This result may be due to taxa switching between these groups, potentially due to genomic or phenotypic plasticity. The difference between selected taxa is larger between lake systems as opposed to differences between HNA and LNA groups, which were not conserved phylogenetically. Thus, our results correspond most with research presented by Vila-Costa et al. (Vila-Costa et al. 2012), in which a taxonomic division was found between HNA and LNA groups, yet this was not rigid and followed seasonal trends. Overall, our results support scenario 4 proposed by Bouvier et al. (Bouvier et al. 2007), where HNA and LNA exhibit a different taxonomy, but this taxonomy changes over time and space and may overlap. With this study, we show that different types of microbial ecological data can be integrated with machine learning to learn about the composition and functioning of bacterial populations in aquatic systems. Future studies on HNA and LNA bacterial groups should use genome-resolved metagenomics, metatranscriptomics, or single-cell genomics to decipher whether the traits that underpin the association of a taxon with a FCM group are related to genomic or phenotypic plasticity.

Methods

Data collection and DNA extraction, sequencing and processing

In this study, we used a total of 173 samples collected from three types of lake systems described previously (Chiang et al. 2018), including: (1) 49 samples from Lake Michigan (2013 & 2015), (2) 62 samples from Muskegon Lake (2013-2015; one of Lake Michigan's estuaries), and (3) 62 samples from twelve inland lakes in Southeastern Michigan (2014-2015). For more details on sampling, please see **Figure 5.1** and the *Field Sampling, DNA extraction, and DNA sequencing and processing* sections within Chiang et al. (Chiang et al. 2018). In all cases, water for microbial biomass samples were collected and poured through a 210 μm and 20 μm bleach sterilized nitex mesh and sequential in-line filtration was performed using 47 mm polycarbonate in-line filter holders (Pall Corporation, Ann Arbor, MI, USA) and an E/S portable peristaltic pump with an easy-load L/S pump head (Masterflex®, Cole Parmer Instrument Company, Vernon Hills, IL, USA) to filter first through a 3 μm isopore polycarbonate (TSTP, 47 mm diameter, Millipore, Billerica, MA, USA) and second through a 0.22 μm Express Plus polyethersulfone membrane filters (47 mm diameter, Millipore, MA, USA). The current study only utilized the 3 - 0.22 μm fraction for analyses.

DNA extractions and sequencing were performed as described in Chiang et al. (Chiang et al. 2018). Fastq files were submitted to NCBI sequence read archive under BioProject accession number PRJNA412984 and PRJNA414423. We analyzed the sequence data using MOTHUR V.1.38.0 (seed = 777; (Schloss et al. 2009b) based on the MiSeq standard operating procedure and put together at the following link: https://github.com/rprops/Mothur_oligo_batch. A combination of the Silva Database (release 123; (Quast et al. 2013b)) and the freshwater TaxAss 16S rRNA database and pipeline (Rohwer et al. 2017b) was used for classification of operational taxonomic units (OTUs).

For the taxonomic analysis, each of the three lake datasets were analyzed separately and treated with an OTU abundance threshold cutoff of at least 5 sequences in 10% of the samples in the dataset (similar strategy to (Weiss et al. 2016)). For comparison of taxonomic abundances across samples, each of the three datasets were then rarefied to an even sequencing depth, which was 4,491 sequences for Muskegon Lake samples, 5,724 sequences for the Lake Michigan samples, and 9,037 sequences for the inland lake samples. Next, the relative abundance at the OTU level was calculated using the *transform_sample_counts()* function in the phyloseq R package (McMurdie and Holmes 2013) by taking the count value and dividing it by the sequencing depth of the sample. For all other taxonomic levels, the taxonomy was merged at certain taxonomic ranks using the *tax_glom()* function in phyloseq (McMurdie and Holmes 2013) and the relative abundance was re-calculated.

Heterotrophic bacterial production measurements

Muskegon Lake samples from 2014 and 2015 were processed for heterotrophic bacterial production using the [^{3}H] leucine incorporation into bacterial protein in the dark method (Kirchman et al. 1985, Simon and Azam 1989). At the end of the incubation with [^{3}H]-leucine, cold trichloroacetic acid-extracted samples were filtered onto 0.2 μm filters that represented the leucine incorporation by the bacterial community. Measured leucine incorporation during the incubation was converted to bacterial carbon production rate using a standard theoretical conversion factor of 2.3 kg C per mole of leucine (Simon and Azam 1989).

Flow cytometry, measuring HNA and LNA

In the field, a total of 1 mL of 20 μm filtered lake water were fixed with 5 μL of glutaraldehyde (20% vol/vol stock), incubated for 10 minutes on the bench (covered with aluminum foil to protect from light degradation), and then flash frozen in liquid nitrogen to later be stored in -80°C freezer until later processing with a flow cytometer. Flow cytometry procedures followed the protocol laid out in Props et al. (Props et al. 2017b), which also uses the samples presented in the current study. Samples were stained with SYBR Green I and measured in triplicate. The lowest number of cells collected after denoising was 2342. HNA and LNA groups were selected using the fixed gates introduced in Prest et al. (Prest et al. 2013) and plotted in **Figure SI 5.17**. Cell counts were determined per HNA and LNA group and averaged over the three replicates (giving rise to HNAcc and LNAcc).

Data analysis

Processed data and analysis code for the following analyses can be found on the GitHub page for this project at https://deneflab.github.io/HNA_LNA_productivity/.

HNA-LNA and HNA-Productivity Statistics and Regressions

We tested the difference in absolute number of cells within HNA and LNA functional groups across running analysis of variance with a post-hoc Tukey HSD test (*aov()* and *TukeyHSD(); stats* R package; (R Core Team 2018)). In addition, we tested the association of HNA and LNA to each other and with productivity by running ordinary least squares regression with the *lm()* (*stats* R package; (R Core Team 2018)).

Ranking correlation

Ranking correlation between variables was calculated using the Kendall rank correlation coefficient, using the *kendalltau()* function in Scipy (v1.0.0) or *cor()* in R (v3.2). The 'tau-b' implementation was used, which is able to deal with ties. Values range from -1 (strong disagreement) to 1 (strong agreement). The same statistic was used to assess the similarity between rankings of variable selection methods.

Centered-log ratio transform

First, following guidelines from Paliy & Shanker, Gloor et al. and Quinn et al.(Paliy and Shankar 2016, Gloor et al. 2017, Quinn et al. 2018), relative abundances of OTUs were transformed using a centered log-ratio (CLR) transformation before variable selection was applied. This means that the relative abundance $x_i$of a taxa was transformed according to the geometric mean of that sample, in which there are p taxa present:

$$x_i' = \log(x_i/(\prod_{j=1}^{p} x_j)^{1/p}) .$$

Zero values were replaced by $\delta = 1/p^2$. This was done using the scikit-bio package (www.scikit-bio.org, v0.4.1).

Lasso & stability selection

Scores were assigned to taxa based on an extension of the Lasso estimator, which is called *stability selection* (Meinshausen and Bühlmann 2010). In the case of $n$samples, the Lasso estimator fits the following regression model:

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} ||y - X\beta||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$,

in which X denotes the abundance table, y the target to predict, which either is HNA cell abundances (HNAcc) or LNA cell abundances (LNAcc), and λ is a regularization parameter which controls the complexity of the model and prevents overfitting. The Lasso performs an intrinsic form of variable selection, as the weights of certain variables will be put to zero.

Stability selection, when applied to the Lasso, is in essence an extension of the Lasso regression. It implements two types of randomizations to assign a score to the variables, and is therefore also called the *Randomized Lasso* (RL). The resulting RL score can be seen as the probability that a certain variable will be included in a Lasso regression model (*i.e.*, its weight will be non-zero when fitted). When performing stability selection, the Lasso is fitted to B different subsamples of the data of fraction $n/2$, denoted as X' and corresponding y'. A second randomization is added by introducing a weakness parameter α. In each model, the penalty λ changes to a randomly chosen value in the set $[\lambda, \lambda/\alpha]$, which means that a higher penalty will be assigned to a random subset of the total amount of variables. The Randomized Lasso therefore becomes:

$$\hat{\beta}^{\lambda} = \text{argmin}_{\beta \in \mathbb{R}^p} ||y' - X'\beta||_2^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{w_j}$$,

where w_j is a random variable which is either α or 1. Next, the Randomized Lasso score (RL score) is determined by counting the number of times the weight of a variable was non-zero for each of the B models and divided by B. Meinshausen and Bühlmann show that, under stringent conditions, the number of falsely selected variables is controlled for the Randomized Lasso when the RL score is higher than 0.5. If λ is varied, one can determine the stability path, which is the relationship between π and λ for every variable. For our implementation, $B = 500$, $\alpha = 0.5$ and the highest score was selected in the stability path for which λ ranged from 10^{-3} until 10^3, logarithmically divided in 100 intervals. The *RandomizedLasso()* function from the scikit-learn machine learning library was used (Pedregosa et al. 2011), v0.19.1).

Random Forests & Boruta

The Boruta algorithm is a *wrapper* algorithm that makes use of Random Forests as a base classification or regression method in order to select all relevant variables in function of a response variable (Kursa and Rudnicki 2010). Similar to stability selection, the method uses an additional form of randomness in order to perform variable selection. Random Forests are fitted to the data multiple times. To remove the correlation to the response variable, each variable gets per iteration a so-called *shadow variable*, which is a permuted copy of the original variable. Next, the Random Forest algorithm is run with the extended set of variables, after which variable importances are calculated for both original and shadow variables. The shadow variable that has the highest importance score is used as reference, and every variable with significantly lower importance, as determined by a Bonferroni corrected t-test, is removed. Likewise, variables containing an importance score that is significantly higher are included in the final list of selected variables. This procedure can be repeated until all original variables are either discarded or included in the final set; variables that remain get the label ‘tentative’ (i.e., after all repetitions it is still not possible to either select or discard a certain variable). We used the boruta_py package to implement the Boruta algorithm (https://github.com/scikit-learn-contrib/boruta_py). Random Forests were implemented using *RandomForestRegressor()* function from scikit-learn (Pedregosa et al. 2011), v0.19.1). Random Forests were run with 200 trees, the number of variables considered at every split of a decision tree was $p/3$ and the minimal number of samples per leaf was set to five. The latter were based on default values for Random Forests in a regression setting (Probst et al. 2018). The Boruta algorithm was run for 300 iterations, variables were selected or discarded at $P < .05$ after performing Bonferroni correction.

Recursive variable elimination

Scores of the Randomized Lasso were evaluated using a recursive variable elimination strategy (Guyon et al. 2002). Variables were ranked according to the RL score. Next, the lowest-ranked variables were eliminated from the dataset, after which the Lasso was applied to predict HNAcc and LNAcc respectively. This process was repeated until only the highest-scored taxa remained. In this way, performance of the Randomized Lasso was assessed from a minimal-optimal evaluation perspective (Nilsson et al. 2007). In other words, the fewest variables that resulted in the highest predictive performance was determined.

Performance evaluation

In order to account for the spatiotemporal structure of the data, a blocked cross-validation scheme was implemented (Roberts et al. 2017). Samples were grouped according the site and year that they were collected. This results in 5, 10 and 16 distinctive groups for the Michigan, Muskegon and Inland lake systems respectively. Predictive models were optimized in function of the R^2 between predicted and true values of held-out groups using a leave-one-group-out cross-validation scheme with the *LeaveOneGroupOut()* function. This results in a cross-validated R^2_{CV} value. For the Lasso, λ was determined using the lassoCV() function, with setting eps=10^{-4} and n_alphas=400. The Random Forest object was optimized using a grid search where max_features was chosen in the interval $[1, \sqrt{p}, 2\sqrt{p}, ..., p]$ (all variables) or $[1, ..., p]$ (Boruta selected variables) and min_samples_leaf in the interval $[1, ..., 5]$, using the *GridSearchCV()* function. The number of decision trees (n_trees) was set to 200. All functions are part of scikit-learn ((Pedregosa et al. 2011); v0.19.1)

Stability of the Randomized Lasso

Similarity of RL scores between lake systems and functional groups was quantified using the Pearson correlation. This was done using the *pearsonr()* function in Scipy (v1.0.0).

Patterns of HNA and LNA OTUs across ecosystems and phylogeny

To visualize patterns of selected HNA and LNA OTUs across the three ecosystems, a heatmap was created with the RL scores of each OTU from the Randomized Lasso regression that were higher than specified threshold values. The heatmap was created with the *heatmap.2()* function (*gplots* R package) using the euclidean distances of the RL scores and a complete linkage hierarchical clustering algorithm (**Figure 5.4**).

Correlations between taxa and productivity measurements

Kendall tau ranking correlations between productivity measurements and individual abundances were calculated on the phylum and OTU level using the *kendalltau()* function from Scipy (v1.0.0). P-values were corrected using Benjamini-Hochberg correction, reported as P_adj. This was done using the *multitest()* function from the Python module Statsmodels ((Seabold and Perktold 2010); v0.5.0).

Phylogenetic tree construction and signal calculation

We calculated the best performing maximum likelihood tree using the GTR-CAT model (-gtr -fastest) model of nucleotide substitution with fasttree (version 2.1.9 No SSE3; (Price et al. 2010)). Phylogenetic signal with both discrete (*i.e.*, HNA, LNA, or both) and continuous traits (*i.e.* the RL score) using the newick tree from FastTree was then used to model phylogenetic signal using Pagel's lambda (discrete trait: fitDiscrete() from the geiger R package (Harmon et al. 2008); continuous trait: phylosig() from the phytools R (Revell 2012)), Blomberg's K (phylosig() function from the phytools R package (Revell 2012)), and Moran's I (abouheif.moran() function from the adephylo R package (Jombart et al. 2010)).

Acknowledgements

PR was supported by Ghent University (BOFSTA2015000501) and MLS was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. DGE 1256260). Part of the computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, the Hercules Foundation and the Flemish Government department EWI. Flow cytometry analysis was supported through a Geconcerteerde Onderzoeksactie (GOA) from Ghent University (BOF15/GOA/006). Support for the sampling of Muskegon Lake was provided by a Community Foundation for Muskegon County grant to the Annis Water Resources Institute.

References

Aanderud, Z. T., J. C. Vert, J. T. Lennon, T. W. Magnusson, D. P. Breakwell, and A. R. Harker. 2016. Bacterial dormancy is more prevalent in freshwater than hypersaline lakes. Frontiers in Microbiology 7:1–13.

Adams, H. E., B. C. Crump, and G. W. Kling. 2014. Metacommunity dynamics of bacteria in an arctic lake: The impact of species sorting and mass effects on bacterial production and biogeography. Frontiers in Microbiology 5:1–10.

Amy, P. S., and R. Y. Morita. 1983. Starvation-survival patterns of sixteen freshly isolated open-ocean bacteria. Applied and Environmental Microbiology 45:1109–1115.

Arnoldini, M., T. Heck, A. Blanco-Fernández, and F. Hammes. 2013. Monitoring of Dynamic Microbiological Processes Using Real-Time Flow Cytometry. PLoS ONE 8:e80117.

Baxter, N. T., J. P. Zackular, G. Y. Chen, and P. D. Schloss. 2014. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. Microbiome 2:1–11.

Biddanda, B. A. 2017. Global Significance of the Changing Freshwater Carbon Cycle Emerging Role of Freshwater in the Global Theater. Eos 98:1–5.

Bouvier, T., P. A. Del Giorgio, and J. M. Gasol. 2007. A comparative study of the cytometric characteristics of High and Low nucleic-acid bacterioplankton cells from different aquatic ecosystems. Environmental Microbiology 9:2050–2066.

Bowman, J. S., L. A. Amaral-zettler, J. J. Rich, C. M. Luria, and H. W. Ducklow. 2017. Bacterial community segmentation facilitates the prediction of ecosystem function along the coast of the western Antarctic Peninsula. ISME Journal 11:1460–1471.

Carini, P., P. J. Marsden, J. W. Leff, E. E. Morgan, M. S. Strickland, and N. Fierer. 2016. Relic DNA is abundant in soil and obscures estimates of soil microbial diversity. Nature Microbiology 2:16242.

Chen, J., N. Chia, K. R. Kalari, J. Z. Yao, M. Novotna, M. M. P. Soldan, D. H. Luckey, E. V. Marietta, P. R. Jeraldo, X. Chen, B. G. Weinshenker, M. Rodriguez, O. H. Kantarci, H. Nelson, J. A. Murray, and A. K. Mangalam. 2016. Multiple sclerosis patients have a distinct gut microbiota compared to healthy controls. Scientific Reports 6:1–10.

Chiang, E., M. L. Schmidt, M. A. Berry, B. A. Biddanda, A. Burtner, T. H. Johengen, D. Palladino, and V. J. Denef. 2018. Verrucomicrobia are prevalent in north- temperate freshwater lakes and display class- level preferences between lake habitats. PLoS ONE 13:1–20.

Coleman, M. L., M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. Delong, S. W. Chisholm, M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. Delong, and S. W. Chisholmlt. 2006. Genomic Islands and the Ecology and Evolution of Prochlorococcus. Science 311:1768–1770.

Corno, G., and K. Jürgens. 2006. Direct and indirect effects of protist predation on population size structure of a bacterial strain with high phenotypic plasticity. Applied and Environmental Microbiology 72:78–86.

Degenhardt, F., S. Seifert, and S. Szymczak. 2017. Evaluation of variable selection methods for random forests and omics data sets. Briefings in Bioinformatics:1–12.

Denef, V. J., L. H. Kalnejais, R. S. Mueller, P. Wilmes, B. J. Baker, B. C. Thomas, N. C. VerBerkmoes, R. L. Hettich, and J. F. Banfield. 2010. Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. Proceedings of the National Academy of Sciences of the United States of America 107:2383–2390.

Gasol, J. M., and P. A. Del Giorgio. 2000. Using flow cytometry for counting natural planktonic bacteria and understanding the structure of planktonic bacterial communities. Scientia Marina 64:197–224.

Gasol, J. M., U. L. Zweifel, F. Peters, A. Jed, U. L. I. Zweifel, and J. E. D. A. Fuhrman. 1999. Significance of Size and Nucleic Acid Content Heterogeneity as Measured by Flow Cytometry in Natural Planktonic Bacteria Significance of Size and Nucleic Acid Content Heterogeneity as Measured by Flow Cytometry in Natural Planktonic Bacteria. Applied and Environmental Microbiology 65:4475–4483.

Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue. 2017. Microbiome datasets are compositional: And this is not optional. Frontiers in Microbiology 8:1–6.

Van der Gucht, K., K. Cottenie, K. Muylaert, N. Vloemans, S. Cousin, S. Declerck, E. Jeppesen, J.-M. Conde-Porcuna, K. Schwenk, G. Zwart, H. Degans, W. Vyverman, and L. De Meester. 2007. The power of species sorting: local factors drive bacterial community composition over a wide range of spatial scales. Proceedings of the National Academy of Sciences of the United States of America 104:20404–20409.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik. 2002. Gene Selection for Cancer Classification using Support Vector Machines. Machine learning 46:389–422.

Harmon, L. J., J. T. Weir, C. D. Brock, R. E. Glor, and W. Challenger. 2008. GEIGER: Investigating evolutionary radiations. Bioinformatics 24:129–131.

Herren, C. M., and K. D. McMahon. 2018. Keystone taxa predict compositional change in microbial communities. Environmental Microbiology:1–34.

Hunt, D. E., L. a David, D. Gevers, S. P. Preheim, E. J. Alm, and M. F. Polz. 2008. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320:1081–1085.

Jia, X., F. Dini-Andreote, and J. Falcão Salles. 2018. Community Assembly Processes of the Microbial Rare Biosphere. Trends in Microbiology xx:1–10.

Jochem, F. J., P. J. Lavrentyev, and M. R. First. 2004. Growth and grazing rates of bacteria groups with different apparent DNA content in the Gulf of Mexico. Marine Biology 145:1213–1225.

Jombart, T., F. Balloux, and S. Dray. 2010. adephylo: New tools for investigating the phylogenetic signal in biological traits. Bioinformatics 26:1907–1909.

Jones, S. E., and J. T. Lennon. 2010. Dormancy contributes to the maintenance of microbial diversity. Proceedings of the National Academy of Sciences 107:5881–5886.

Kirchman, D., E. K’nees, and R. Hodson. 1985. Leucine incorporation and its potential as a measure of protein synthesis by bacteria in natural aquatic systems. Applied and Environmental Microbiology 49:599–607.

Kursa, M. B., and W. R. Rudnicki. 2010. Feature Selection with the Boruta Package. Journal Of Statistical Software 36:1–13.

Lebaron, P., P. Servais, H. Agogué, C. Courties, and F. Joux. 2001. Does the High Nucleic Acid Content of Individual Bacterial Cells Allow Us to Discriminate between Active Cells and Inactive Cells in Aquatic Systems? Applied and Environmental Microbiology 67:1775–1782.

Lebaron, P., P. Servais, a.-C. Baudoux, M. Bourrain, C. Courties, and N. Parthuisot. 2002. Variations of bacterial-activity with cell size and nucleic acid content assessed by flow cytometry. Aquat. Microb. Ecol. 28:131–140.

Lennon, J. T., and S. E. Jones. 2011. Microbial seed banks: the ecological and evolutionary implications of dormancy. Nature reviews. Microbiology 9:119–30.

Li, H. 2015. Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis. Annual Review of Statistics and Its Application 2:73–94.

Lin, W., P. Shi, R. Feng, and H. Li. 2014. Variable selection in regression with compositional covariates. Biometrika 101:785–797.

Louca, S., M. Doebeli, and L. W. Parfrey. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6:1–12.

Ma, J., A. L. Prince, D. Bader, M. Hu, R. Ganu, K. Baquero, P. Blundell, R. Alan Harris, A. E. Frias, K. L. Grove, and K. M. Aagaard. 2014. High-fat maternal diet during pregnancy persistently alters the offspring microbiome in a primate model. Nature Communications 5:1–11.

McCarthy, A., E. Chiang, M. L. Schmidt, and V. J. Denef. 2015. RNA Preservation Agents and Nucleic Acid Extraction Method Bias Perceived Bacterial Community Composition. Plos One 10:e0121659.

McMurdie, P. J., and S. Holmes. 2013. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS ONE 8:e61217.

Meinshausen, N., and P. Bühlmann. 2010. Stability selection. Journal of the Royal Statistical Society. Series B: Statistical Methodology 72:417–473.

Morán, X. A. G., H. W. Ducklow, and M. Erickson. 2011. Single-cell physiological structure and growth rates of heterotrophic bacteria in a temperate estuary (Waquoit Bay, Massachusetts). Limnology and Oceanography 56:37–48.

Morán, X., A. Bode, L. Suárez, and E. Nogueira. 2007. Assessing the relevance of nucleic acid content as an indicator of marine bacterial activity. Aquatic Microbial Ecology 46:141–152.

Newton, R. J., S. E. Jones, A. Eiler, K. D. McMahon, and S. Bertilsson. 2011. A guide to the natural history of freshwater lake bacteria. Page Microbiology and molecular biology reviews.

Nilsson, R., J. M. Peña, J. Björkegren, and J. Tegnér. 2007. Consistent Feature Selection for Pattern Recognition in Polynomial Time. The Journal of Machine Learning Research 8:589–612.

Pagel, M. 1999. Inferring the historical patterns of biological evolution. Nature 401:877–884.

Paliy, O., and V. Shankar. 2016. Application of multivariate statistical techniques in microbial ecology. Molecular Ecology 25:1032–1057.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. JOURNAL OF MACHINE LEARNING RESEARCH 12:2825–2830.

Prest, E. I., F. Hammes, S. Kötzsch, M. C. M. van Loosdrecht, and J. S. Vrouwenvelder. 2013. Monitoring microbiological changes in drinking water systems using a fast and reproducible flow cytometric method. Water Research 47:7131–7142.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE 5.

Probst, P., M. Wright, and A.-L. Boulesteix. 2018. Hyperparameters and Tuning Strategies for Random Forest. arXiv:preprint.

Proctor, C. R., M. D. Besmer, T. Langenegger, K. Beck, J.-C. Walser, M. Ackermann, H. Bürgmann, and F. Hammes. 2018. Phylogenetic clustering of small low nucleic acid-content bacteria across diverse freshwater ecosystems. The ISME Journal.

Props, R., M. L. Schmidt, J. Heyse, H. A. Vanderploeg, N. Boon, and V. J. Denef. 2017. Flow cytometric monitoring of bacterioplankton phenotypic diversity predicts high population-specific feeding rates by invasive dreissenid mussels. Environmental Microbiology 00.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Research 41:590–596.

Quinn, T. P., I. Erb, M. F. Richardson, and T. M. Crowley. 2018. Understanding sequencing data as compositions: an outlook and review. Bioinformatics:1–9.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. Vienna, Austria.

Ramseier, M. K., U. von Gunten, P. Freihofer, and F. Hammes. 2011. Kinetics of membrane damage to high (HNA) and low (LNA) nucleic acid bacterial clusters in drinking water by ozone, chlorine, chlorine dioxide, monochloramine, ferrate(VI), and permanganate. Water Research 45:1490–1500.

Read, D. S., H. S. Gweon, M. J. Bowes, L. K. Newbold, D. Field, M. J. Bailey, and R. I. Griffiths. 2015. Catchment-scale biogeography of riverine bacterioplankton. ISME Journal 9:516–526.

Revell, L. J. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3:217–223.

Revell, L. J., L. J. Harmon, and D. C. Collar. 2008. Phylogenetic signal, evolutionary process, and rate. Systematic biology 57:591–601.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929.

Rohwer, R. R., J. J. Hamilton, R. J. Newton, and K. D. McMahon. 2017. TaxAss: Leveraging Custom Databases Achieves Fine-Scale Taxonomic Resolution. bioRxiv:214288.

Schattenhofer, M., J. Wulf, I. Kostadinov, F. O. Glöckner, M. V. Zubkov, and B. M. Fuchs. 2011. Phylogenetic characterisation of picoplanktonic populations with high and low nucleic acid content in the North Atlantic Ocean. Systematic and Applied Microbiology 34:470–475.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, and C. F. Weber. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology 75:7537–7541.

Schubert, A. M., M. a M. Rogers, C. Ring, J. Mogle, J. P. Petrosino, V. B. Young, D. M. Aronoff, and P. D. Schloss. 2014. Microbiome Data Distinguish Patients with Clostridium difficile Infection and Non- C . difficile -Associated Diarrhea from Healthy. mBio 5:1–9.

Seabold, S., and J. Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. Proc of the 9th Python in Science Conf.:57–61.

Servais, P., E. O. Casamayor, C. Courties, P. Catala, N. Parthuisot, and P. Lebaron. 2003. Activity and diversity of bacterial cells with high and low nucleic acid content. Aquatic Microbial Ecology 33:41–51.

Servais, P., C. Courties, P. Lebaron, and M. Troussellier. 1999. Coupling bacterial activity measurements with cell sorting by flow cytometry. Microbial Ecology 38:180–189.

Shade, A., S. E. Jones, J. G. Caporaso, J. Handelsman, R. Knight, N. Fierer, and J. a Gilbert. 2014. Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. mBio 5:e01371-14.

Shapiro, B. J., and M. F. Polz. 2014. Ordering microbial diversity into ecologically and genetically cohesive units. Trends in Microbiology 22:235–247.

Sherr, E. B., B. F. Sherr, and K. Longnecker. 2006. Distribution of bacterial abundance and cell-specific nucleic acid content in the Northeast Pacific Ocean. Deep-Sea Research Part I: Oceanographic Research Papers 53:713–725.

Simon, M., and F. Azam. 1989. Protein content and protein synthesis rates of planktonic marine bacteria. Marine Ecology Progress Series 51:201–213.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. BMC bioinformatics 9:307.

Tada, Y., and K. Suzuki. 2016. Changes in the community structure of free-living heterotrophic bacteria in the open tropical Pacific Ocean in response to microalgal lysate-derived dissolved organic matter. FEMS Microbiology Ecology 92:1–13.

Vila-Costa, M., J. M. Gasol, S. Sharma, and M. A. Moran. 2012. Community analysis of high- and low-nucleic acid-containing bacteria in NW Mediterranean coastal waters using 16S rDNA pyrosequencing. Environmental Microbiology 14:1390–1402.

Vives-Rego, J., P. Lebaron, and Caron Nebe-von. 2000. Current and future applications of flow cytometry in aquatic microbiology. FEMS microbiology reviews 24:429–448.

Wang, Y., F. Hammes, N. Boon, M. Chami, and T. Egli. 2009. Isolation and characterization of low nucleic acid (LNA)-content bacteria. ISME Journal 3:889–902.

Wang, Y., F. Hammes, K. De Roy, W. Verstraete, and N. Boon. 2010. Past, present and future applications of flow cytometry in aquatic microbiology. Trends in Biotechnology 28:416–424.

Weiss, S., W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight. 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME Journal 10:1669–1681.

Widder, S., R. J. Allen, T. Pfeiffer, T. P. Curtis, C. Wiuf, W. T. Sloan, O. X. Cordero, S. P. Brown, B. Momeni, W. Shou, H. Kettle, H. J. Flint, A. F. Haas, B. Laroche, J. U. Kreft, P. B. Rainey, S. Freilich, S. Schuster, K. Milferstedt, J. R. Van Der Meer, T. Grobkopf, J. Huisman, A. Free, C. Picioreanu, C. Quince, I. Klapper, S. Labarthe, B. F. Smets, H. Wang, O. S. Soyer, S. D. Allison, J. Chong, M. C. Lagomarsino, O. A. Croze, J. Hamelin, J. Harmand, R. Hoyle, T. T. Hwa, Q. Jin, D. R. Johnson, V. de Lorenzo, M. Mobilia, B. Murphy, F. Peaudecerf, J. I. Prosser, R. A. Quinn, M. Ralser, A. G. Smith, J. P. Steyer, N. Swainston, C. E. Tarnita, E. Trably, P. B. Warren, and P. Wilmes. 2016. Challenges in microbial ecology: Building predictive understanding of community function and dynamics. ISME Journal 10:2557–2568.

Woodhouse, J. N., A. S. Kinsela, R. N. Collins, L. C. Bowling, G. L. Honeyman, J. K. Holliday, and B. A. Neilan. 2016. Microbial communities reflect temporal changes in cyanobacterial composition in a shallow ephemeral freshwater lake. ISME Journal 10:1337–1351.

Zaura, E., B. W. Brandt, A. Prodan, M. J. Teixeira De Mattos, S. Imangaliyev, J. Kool, M. J. Buijs, F. L. Jagers, N. L. Hennequin-Hoenderdos, D. E. Slot, E. A. Nicu, M. D. Lagerweij, M. M. Janus, M. M. Fernandez-Gutierrez, E. Levin, B. P. Krom, H. S. Brand, E. C. Veerman, M. Kleerebezem, B. G. Loos, G. A. Van Der Weijden, W. Crielaard, and B. J. Keijser. 2017. On the ecosystemic network of saliva in healthy young adults. ISME Journal 11:1218–1231.

Zimmerman, R., R. Iturriaga, and J. Becker-Birck. 1978. Simultaneous determination of the total number of aquatic bacteria and the number thereof involved in respiration. Applied and Environmental Microbiology 36:926–935.

Table 5.1. Top-scoring OTUs according to the randomized lasso.

Top-scoring OTUs according to the RL per functional population and lake ecosystem. Selection according to the Boruta algorithm is given in addition to the RL score. Descriptive statistics by means of the Kendall rank correlation coefficient (KRCC) have been added with level of significance in function of the HNA/LNA population. Full taxonomy of the OTUs is given in **Table SI 5.3.**

Lake system	**Functional group**	**OTU**	**RL score**	**Boruta selected**	**Kendall's tau (HNA)**	**P-value (HNA)**	**Kendall's tau (LNA)**	**P-value (LNA)**
Inland	HNA	OTU369	0.382	yes	-0.43	<0.001	-0.28	0.0012
	LNA	OTU555	0.384	no	0.089	N.S.	0.22	0.011
Michigan	HNA	OTU025	0.362	yes	0.46	<0.001	0.41	<0.001
	LNA	OTU168	0.428	yes	0.26	0.0092	0.4	<0.001
Muskegon	HNA	OTU173	0.462	yes	0.5	<0.001	0.2	0.019
	LNA	OTU029	0.568	no	0.26	0.0029	0.49	<0.001

Figure 5.1. Association of flow cytometry functional groups with each other and productivity.

(A) Correlation between HNA cell counts and LNA cell counts across the three freshwater lake ecosystems. **(B-D)** Muskegon Lake bacterial heterotrophic production and its correlation with **(B)** HNA cell counts, **(C)** LNA cell counts, and **(D)** total cell counts. The grey area in plots A, B, and D represents the 95% confidence intervals.

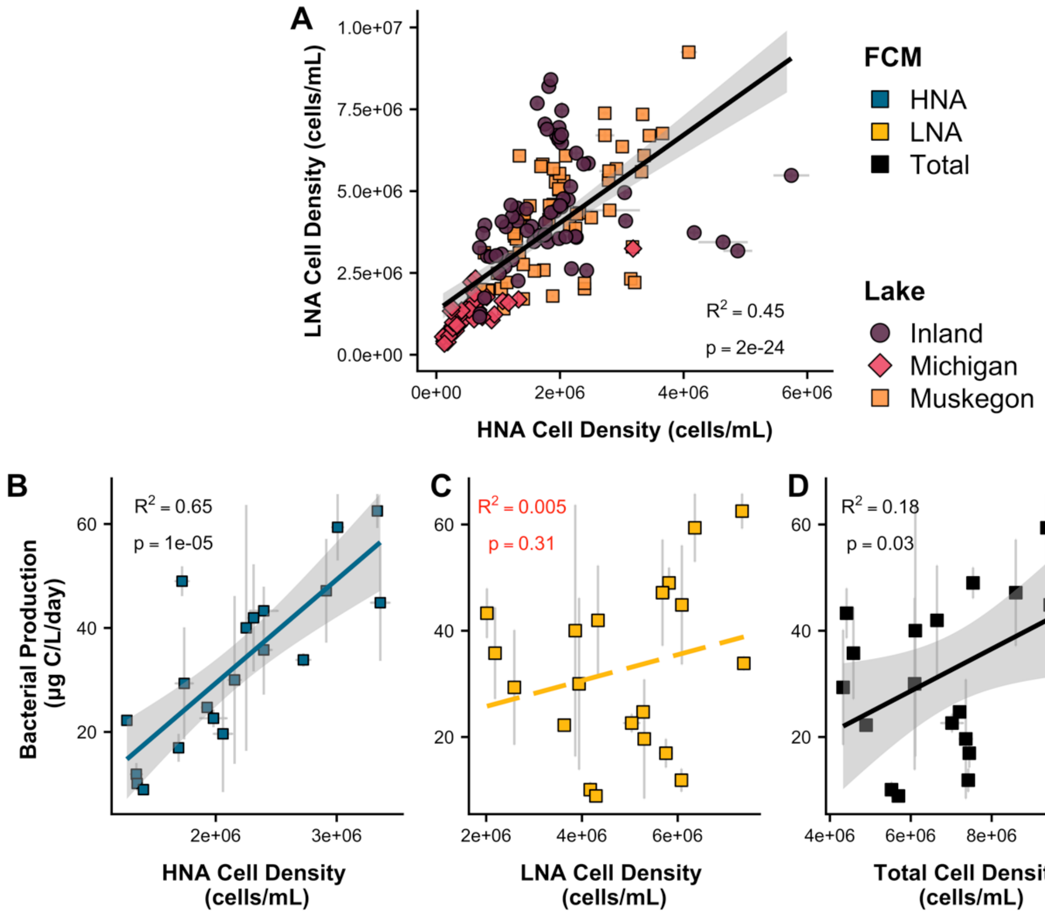


Figure 5.2. R^2_{CV} in function of the number of OTUs, which were iteratively removed based on the RL score and evaluated using the Lasso at every step.

The solid (HNA) and dashed (LNA) vertical lines corresponds to the threshold (i.e., number of OTUs) which resulted in a maximal R^2_{CV}. **(A)** Inland system ($R^2_{CV,max} = 0.92$), HNAcc; **(B)** Lake Michigan ($R^2_{CV,max} = 0.53$), HNAcc; **(C)** Muskegon lake, HNAcc ($R^2_{CV,max} = 0.85$); **(D)** Inland system, LNAcc ($R^2_{CV,max} = 0.87$); **(E)** Lake Michigan, LNAcc ($R^2_{CV,max} = 0.79$); **(F)** Muskegon lake, LNAcc ($R^2_{CV,max} = 0.91$).

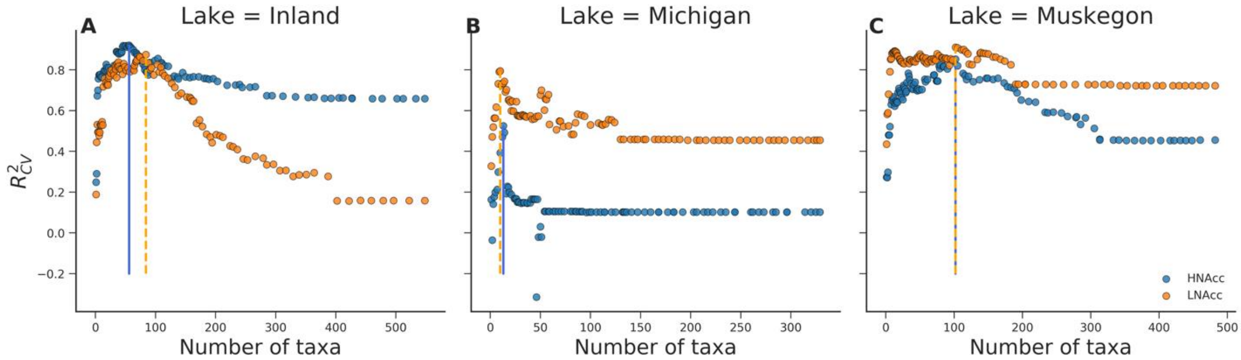


Figure 5.3. Evaluation of HNAcc and LNAcc predictions using the Lasso at all taxonomic levels.

Evaluation of HNAcc and LNAcc predictions using the Lasso at all taxonomic levels. for Muskegon lake, expressed in terms of R^2_{CV}, using different subsets of taxonomic variables. Subsets were determined by iteratively eliminating the lowest-ranked taxonomic variables based on the RL score.

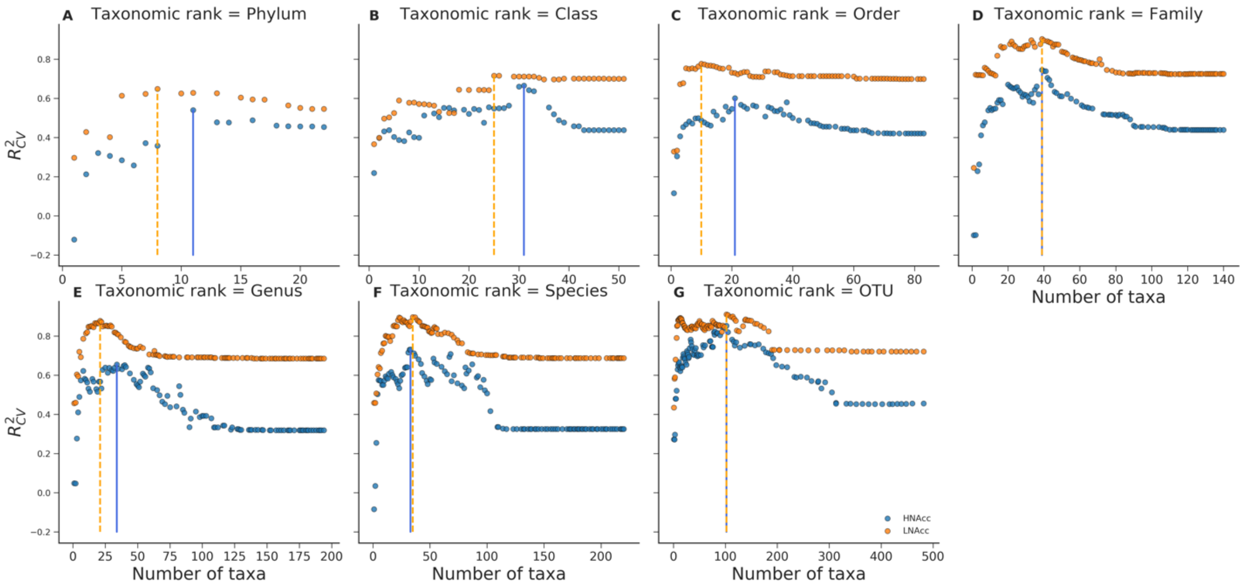


Figure 5.4. Hierarchical clustering of the RL scores.

Hierarchical clustering of the RL score for the top 10 selected OTUs within each lake system and FCM functional groups with the selected OTU (rows) across HNA and LNA groups within the three lake systems (columns).

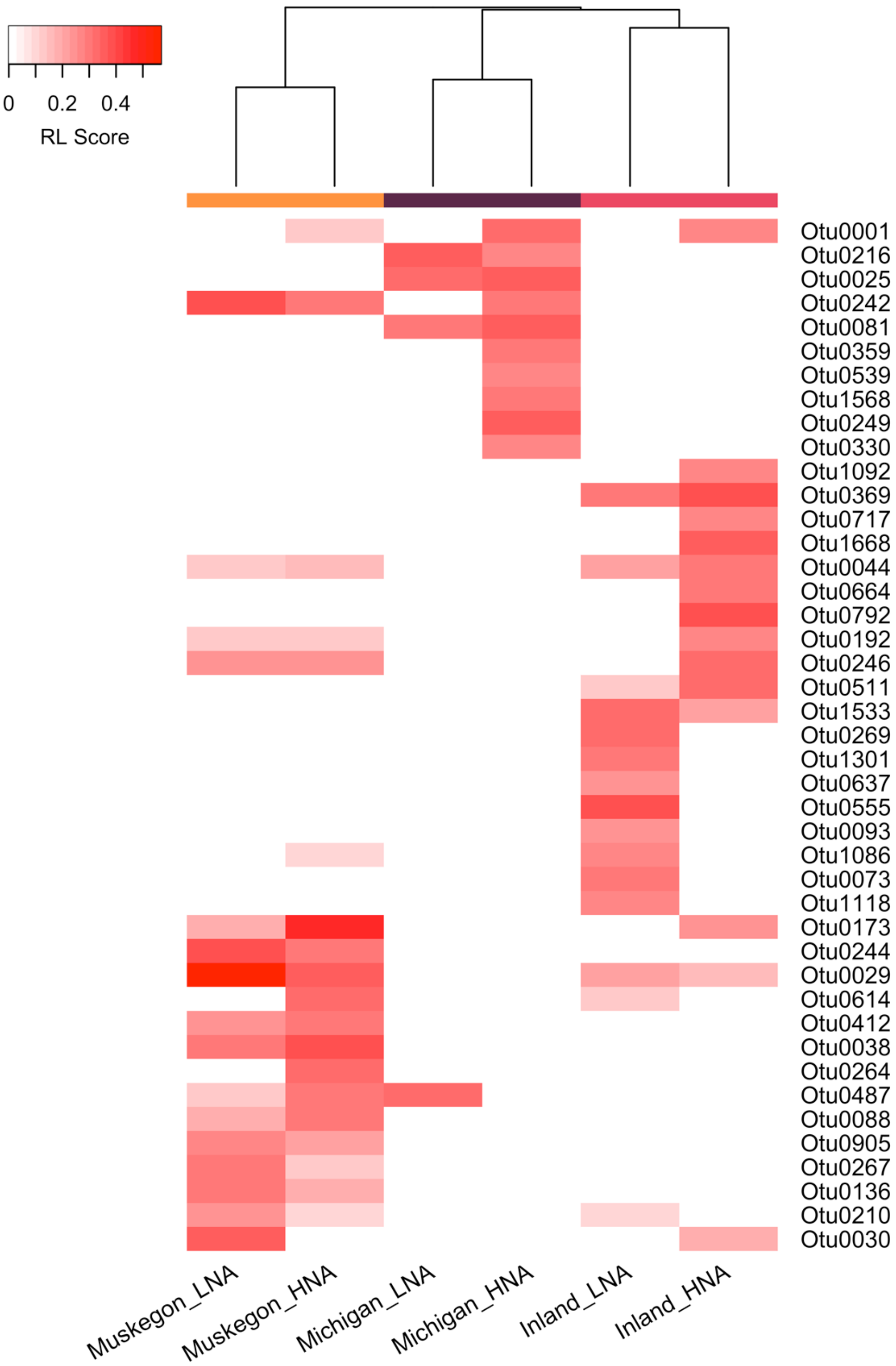


Figure 5.5. Phylogenetic tree with all HNA and LNA selected OTUs

Phylogenetic tree with all HNA and LNA selected OTUs from each of the three lake systems with their phylum level taxonomic classification and association with HNA, LNA or to both groups based on the RL score threshold values.

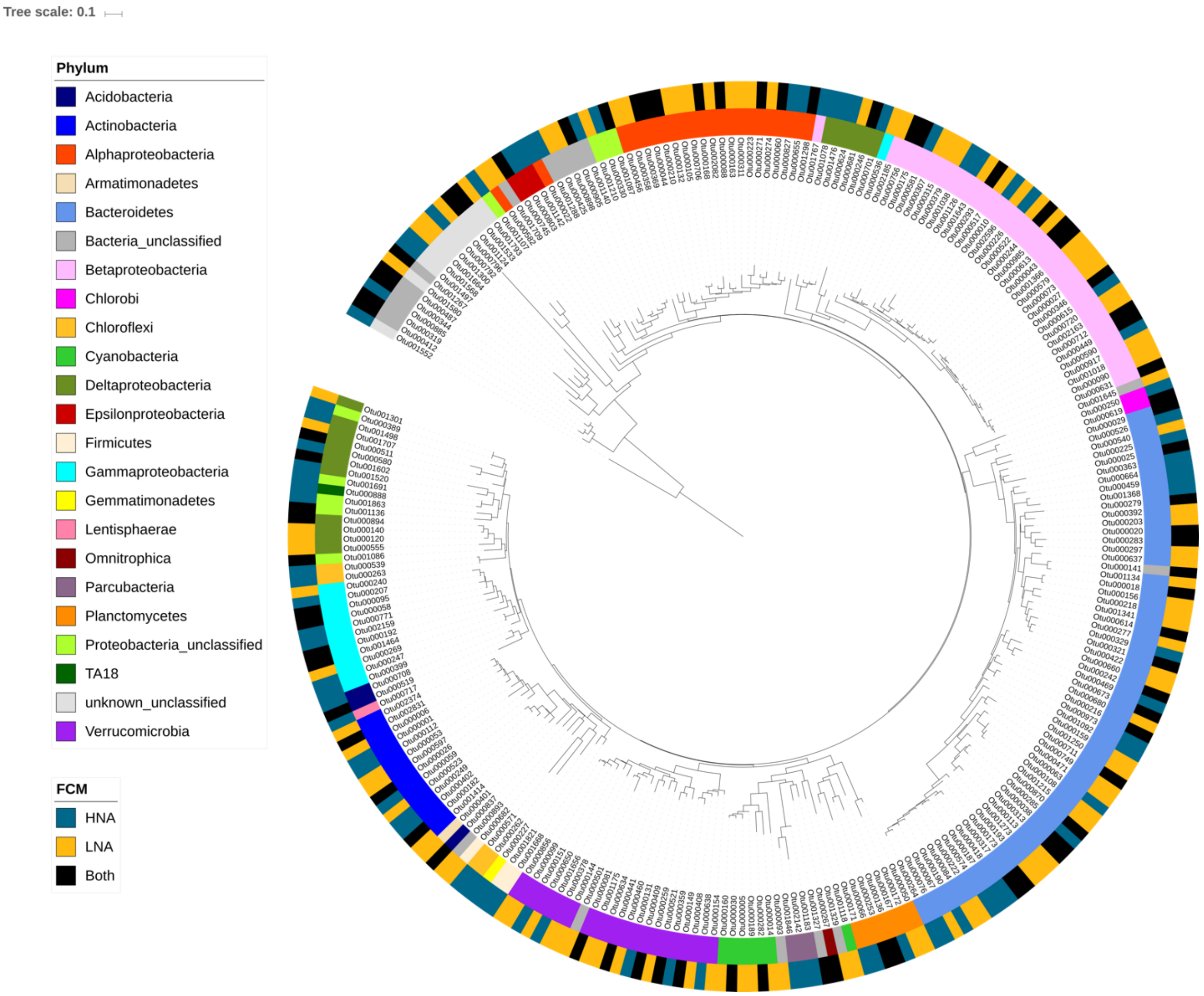


Figure 5.6. Correlation of the relative abundances of the top three OTUs selected by the RL and flow cytometry space.

Correlation (Kendall’s tau-b) between the relative abundances of the top three OTUs selected by the RL and the densities in the cytometric fingerprint. The fluorescence threshold used to define HNA and LNA populations is indicated by the dotted line.

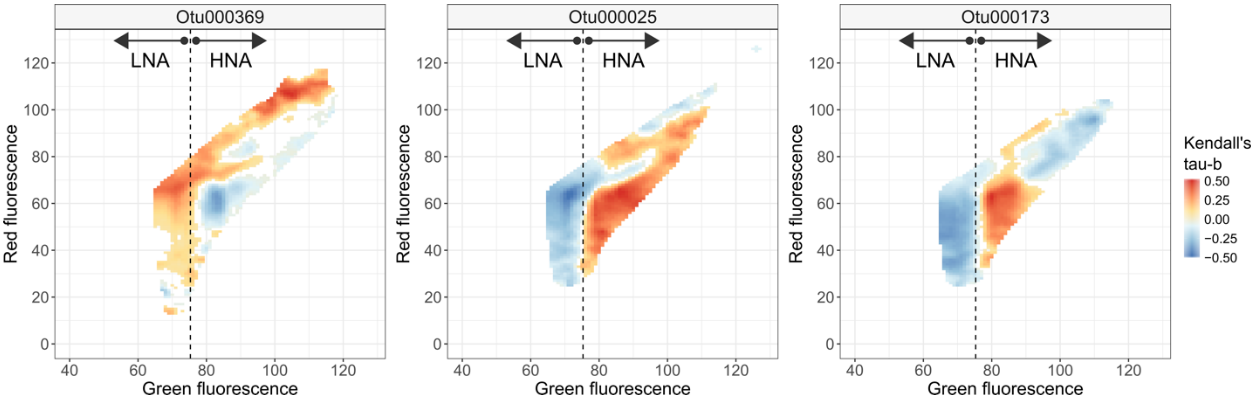


Table SI 5.2. Evaluation of Random Forest (RF) predictions.
Expressed in R^2_{CV} for all OTUs versus those selected (*i.e.*, sel) by the Boruta algorithm.

Lake system	Functional group	RF(all)	RF(sel)
Inland	HNAcc	0.53	0.71
Inland	LNAcc	0.21	0.48
Michigan	HNAcc	0.28	0.42
Michigan	LNAcc	0.40	0.59
Muskegon	HNAcc	0.45	0.59
Muskegon	LNAcc	0.66	0.77

Table SI 5.3. Full taxonomy of top-ranked OTUs according to the Randomized Lasso.

Lake system	Functional group	OTU	Phylum	Class	Order	Family	Genus	Species
Inland	HNAcc	OTU369	Proteobacteria	Alphaproteobacteria	Rhodospirrlalleles	alfVIII	alfVIII unclassified	Unclassified
Inland	LNAcc	OTU555	Proteobacteria	Deltaproteobacteria	Bdellovibrionales	Bdellovibrionacea	OM27_clade	Unclassified
Michigan	HNAcc	OTU025	Bacteroidetes	Cytophagia	Cytophagales	bacIII	bacIII-A	Unclassified
Michigan	LNAcc	OTU168	Proteobacteria	Alphaproteobacteria	Rhizobiales	alfVII	alfVII unclassified	Unclassified
Muskegon	HNAcc	OTU173	Bacteroidetes	Flavobacteriia	Flavobacteriales	bacII	bacII-A	Unclassified
Muskegon	LNAcc	OTU029	Bacteroidetes	Cytophagia	Cytophagales	bacIII	bacIII-B	Algor

Figure SI 5.7. Relationship between the RL score and relative abundance per OTU.

Scatter plot of RL score versus the average relative abundance of every OTU for HNAcc (blue points, **A**, **B**, and **C**) and LNA (orange points, **D**, **E**, and **F**) for each lake system: Inland Lakes (**A** and **D**), Lake Michigan (**B** and **E**), and Muskegon Lake (**C** and **F**).

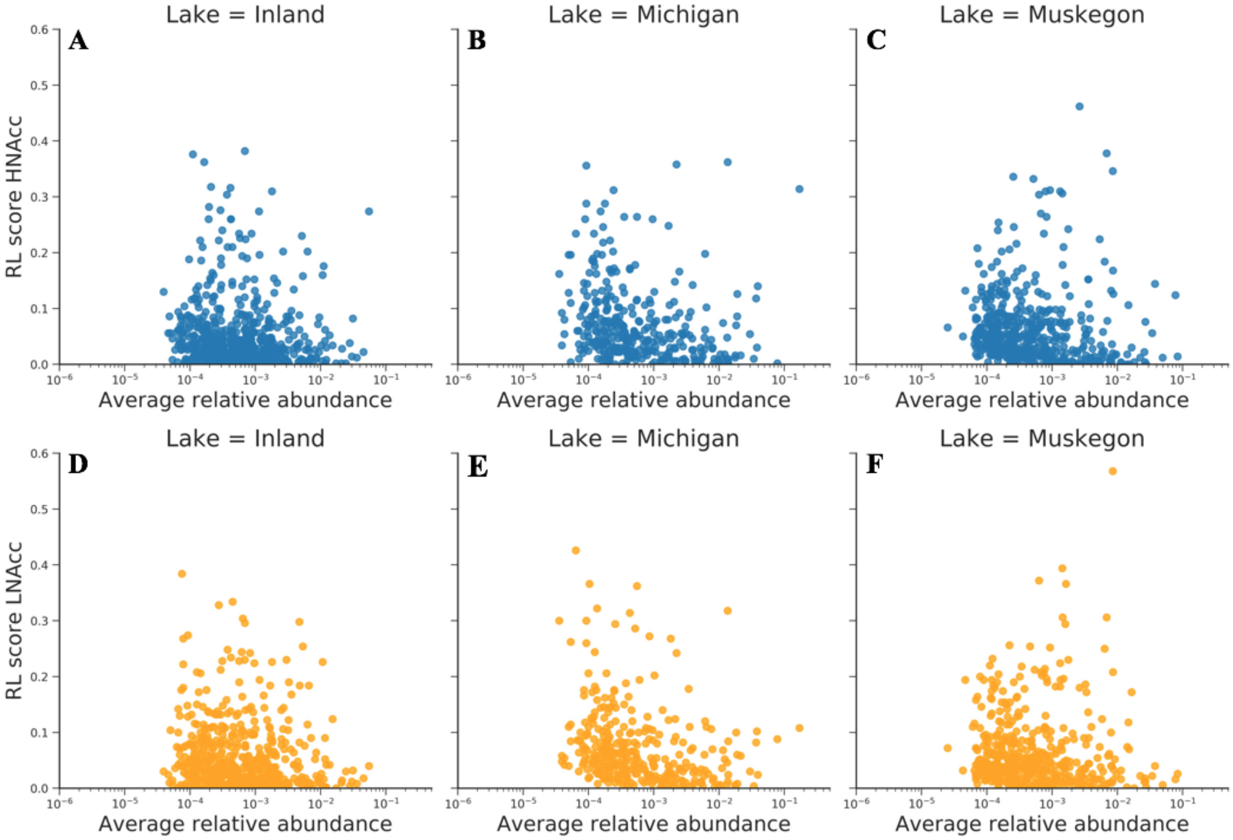


Figure SI 5.8. Comparison of predictions of HNAcc and LNAcc versus relative fractions.

Performed for Muskegon Lake at the OUT level expressed in terms of R^2_{CV}. The subset of taxonomic variables was iteratively reduced using a recursive variable elimination strategy, based on the RL score. Lowest-scored variables were removed at every step, after which the base model (*i.e.*, the Lasso) was used to model and predict cell counts or relative abundances. Predictions for HNA and LNA relative abundances overlap (red and green dots).

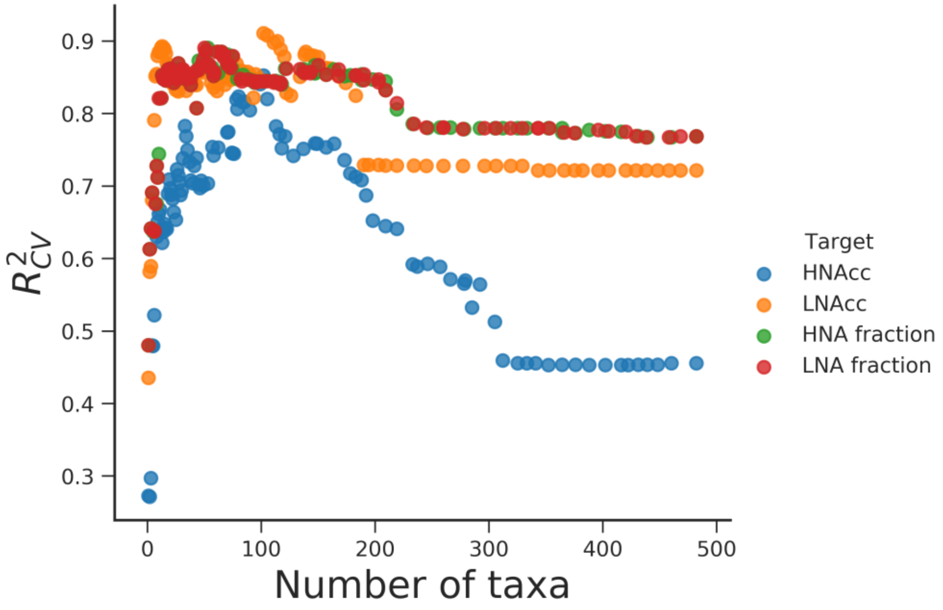


Figure SI 5.9. The R^2_{CV} of HNA and LNA OTUs selected by the RL.

Prediction of HNAcc (**A**) and LNAcc (**B**) for Muskegon Lake at the OTU level expressed in terms of R^2_{CV} using relative abundances (*i.e.*, “compositional”) and CLR transformed (*i.e.*, “CLR transformed”). The subset of taxonomic variables was iteratively reduced using a recursive variable elimination strategy, based on the RL score. Lowest-scored variables were removed at every step, after which the base model (*i.e.*, Lasso) were used to model and predict HNAcc and LNAcc.

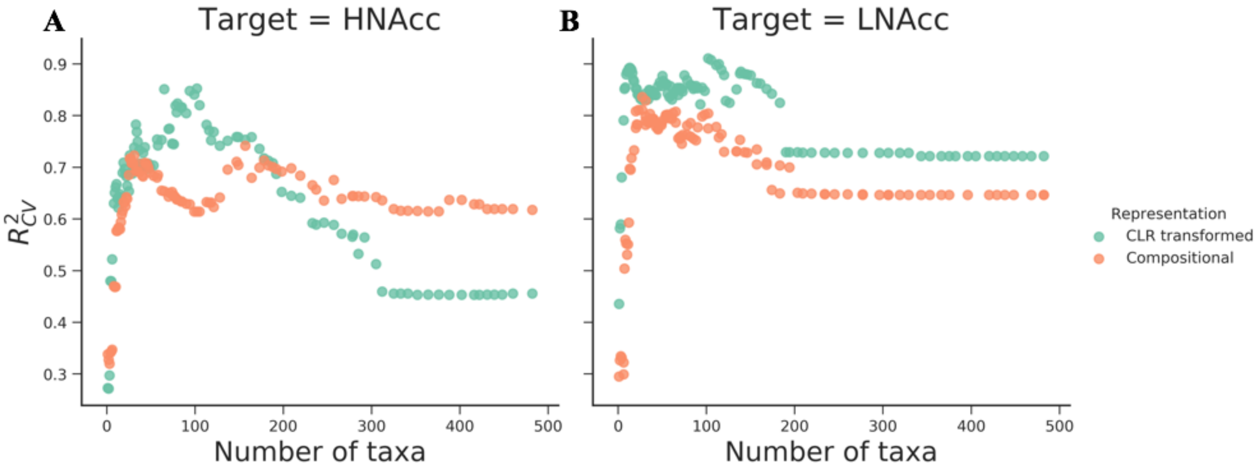


Figure SI 5.10. Distribution of the RL score for all three lake systems.

Inland Lakes (**A**), Lake Michigan (**B**), and Muskegon Lake (**C**) and all taxonomic levels in function of HNAcc.

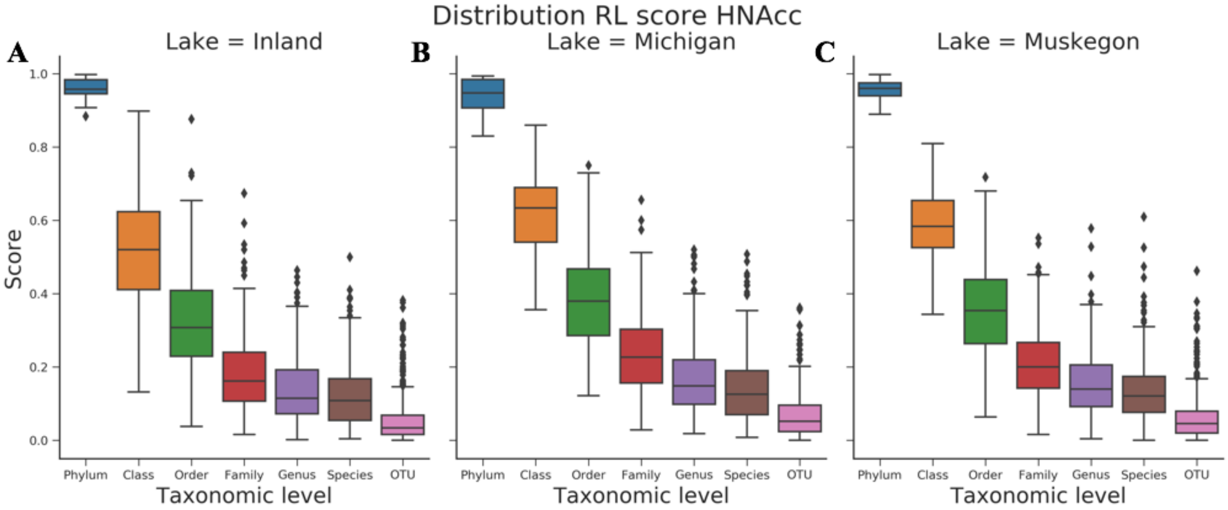


Figure SI 5.11. Selected OTUs in red according to the Boruta algorithm for each lake system and functional group.

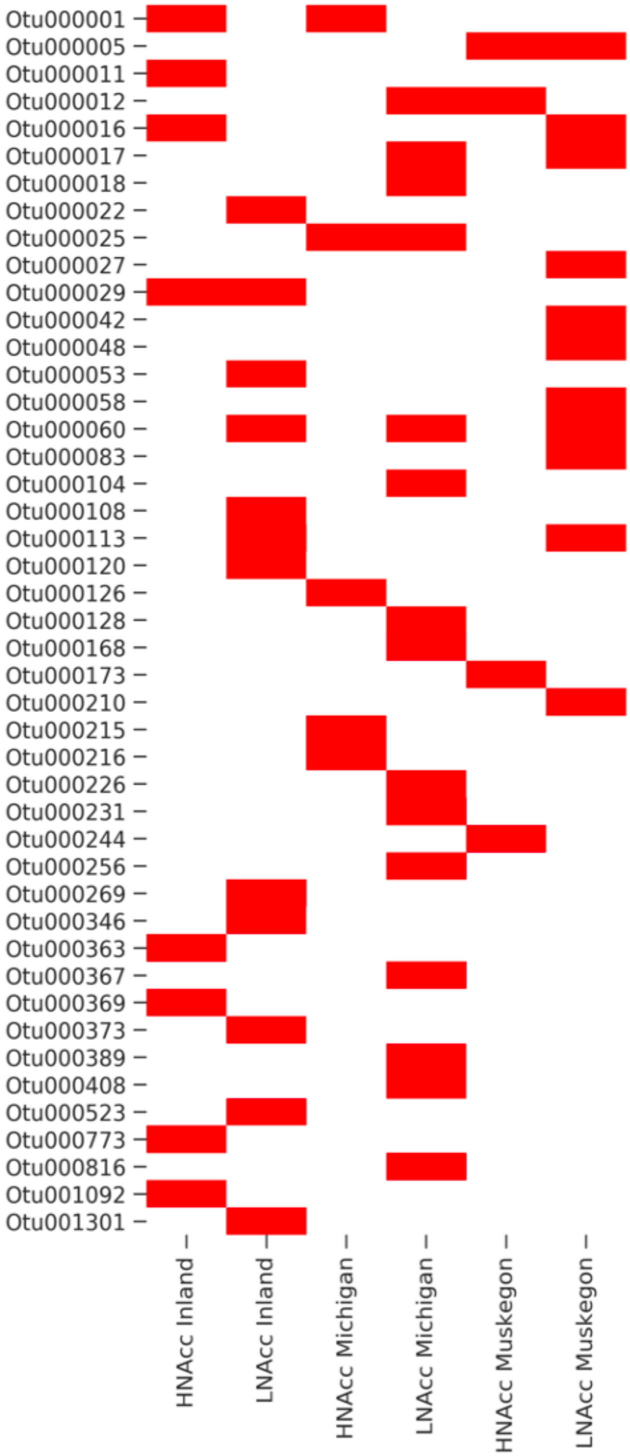


Figure SI 5.12. Fraction of selected OTUs using the Boruta algorithm for HNA and LNA.

Relative fraction of selected OTUs using the Boruta algorithm for HNAcc (blue points, **A, B,** and **C**) and LNAcc (orange points, **D, E,** and **F**) for each lake system: : Inland Lakes (**A** and **D**), Lake Michigan (**B** and **E**), and Muskegon Lake (**C** and **F**).

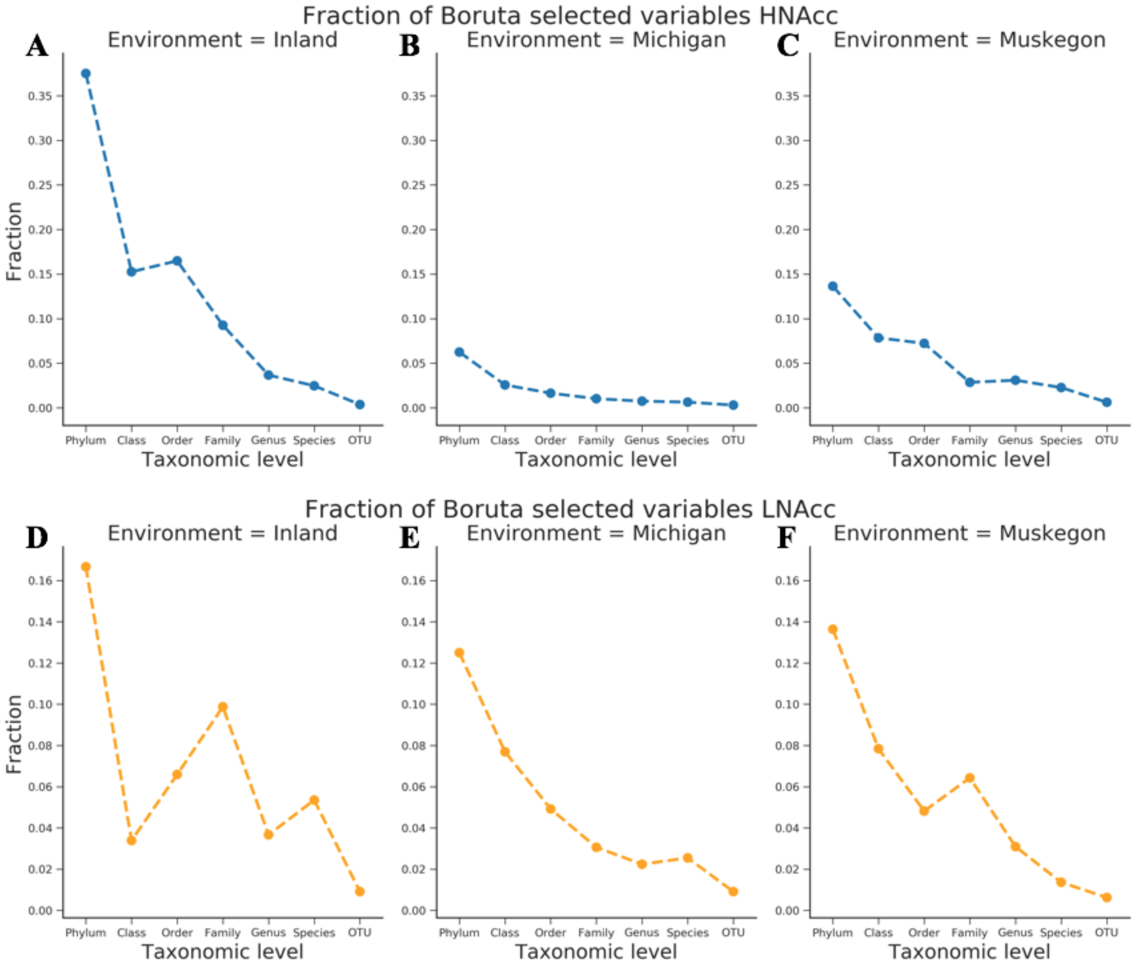


Figure SI 5.13. Comparison of Random Forest predictions using Boruta selected and the Randomized Lasso.

Comparison of Random Forest predictions (grey dashed line) using Boruta selected OTUs versus predictions using the Lasso and RL score for HNAcc (blue points, **A, B,** and **C**) and LNAcc (orange points, **D, E,** and **F**) for each lake system: Inland Lakes (**A** and **D**), Lake Michigan (**B** and **E**), and Muskegon Lake (**C** and **F**).

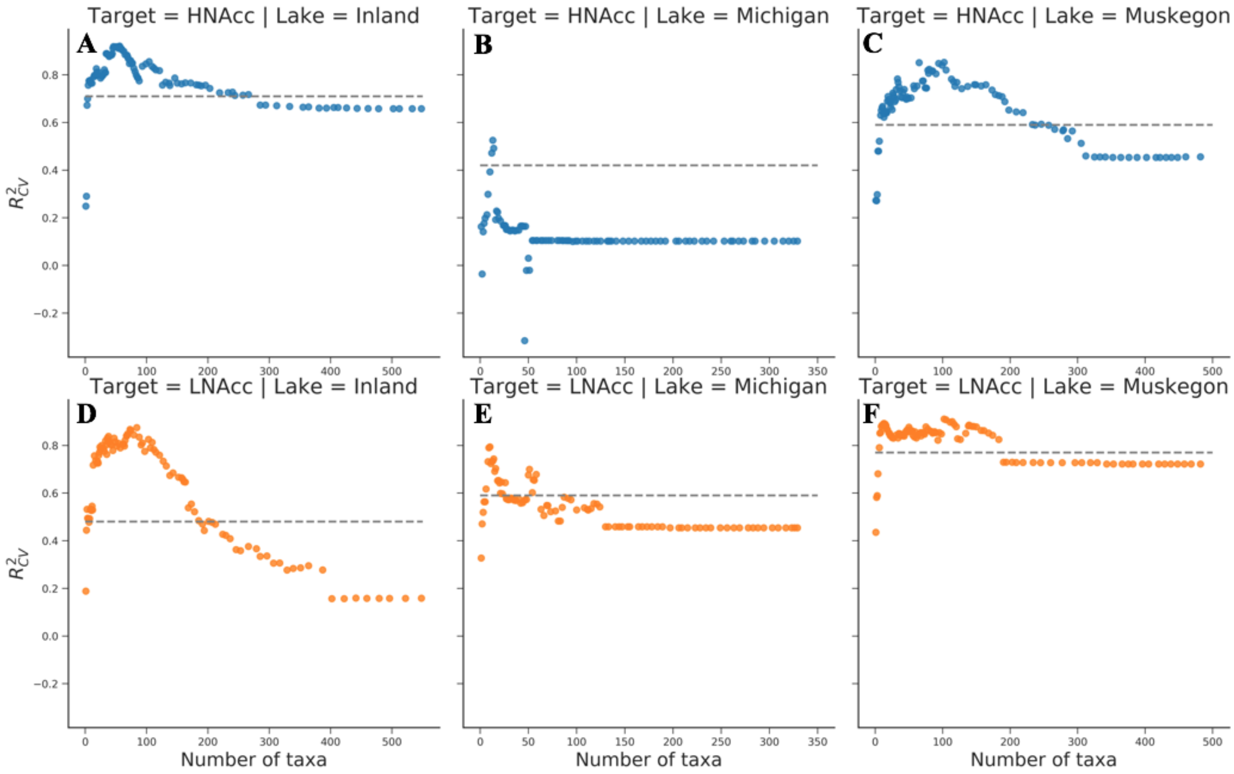


Figure SI 5.14. Pearson correlations between HNAcc and LNAcc RL scores.

Pearson correlations between RL scores assigned to OTUs in function of HNAcc and LNAcc between lake systems. Only those OTUs were considered that were present in all lake systems, which were 190 in total. Values are bolded if P < 0.05.

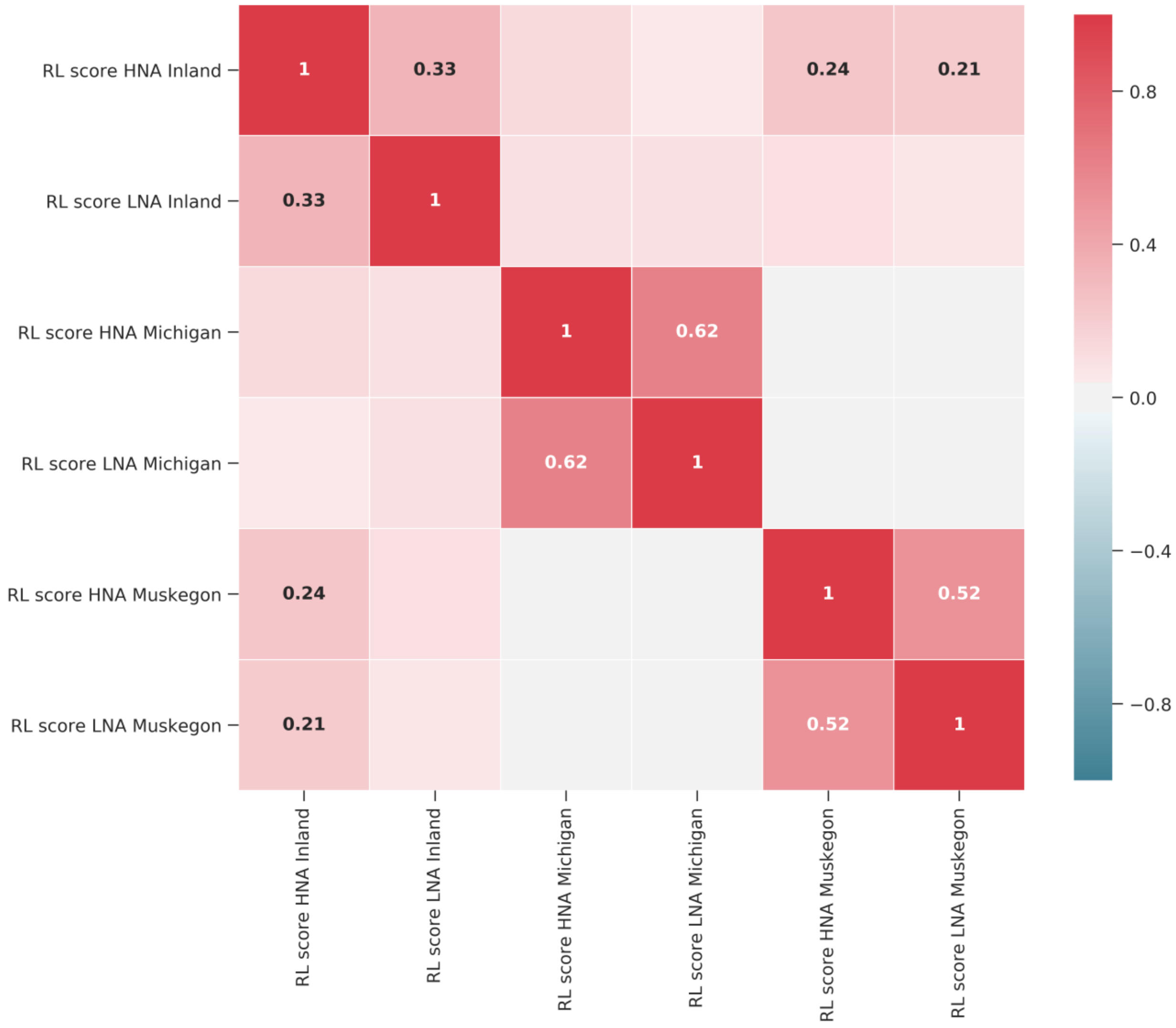


Figure SI 5.15. Phylogenetic tree with RL-selected OTUs and their HNA and LNA RL scores.

Phylogenetic tree with all HNA and LNA selected OTUs from each of the three lake systems with their phylum level taxonomic classification (first row) and association with HNA, LNA, or both groups based on the RL score threshold values (las row. Second (HNA) and third rows (LNA) display the RL scores.

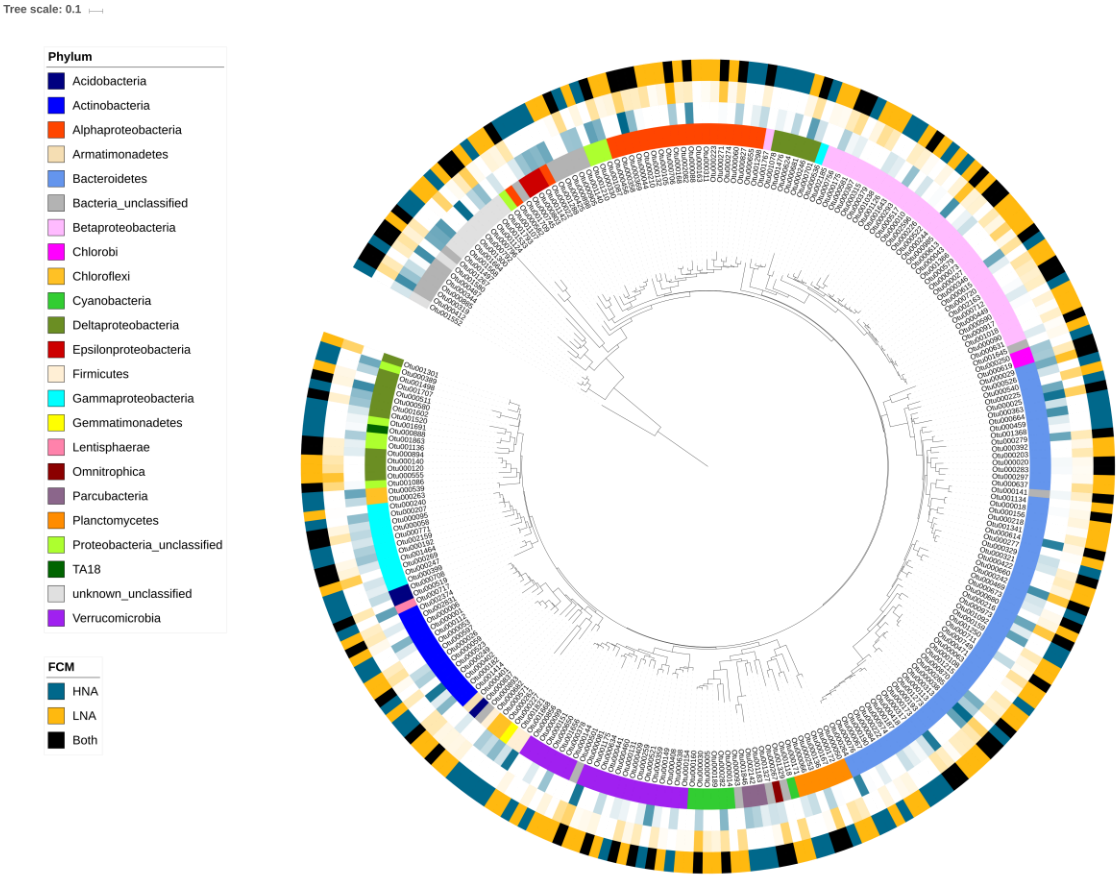


Figure SI 5.16. Kendall's tau of individual OTU abundances and productivity measurements versus the RL score determined in function of HNAcc.

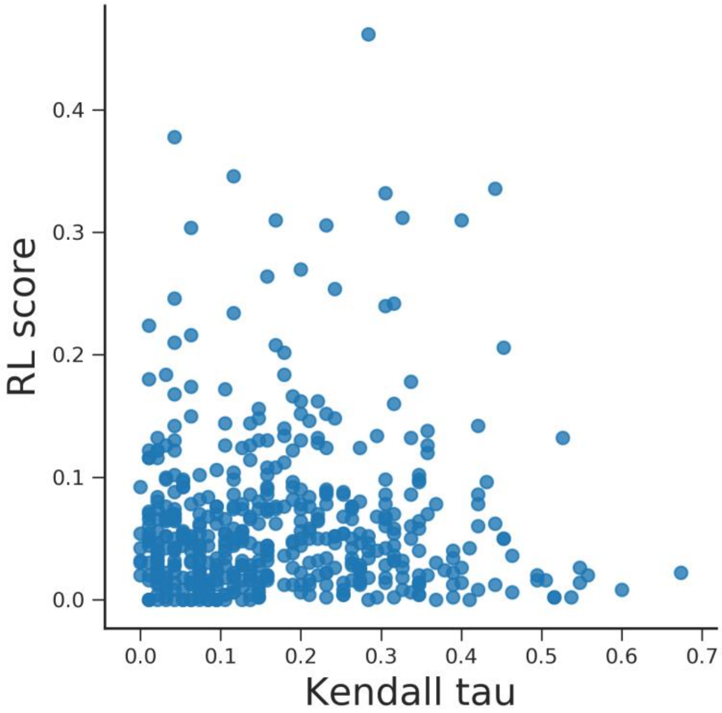
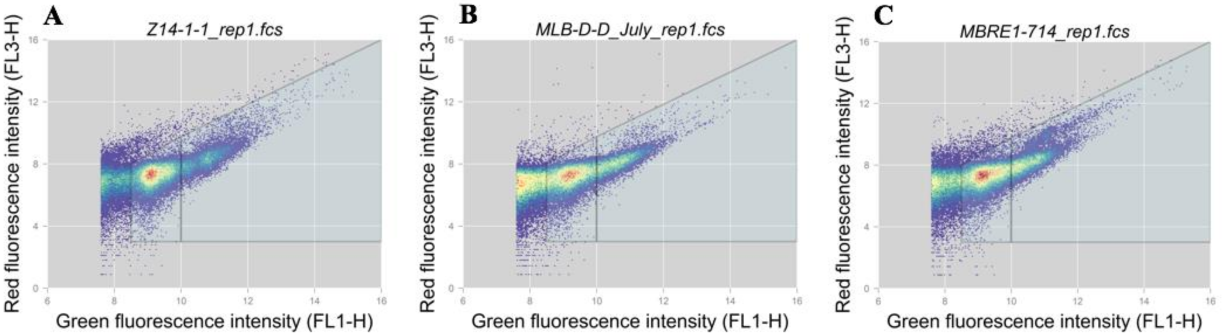


Figure SI 5.17. Visualization of the flow cytometry gating strategy.

Examples of the gating strategy to determine HNAcc and LNAcc for the three lake systems. The gating strategy is performed in the arcsinh(x) transformed bivariate space of the FL1-H and FL3-H channel, following guidelines of Preset et al., 2013.



Chapter VI:

Conclusions

Reflection

In my dissertation, I used an observational approach to investigate the patterns of bacterial community diversity, composition, and genomic structure as they relate heterotrophic production which is an important ecosystem function in freshwater lakes. The goal of my dissertation was to connect major ideas in the broader field of community ecology to a microbiological context using freshwater lakes as the testing grounds.

In **Chapter II**, I sampled the particle-associated and free-living bacterial community composition of surface and bottom lake layers within 11 lakes that had high- and low-productivity. I found that independent of lake nutrient level and surface or bottom lake layer, particle-associated and free-living bacterial community composition differed and there were larger inter-lake differences in the particle-associated communities than for free-living communities. These community differences were consistent in all comparisons at the phylum-level, emphasizing that large changes in community composition are likely the result of deeply conserved traits within specific phylogenetic groups.

In **Chapter III**, I collected samples taken from a freshwater, estuarine lake with a large, natural productivity gradient to assess whether microhabitat played a role in shaping bacterial biodiversity-ecosystem function relationships. I found that particle-associated communities were more diverse (*i.e.,* higher species richness) and displayed a positive, linear relationship between diversity and heterotrophic production, whereas no relationship was found in the free-living communities. The relationship for particle-associated communities strengthened when evenness was accounted for, indicating that bacterial abundance matters for productivity. Additionally, I found that there was a negative relationship between phylogenetic diversity and per-capita heterotrophic production, suggesting that communities with more closely related species have higher per-capita production. This highlights that specific ecological processes may occur on particulate matter that do not occur in the open water, which I suggest arise due to closely related taxa in particles complementing each other's metabolic functions.

In **Chapter IV**, I used the samples from Chapter III to further investigate the potential genomic characteristics that underpin differences in the diversity, community composition, and ecosystem functioning of particle-associated and free-living habitats. I found that bacteria which specializing in particle-associated microhabitats have larger genome sizes, a lower proportion of coding DNA, higher GC content, and higher nitrogen content. I also found that the number of unique genes was higher in particle-associated genomes. However, per region of the genome, the particle-associated bacteria had both fewer genes and lower gene diversity compared to free-living bacteria who "packed more in" to their genomes. This major difference in genome structure based on microhabitat suggests that an organism's ecology is strongly linked to evolutionary processes that occur within their genomes.

Finally in **Chapter V**, with samples taken from three types of freshwater lake systems, I associated bacterial composition using machine learning with two important flow cytometry groups that are ubiquitous across all aquatic systems: high- (HNA) and low-nucleic acid (LNA) functional groups. Even though HNA bacteria were only half as abundant as LNA bacteria in my samples, they were very strongly correlated with heterotrophic production. This implies that the smaller pool of HNA bacteria may play a disproportionately large role in the freshwater carbon flux. In addition, I found that very few taxa specialized in either the HNA or LNA groups but were instead highly system specific. This suggests that it is unlikely that there are universal bacterial taxa within the HNA or LNA groups across aquatic systems. Rather, HNA functional groups likely represent the subset of the community that are actively contributing to ecosystem functioning at the time of sample collection.

Some unknowns and potential limitations

Here, I have defined the spatial niche of bacteria using sequential in-line filtering with filters of two different pore sizes (3 um and 0.22 um) to define the particle-associated and free-living microhabitats in my studies. This is a very common method used in the aquatic sciences to study bacteria (Bidle and Fletcher 1995, Crump et al. 1998, Acinas et al. 1999, Besemer et al. 2005, Eloe et al. 2011, Jackson et al. 2014, Sunagawa et al. 2015). This approach assumes that larger pore size filters collect detrital particles, microeukaryotes, and (potentially) chemotactic organisms whereas the plankton are captured on the smaller filter. However, there are fundamental drawbacks to this approach. Because the filters are usually used to destructively collect DNA or RNA, information about the material collected on the filter is typically ignored and not analyzed for particle source, type, or density of particles per liter. A lack of knowledge of the source and density of particles has important ecological ramifications. For example, particle-associated bacteria on a detrital particle likely have different ecological and evolutionary strategies compared to bacteria that are facultative or obligate symbionts with algal cells. Finally, information on particle source and load will help provide more accurate scaling up of bacterial processes for ecosystem models.

Future work on social and community interactions within bacterial communities attached to particulate matter will also be important for understanding ecological relationships. For example, in an experimental study on model marine particles, Datta et al. (2016) found that there was rapid community succession from primary particle degraders to secondary consumers. The secondary consumers may decrease the particle degradation rate while increasing the bacterial biomass attached to particulate matter. This detail influences the ecological interpretations (and therefore assumptions) that researchers make about an organism's niche and metabolic function. In addition, future work on bacterial social interactions may have relevant implications for bacterial community dynamics. Ganesh et al. (2014) found that genes mediating social interactions (*i.e.,* cell-to-cell transfer, antibiotic resistance, genetic mobile elements, viruses, adhesion, and motility) were upregulated on particulate matter in an oxygen minimum zone. These types of social and community interactions will also have large impacts on scaling up microbial processes.

Synthesis

My dissertation works to step beyond the traditional descriptive approach in microbial ecology, which asks “who’s there?” and “what are they doing?” (**Figure 1.1A**). Rather, I work to integrate theories and ideas from the broader field of community ecology into bacterial systems (**Figure 1.1B**). The work presented here shows that bacterial systems fit into a community ecology framework. For example, niche theory and biodiversity-ecosystem function relationships, which were originally developed in plant and animal systems, are indeed relevant to the bacterial realm. Asking community ecology questions about bacterial systems does have its advantages. Bacterial communities are a useful system for developing ideas regarding the interplay between ecology and evolution. This is not only because bacterial ecology and evolution can work on similar time scales but also due to the way bacterial communities are sampled (*i.e.*, through molecular approaches), which allows for evolutionary interpretations of ecological patterns. As an example, my future research will focus on the ecological role of horizontal gene transfer and determine the impact that nutrient concentration and particle-association have on the dynamics of gene transfer.

References

Acinas, S. G., J. Antón, and F. Rodríguez-Valera. 1999. Diversity of Free-Living and Attached Bacteria in Offshore Western Mediterranean Waters as Depicted by Analysis of Genes Encoding 16S rRNA. Applied and Environmental Microbiology 65:514–522.

Besemer, K., M. Moeseneder, J. Arrieta, G. Herndl, and P. Peduzzi. 2005. Complexity of bacterial communities in a river-floodplain system (Danube, Austria). Applied and environmental microbiology 71:609–620.

Bidle, K. D., and M. Fletcher. 1995. Comparison of free-living and particle-associated bacterial communities in the chesapeake bay by stable low-molecular-weight RNA analysis. Applied and Environmental Microbiology 61:944–952.

Crump, B. C., J. A. Baross, and C. A. Simenstad. 1998. Dominance of particle-attached bacteria in the Columbia River estuary, USA. Aquatic Microbial Ecology 14:7–18.

Datta, M. S., E. Sliwerska, J. Gore, M. F. Polz, and O. X. Cordero. 2016. Microbial interactions lead to rapid micro-scale successions on model marine particles. Nature Communications 7:11965.

Eloe, E. A., C. N. Shulse, D. W. Fadrosh, S. J. Williamson, E. E. Allen, and D. H. Bartlett. 2011. Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. Environmental Microbiology Reports 3:449–458.

Ganesh, S., D. J. Parris, E. F. DeLong, and F. J. Stewart. 2014. Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. The ISME journal 8:187–211.

Jackson, C. R., J. J. Millar, J. T. Payne, and C. A. Ochs. 2014. Free-Living and Particle-Associated Bacterioplankton in Large Rivers of the Mississippi River Basin Demonstrate Biogeographic Patterns. Applied and Environmental Microbiology 80:7186–7195.

Sunagawa, S., L. P. Coelho, S. Chaffron, J. R. Kultima, K. Labadie, G. Salazar, B. Djahanschiri, G. Zeller, D. R. Mende, A. Alberti, F. M. Cornejo-Castillo, P. I. Costea, C. Cruaud, F. D'Ovidio, S. Engelen, I. Ferrera, J. M. Gasol, L. Guidi, F. Hildebrand, F. Kokoszka, C. Lepoivre, G. Lima-Mendez, J. Poulain, B. T. Poulos, M. Royo-Llonch, H. Sarmento, S. Vieira-Silva, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, C. Bowler, C. de Vargas, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, O. Jaillon, F. Not, H. Ogata, S. Pesant, S. Speich, L. Stemmann, M. B. Sullivan, J. Weissenbach, P. Wincker, E. Karsenti, J. Raes, S. G. Acinas, P. Bork, E. Boss, C. Bowler, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. Sieracki, and D. Velayoudon. 2015. Structure and function of the global ocean microbiome. Science 348:1261359.