THE UNIVERSITY OF CHICAGO

THE DYNAMICS OF BACTERIAL COMMUNITIES ASSOCIATED WITH *ARABIDOPSIS THALIANA*

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON MICROBIOLOGY

BY
MATTHEW PERISIN

CHICAGO, ILLINOIS
MARCH 2016

"The brick walls are there for a reason. The brick walls are not there to keep us out; the brick walls are there to give us a chance to show how badly we want something. The brick walls are there to stop the people who don't want it badly enough. They are there to stop the *other* people.

- Randy Pausch, *The Last Lecture*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

During my the course of my dissertation research I have benefited greatly from personal and professional help from advisors, administrators, colleagues, friends, and family. I am indebted to my advisor, Joy Bergelson, for taking a chance on a biochemist wanting to convert to an ecologist. Joy gave me the freedom to define and conduct research projects, that at times was overwhelming, but looking back made me a strong and independent scientist. From a writing standpoint, I am extremely grateful to have learned from the best. I have strived to emulate Joy's concise and impactful style with this dissertation.

My committee has also provided great mentorship and advice over the years. Stefano Allesina has been a fantastic collaborator and gave me a base in theoretical ecology and scientific computing. Howard Shuman has inspired and challenged me to think outside the box. Sean Crosson has demonstrated the importance of excitement and a sense of humor while doing awesome and impactful science. Jack Gilbert has provided encouragement, research advice, and career guidance all at the drop of a hat. I am continually amazed at his ability to do all the things at breakneck speed and with such passion and excitement.

Being a committee of microbiology student working in an ecology & evolution lab while on a genetics and regulation training grant, I have benefited from countless administrative staff. Thanks to Jeff Wisniewski, Bonnie Brown, Connie Homan, and Noami Perez in the E&E office for all the work you do to allow students to focus on research without having any worries about administrative issues. Special thanks to Alison Anastasio for always lending an ear to personal and administrative concerns and for just being awesome. Thanks to Ian Miller and Garnett Kirk for always taking time to help with computer issues. There has always been flux in the microbiology department, but thanks to Noreen Bentley for help organizing my defense. From the genetics training grant, I am lucky to have gotten to know Sue Levison, who has always been a huge help.

The Bergelson Lab has been a great place to work thanks to all past and present members. During my rotation, Luke Barrett initially got me excited in plant-microbe interactions.

Wayan Muliyati was a great lab manager and bus buddy. Thanks to all lab members for brainstorming and moral support, including Madlen Vetter, Leah Johnson, Matt Horton, Chris Meyer, Sumitha Nallu, Hana Lee, and Kat Beilsmith. I am thankful to have started in the lab as three other students that have provided constant support and feedback. Thanks Talia Karasov, Alice MacQueen, and Laura Merwin. I am also thankful for technician and transfer student assistance from Carlos Sahagun, Roderick Woolley, and Mathieu Videlier. I have collaborated with most of the students and postdocs that have come through the lab, but I am especially thankful for the collaboration with Ben Brachi. I have expanded my technical and conceptual skills thanks to him. A special thanks also to Tim Morton who came in as the new lab manager and brought excitement and enthusiasm everyday no matter the circumstances. He truly is fantastic.

This grad school journey would have been much more difficult without great classmates and friends. Thanks to my microbiology cohort of Bryan, Mike, Sao-Mai, and Tristan for support and general comradery. Thanks to everybody on the softball teams and the μBio Club founders, Mike Gebhardt, Sean Gibbons, Vivian Choi, and Charlie Wright. Thanks to my friends back home for support and not constantly asking me when I would graduate.

I am extremely grateful for my family. Steve and Tom have supported me in so many ways. There are no words to describe how thankful I am for their generosity. My cousins Andy, Derek, and Jeremy have been there for me no matter what. Aunt Susie and Uncle Marc have provided unconditional love and support. Thanks to my grandparents for giving me the smarts and determination to succeed. Though I have lost a lot during grad school, I am thankful to have gained a whole new family. Thanks to my in-laws Becky and Tim for moral support, and to the whole family for good times before and good times to come. My sister will probably take some credit for this dissertation, and some is deserved. She set the bar high and I have always tried to reach or surpass it. My dad has supported me all along and given me the freedom to forge my own path. To my mom and Aunt Judy, I miss you both so much. I know you both will be looking down with big smiles at my defense. Finally,

I am thankful for my beautiful, talented, and supportive wife Casey. We have both been on this rollercoaster ride and it is safe to say that the highs would not have been as high and the lows would have been lower without you.

ABSTRACT

Plants are colonized inside and out by a diverse array of microbes. These inhabitants can have pathogenic, beneficial, or commensal relationships with the plant. Historically, studies have focused on individual microbes and their effects on plant phenotypes and fitness; however, recent evidence points towards a microbial community context for these interactions. To fully describe host-pathogen interactions, we need to better understand not just the direct interactions between host and microbe, but also between microbe and microbe in the host-associated microbiota. Greater understanding of the ecological forces that structure these communities will lead to effective strategies for promotion of plant health.

To uncover these forces, I have studied the bacterial communities associated with *Arabidopsis thaliana* and taken a multi-faceted approach involving observation of controlled and natural communities. In Chapter 2, I describe the benefits and pitfall of three different approaches for quantifying mixed species communities *in vitro* and *in planta*. In Chapter 3, I characterize the temporal dynamics and predictability of communities across two axes of complexity, environmental heterogeneity and species richness. In Chapter 4, I examine the spatial and temporal dynamics of communities associated with natural *A. thaliana*.

Taken together these chapters demonstrate remarkable disparities in the ability to predict bacterial community succession under controlled versus field conditions. For mixtures of endophytic leaf isolates *in vitro*, I observed large dissimilarities in community compositions between replicates, even given a homogenous liquid media environment. In contrast, for field grown plants, I observed consistent differences in community composition by tissue that followed similar successional trajectories within tissue over two years at two sites. This predictability discrepancy is likely due to differences in the types, timing, and number of selective forces in controlled versus field environments. Under controlled conditions, bacteria encountered rich liquid or solid media or gnotobiotic plants rich in unoccupied niches. Bacterial competition for nutrients was likely the main force structuring the community, with variable outcomes for each replicated community. In contrast, bacterial communities asso-

ciated with field grown plants followed remarkably repeatable successional patterns. Communities associated with vegetative roots initially diverged from soil communities in terms of composition, and upon flowering and senescence, root communities became more soil-like. Comparison of operational taxonomic unit (OTU) profiles between tissues indicated that the roots first filter OTUs from the surrounding soil that become differentially enriched in each aboveground tissue.

The natural environment imposes a variety of selective forces on bacterial community composition, including fluctuations in abiotic conditions, host metabolite and defense factors, and interspecific competition. Based on my *in vitro* findings competitive outcomes can be difficult to predict *a priori*; however, host and environmental forces seem to impose stronger and more consistent selection on community composition. These findings have implications for the use of pro and prebiotics for microbiome manipulations. Given that bacterial competition is difficult to predict even under the simplest of conditions, introducing probiotic strains to compete with native strains is likely to be ineffective. Instead, selecting for the growth of native, desired taxa by specific nutrient addition may be more successful. Prebiotic strategies are more in line with natural processes in which the plant produces nutrient and defense metabolites that select for specific taxa. Manipulating microbiomes in the soil prior to planting is crucial. The roots will filter strains from the initial soil community that will then colonize the entire plant.

CHAPTER 1

INTRODUCTION

Bacterial diversity is astonishing. The number of bacteria inhabiting the surface and interior of leaves (10^{26}) outnumbers the estimated number of stars in the known universe by four orders of magnitude (Lindow & Brandl, 2003), and less than one gram of soil can harbor thousands of different bacterial species (Schloss & Handelsman, 2006). The ubiquitous distribution of bacteria from the deep oceans to human intestines is of upmost importance for sustaining life on Earth. Bacteria maintain fundamental geochemical processes (Falkowski *et al.*, 2008), and human-associated bacteria train our immune system and break down complex carbohydrates into usable forms (Round & Mazmanian, 2009; Grice & Segre, 2012). Plant-associated bacteria fix nitrogen into usable forms and solubilize nutrients such as phosphate (Cocking, 2003; Rodríguez *et al.*, 2006). Bacteria are so tightly intertwined in all forms of life, that Gilbert & Neufeld (2014) prognosticated that the absence of all bacteria would result in collapse of human civilization within a year.

In general, microbes do not exert these effects in monoculture, but instead as diverse multi-species communities, commonly referred to as microbiomes or microbiotas. To understand the maintenance of this diversity, we first need to catalogue communities occupying diverse niches across the globe, which has been the objective of the Earth Microbiome Project (Gilbert *et al.*, 2014). For host-associated microbial communities, there have been large scale studies to understand diversity relationships with disease and what constitutes a normal microbiome (Huttenhower *et al.*, 2012). Human and mice microbiome studies have linked microbial consortia to inflammatory bowel disease, colon cancer, obesity, and even autism (Kinross *et al.*, 2011; Hsiao *et al.*, 2013). These studies have added an additional layer of complexity to Koch's Postulates. Instead of linking one pathogen to one disease, researchers are linking communities of microbes to disease. There is now a community context to each disease and insights into complex relationships not only between pathogens and hosts, but also between pathogens and commensals or between commensals and hosts is required to

fully understand disease progression. The power of community-level approaches is evidenced by the efficacy of fecal microbiota transplants for treatment of recurrent *Clostridium difficile* infections (van Nood *et al.*, 2013).

Though the majority of host-associated microbiome research has focused on humans, the same techniques have been applied to plant-associated microbial communities. The plant microbiome field is accelerating at a rapid pace, marked by increased interests and collaborations in industry (Fox, 2014). Plant-microbe research has traditionally focused on single pathogens that cause specific diseases (Mansfield *et al.*, 2012); however, there is now an appreciation that microbes affect plant phenotypes and fitness not in isolation but in complex communities, in which interactions may be diffuse (Karasov *et al.*, 2014). Additionally, bacterial consortia in the soil have been observed to affect flowering time (Wagner *et al.*, 2014; Panke-Buisse *et al.*, 2015), demonstrating a community level modulation of host phenotype. Many parallels have been drawn between human and plant-associated microbiome research (Hacquard *et al.*, 2015), with the hope of uncovering the ecological mechanisms by which pathogens infiltrate a community and interact with the host to cause disease. If we can better understand the ecological forces that structure microbial communities, then we can design strategies to promote host health and resist pathogen invasion.

1.0.1 Community ecology theory and microbes

To gain insights into these ecological forces that maintain remarkably diverse microbial communities, we must first remember and apply classic community ecology theory. Before ever considering the remarkable estimates of bacterial diversity, ecologists were baffled at the extent of diversity in nature, especially given the competitive exclusion principle (Hutchinson, 1961). The competitive exclusion principle states that two species residing in the same niche will compete to the extinction of the inferior competitor (Gause, 1934). The niche differentiation theory first emerged to reconcile observed diversity with competitive exclusion (Hutchinson, 1961). At the core of this theory is the idea that no two species occupy the

same niche due to ecological differences in resource use, predators and enemies, and abilities to survive over temporal and spatial environmental gradients (Hutchinson, 1961). Within a community, Lotka-Volterra equations have provided a basis for understanding coexistence (Lotka, 1920; Volterra, 1926). A species cannot competitively exclude another species if its own growth negatively impacts itself more than the competitor; therefore, the mechanisms that govern coexistence are stabilizing in that they increase negative intraspecific interactions relative to negative interspecific interactions (Lotka, 1920; Volterra, 1926; Chesson, 2000). Species niche differences stabilize the community by overcoming differences in competitive ability (Hutchinson, 1961; Levine & HilleRisLambers, 2009). For example, consider resource partitioning. If a species depends on a particular resource, than its consumption of that resource negatively impacts its own resources relative to competitors that utilize another resource. Therefore, niche differentiation theory predicts that a species growth rate will be greater when it is at a low relative to high density since its own resource will become limiting as its density increases (Chesson, 2000).

The role of niches in species coexistence has been challenged by Hubbell's neutral theory, which assumes that all species are ecologically equivalent in regards to birth, death, dispersal, and speciation rates (Hubbell, 2001). In this theory, community structure is determined by stochastic processes and dispersal limitation (Hubbell, 2001). This theory greatly simplifies community structure, yet has been used effectively to explain diversity in a variety of habitats (Chave & Leigh, 2002; Woodcock *et al.*, 2007). Alonso *et al.* (2006) emphasize the parsimony and liken neutral theory to the ideal gas law in that ideal gases like ecologically equivalent species may not exist, but the utility of the theory cannot be denied. In this theory there is a source metacommunity of individuals that are equivalent in terms of birth/death rate and the relative species abundances are random. Local community structure is determined by migration from the metacommunity; therefore, more abundant species at the metacommunity level will be more abundant at the local level (Hubbell, 2001; Chave, 2004). Differences in environmental conditions are not predicted to structure communities since species have equal

fitness under all conditions.

While theories on species coexistence have primarily emerged from observations of macroorganisms, microbial communities offer unique opportunities to test the premises of neutral and niche theories as microbes can be manipulated in well controlled settings *in vitro* and monitored in natural settings with sequencing based approaches. Interestingly, few studies directly compare *in vitro* results to natural communities. Before sequencing technology took off, studies of bacterial coexistence relied on bacteria that could be cultured. In support of niche theory, Wilson & Lindow (1994) examined coexistence in epiphytic bacterial populations by comparing *in vitro* carbon utilization profiles and found that species with similar profiles were less likely to coexist. Recent studies measuring community structure by 16S rRNA gene amplicon sequences or fingerprints have found that neutral models are sufficient to explain species distributions (Woodcock *et al.*, 2007; Finkel *et al.*, 2011); however, other reports note that neutral theory can explain some variation, but describing the full community is not possible without adding environmental parameters (Van der Gucht *et al.*, 2007; Dumbrell *et al.*, 2009; Logue & Lindström, 2010; Langenheder & Székely, 2011).

1.0.2 Study system: The bacterial communities associated with Arabidopsis thaliana

A. thaliana is a small flowering annual dicotyledonous plant native to Europe and Asia, and was introduced to North America during the early stages of European colonization (O'Kane & Al-Shehbaz, 1997). In Eurasia, *A. thaliana* displays fine-scale population structure, with isogenic individuals found at single sites. In contrast, North American accessions from across the continent are isogenic (Platt *et al.*, 2010). *A. thaliana* seeds in the Midwestern United States first germinate in the fall and then overwinter as a young rosette. In the spring, rosettes undergo vegetative growth until they bolt and flower. The life cycle is completed by senescence in late spring. One major benefit to using this plant to study microbial communities is that *A. thaliana* can also be grown gnotobiotic on agar plates under sterile

conditions. In nature, *A. thaliana* is home to a diverse array of bacteria that colonize both the surface of leaves and roots (epiphytes) and the intercellular spaces inside leaves and roots (endophytes). Collectively the epiphytes and endophytes of the leaves are referred to as the phyllosphere whereas the epiphytes and endophytes of the roots are referred to as the rhizosphere. An additional advantage of this system is that most abundant bacterial species in the leaves and roots are cultivatable (Kniskern *et al.*, 2007; Traw *et al.*, 2007; Bodenhausen *et al.*, 2013).

In my dissertation, I characterize the dynamics of these bacterial communities under controlled and natural conditions. Cultivatable isolates allowed me to remove host effects and zero in on bacterial competition and environmental factors that structure communities. By observing natural bacterial communities, I uncovered temporal and spatial factors that affect community composition.

1.0.3 Chapter summaries

In Chapter 2, I describe the benefits and pitfalls to three approaches for quantifying *in vitro* and *in planta* bacterial communities. To characterize bacterial interspecific interactions outside the plant, I first chromosomally inserted fluorescent protein genes into endophytic leaf isolates. These fluorescent tags allowed for high-density time series measurements of absolute organismal abundances for up to two isolates. From pairwise competitions of labeled isolates, I observed a strict competitive hierarchy. Fluorescent marker insertions were only feasible for a limited number of isolates, though. To quantify more species rich communities, I adopted a 16S rRNA gene (16S) amplicon sequencing approach. A caveat of this approach is that due to variation in ribosomal copy numbers, bacterial amplicon abundance does not necessarily equal organismal abundance. To resolve this complication for my controlled communities, I whole genome sequenced each isolate and estimated 16S rRNA gene copy numbers by developing a novel approach to compare read depths for 16S regions to single-copy gene regions. The development of this bioinformatic pipeline led to the publication "16Stimator: statisti-

cal estimation of ribosomal gene copy numbers from draft genome assemblies" (Perisin *et al.*, 2015). Amplicon sequencing, however, requires destructive sampling which complicates inference of bacterial spatial distributions. Traditionally, fluorescence bandpass microscopy has been used to visualize bacterial localization, but this method is obsolete for visualization of multiple isolates in a plant environment with strong autofluorescence. In collaboration with Gary Borisy's laboratory at the Marine Biological Laboratory and Nationwide Histology, I developed leaf sectioning, fluorescence *in situ* hybridization, and spectral imaging techniques to successfully visualize and discriminate fourteen species *in vitro* and a single isolate *in planta*.

In Chapter 3, I characterized the role of environmental and species richness factors in structuring controlled bacterial communities. By inoculating random subsets of twelve endophytic species into three increasingly complex environments (liquid media, solid media, gnotobiotic *A. thaliana*) at specified richness levels, I observed strong environmental and richness effects on community composition. I assessed the predictability of succession by pairwise multivariate dissimilarities of replicate communities. *In planta* communities displayed the highest dissimilarities, followed by liquid then solid media communities. By comparison of average abundances across replicates, I uncovered clear species' abundance hierarchies in which the order of isolates differed by environment. Two species that were barely detected in either liquid or solid media, rose to high abundances *in planta*. These species, *Pseudomonas viridiflava* and *Curtobacterium flaccumfaciens* had opposite effects on plant mass.

In Chapter 4, I characterized the spatial and temporal dynamics of bacterial communities associated with natural *A. thaliana* in Southwest Michigan. I conducted this field study over two years at two geographic sites, and collected soil in addition to root, leaf, stem, and silique tissues. The type of tissue had the largest effect on bacterial community composition. By collecting samples over the plant life cycle, I also observed root associated communities initially shift away from soil communities, in terms of composition, but upon flowering and

senescence, these communities became more soil-like. Additionally, I observed site effects that varied in magnitude by tissue type and suggested that these tissues act as sieves. By comparing overlaps in operational taxonomic unit (OTU) abundances, I uncovered that the roots primarily filter OTUs from the soil. Aboveground tissues then enrich for specific genera and OTUs from this root-filtered pool.

CHAPTER 2

APPROACHES TO CHARACTERIZE BACTERIAL INTERACTIONS IN MIXED SPECIES COMMUNITIES

2.1 Abstract

Microbes are crucial to life on earth and they exert their effects not in monoculture, but in complex multi-species communities. Methods to interrogate interactions between community members have traditionally relied on culture based methods until the recent sequencing technology revolution revealed the impressive diversity of the microbial world. These sequencing technologies also have pros and cons, though. In this chapter, I explore the power and limitations of three approaches to characterizing bacterial interactions in mixed species communities. By chromosomally inserting fluorescent protein genes and competing strains *in vitro*, I collected dense time series abundance data via a plate reader assay. Comparison of each isolate's growth when cultured with other strains revealed a strict competitive hierarchy. This approach is limited to only a handful of strains, which led to an approach shift to 16S rRNA gene (16S) amplicon sequencing to characterize controlled communities. Differences in 16S copy number convolute organismal abundance measurements; however, and I resolved this caveat with whole genome sequencing and the development of "16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies" (Perisin *et al.*, 2015). The destructive sampling required by both culture and sequencing based approaches also destroys the community's spatial structure. Uncovering the spatial organization of bacterial communities within the plant leaf is especially challenging given strong autofluorescence. To overcome these hurdles, I have developed novel fluorescence *in situ* hybridization (FISH) and spectral imaging methods to successfully visualize bacterial cells *in planta*.

2.2 Introduction

The influence of microbes pervades every facet of life on earth. Microbes, especially bacteria, drive fundamental geochemical processes that sustain life (Falkowski *et al.*, 2008) and are tightly associated with living hosts in such a way that we now need to consider each person and plant as a superorganism (Bordenstein & Theis, 2015; Vandenkoornhuyse *et al.*, 2015). Bacteria exert these effects not in pure monoculture, as commonly studied in the laboratory, but as members of complex, multi-species communities, constantly cooperating and competing with each other and higher order organisms. In fact, many macro-ecological principles apply to microbial communities, with application generating new micro-ecology insights (Barberán *et al.*, 2014; O'Dwyer *et al.*, 2015). Microbial ecology has generally lagged behind macro-ecology for the simple reason that microbes are difficult to observe. We feel their presence everywhere we look, from breaking down complex carbohydrates in our guts, to fixing nitrogen into usable forms for plants, but these organisms are invisible to the human eye. Culture-based approaches opened up a whole new world of unexplored life that has recently undergone a massive expansion due to marker gene and metagenomic sequencing technologies (Sogin *et al.*, 2006; Rinke *et al.*, 2013).

These sequencing techniques are not always superior to traditional culture-based methods, though. Sequencing has largely been used as a top-down approach, observing the entire community and inferring interactions in comparison to ecological models (Ofiteru *et al.*, 2010). To tease apart interactions between bacteria, bottom-up strategies involving culture-based methods provide more controlled, high-throughput, and less costly means to observe and assess bacterial competition and cooperation. In this work, we sought to characterize the competitive interactions between leaf endophytic bacteria associated with *Arabidopsis thaliana*. We took advantage of cultivatable isolates that have been found to be among the most naturally abundant species by both culture-dependent and independent means (Kniskern *et al.*, 2007; Traw *et al.*, 2007; Bodenhausen *et al.*, 2013). To allow for dense time series measurements of bacterial abundances, we chromosomally inserted fluorescent

protein genes and followed species abundances by measuring fluorescence intensities with a plate reader. Using eYFP and mCherry markers, we competed labeled species in all pairwise combinations and observed a strict competitive hierarchy.

We next sought to characterize more than two species at a time. The fluorescent marker gene approach is limited to species that can be genetically manipulated and only a few non-overlapping fluorescent protein emissions can be simultaneously measured. Many of these isolates have similar colony morphologies, making CFU plating a non-option. For communities of greater than two species, we chose to characterize community succession by 16S rRNA gene amplicon sequencing. This technique requires destructive sampling but can also be made to be high-throughput. We present the results of this approach to characterize controlled communities in different controlled environments and over a range species richness levels in Chapter 2. In this chapter, we address a key caveat of 16S amplicon sequencing, which is that amplicon abundances do not necessarily equal organismal abundances.

The 16S rRNA gene (16S) dominates marker gene studies of bacterial and archaeal diversity, due to its ubiquity, sequence conservation, and variable regions that allow taxonomic discrimination. However, given that 16S copy number can vary from 1-15 copies (Lee *et al.*, 2009), organismal abundances could be grossly misestimated, producing incorrect diversity estimates (Kembel *et al.*, 2012).

Two methods have emerged to normalize 16S amplicon counts based on copy numbers (Kembel *et al.*, 2012; Langille *et al.*, 2013). Both methods compare observed 16S sequences to a reference database of closed genomes to find related isolates with known copy numbers. An ancestral state is calculated and used to predict the 16S copy number of the organism in question. The accuracy of these methods depends on the relatedness between the observed and reference sequences, with prediction accuracy sharply declining with increasing genetic distance (Langille *et al.*, 2013). A larger and more diverse reference database would improve 16S copy number predictions, especially for samples originating from poorly characterized environments.

There are currently ∼3,200 closed and ∼26,000 draft bacterial genomes deposited in NCBI. Copy numbers for closed genomes are accessible in the Ribosomal RNA Database (*rrn*DB) (Stoddard *et al.*, 2014), but no such database exists for draft genomes due to difficulties in resolving copy numbers of highly conserved, repetitive genomic regions. Assemblies frequently contain the 16S rRNA gene in a single, overrepresented contiguous sequence fragment (contig). Although computational methods have emerged to identify and quantify copy number variation using next generation sequencing data (Zhao *et al.*, 2013; Periwal & Scaria, 2014), these approaches are either designed for polyploids or do not provide the necessary sensitivity for 16S copy number quantification. Of these methods, read depth approaches have shown the most promise (Brynildsrud *et al.*, 2015; Greenblum *et al.*, 2015), but they have not been optimized for draft genomes comprising many contigs.

Building on read depth approaches, we demonstrate accurate and precise quantification of 16S copy number from draft genomes using the 16Stimator pipeline (Figure 2.2). Accuracy was assessed by application to *de novo* assemblies of 12 endophytic bacterial isolates collected from *A. thaliana* leaves, and copy number confirmation by qPCR. We then apply our method to the raw sequence data of draft genomes deposited in NCBI. Thus, we have drastically increased the number of species with estimated 16S copy numbers, thereby improving phylogenetically based abundance estimation.

Even though we have developed a method to correct the 16S copy number caveat of 16S amplicon sequencing of controlled communities, there is still the caveat that the interactions which we infer from sequencing or culture-based methods clearly depend on spatial configuration of community members. This spatial information is lost in destructive sampling for DNA extraction for amplicon sequencing or in tissue homogenization for CFU plating. The leaf environment contains a wide variety of niches, for which bacterial preferences have mainly been explored 1-2 species at a time with fluorescently tagged strains (Monier & Lindow, 2005; Godfrey *et al.*, 2010). Fluorescence imaging *in planta* with bandpass filters is difficult given intense autofluorescence. Recently, Remus-Emsermann *et al.* (2014) successfully

bypassed plant autofluorescence by removing leaf surface associated bacteria with adhesive tape, subjecting these bacteria to fluorescence *in situ* hybridization (FISH), and visualizing spatial distributions with fluorescent bandpass imaging. Standard bandpass filter imaging is limited to simultaneous visualization of only a few non-overlapping fluorophores and becomes convoluted for endophyte visualization due to strong plant autofluorescence. Spectral imaging provides an attractive solution to both of these problems. In spectral imaging, the entire emission spectrum is recorded for each pixel and can then be classified as matching the spectrum from a single fluorophore, combinations of fluorophores, or plant autofluorescence. With Combinatorial Labeling and Spectral Imaging FISH (CLASI-FISH), Valm *et al.* (2011) were able to label, image, and differentiate 15 taxa in dental plaque biofilms. In this work, we adapted CLASI-FISH to label and image 13 different endophytic bacteria *in vitro*. We also developed leaf fixation, sectioning, and FISH protocols to successfully visualize *P. syringae* within the plant. Spectral imaging allowed differentiation of the *P. syringae* specific probe fluorescence from background plant autofluorescence.

2.3 Methods

2.3.1 Endophyte collection

Endophytic bacteria (Table 2.1) were isolated from within leaves of *Arabidopsis thaliana* or *Draba verna* from Southwest Michigan (Kniskern *et al.*, 2007; Traw *et al.*, 2007; Barrett *et al.*, 2011). Isolates were cultured and maintained in King's B media (King *et al.*, 1954).

2.3.2 Fluorescently labeled bacteria

Chromosomal integration of fluorescent marker genes

Escherichia coli strains containing mini-Tn7 vectors with fluorescent protein and selectable antibiotic resistance genes (Table 2.2) were obtained from the laboratory of Dr. Herbert

Schweizer at Colorado State University (Choi & Schweizer, 2006). To introduce these plasmids into endophytic bacterial strains (Table 2.3), either four parental mating or electroporation was used. Upon introduction of the mini-Tn7 and transposase (pTNS3) plasmids, the transposase will excise the mini-Tn7 construct from the plasmid and insert it in the bacterial chromosome at an *att*Tn7 site. These sites are generally found in one location in the chromosome, downstream of the highly conserved glucosamine-6-phosphate synthetase (*glmS*) gene (Choi & Schweizer, 2006). For four parental mating, overnight cultures of *E. coli* containing the conjugative plasmid RK600 (300 μL), *E. coli* containing the mini-Tn7 plasmid (300 μL), *E. coli* containing the transposase plasmid (pTNS3) (300 μL), and the endophytic isolate (100 μL) were mixed together and pelleted by centrifugation at 8000 RPM for 2 min. The cell pellet was washed with 1 mL KB and then resuspended in 30 μL KB. This mixture was spotted onto solid KB agar plates and incubated overnight at 28°C. The following day, a glass scraper was used to remove the cell blob. Cells were resuspended in 150 μL 10 mM $MgSO_4$ and spread plated on KB/gentamicin plates to select for integration of the mini-Tn7 construct. For electroporation, mini-Tn7 and transposase plasmids were first isolated from respective *E. coli* strains. The endophytic bacterial strain was grown overnight in KB media. A 6 mL culture was split into four eppendorf tubes, and cells were pelleted by centrifugation at 8000 RPM for 2 min. Cell pellets were washed two times with 300 mM sucrose (filter sterilized through 0.2 μm pores). Cells were resuspended in 25 μL per tube and combined. Mini-Tn7 and transposase plasmids (500 ng each) were added to cells in a 2 mm gap electroporation cuvette (MidSci, St. Louis, MO, USA). Cells were electroporated at 2.5 KV (BioRad MicroPulser, Hercules, CA, USA), and 1 mL of KB media was immediately added. This culture was allowed to recover and grow at 28°C for 4 h with 200 RPM shaking. Cells were pelleted by centrifugation at 8000 RPM for 2 min, resuspended in 100 μL KB, and spread plated on KB/gentamicin plates. To confirm successful integration of the full mini-Tn7 construct, colonies were screened for fluorescence using Crime Lite 2 flashlights (Foster and Freeman, Sterling, VA, USA) (Blue 420-470 nm for eCFP, Blue/green 450-510

nm for eYFP, Green 500-560 nm for mCherry).

Pairwise competition *in vitro*

Single colonies of each fluorescently labeled isolate were picked from KB plates and grown overnight in 5 mL KB at 28°C with 200 RPM shaking. Cells were pelleted by centrifugation at 8000 RPM for 2 min. Pellets were washed with 1 mL KB, resuspended in 1 mL KB, and diluted to an absorbance at 600 nm of 0.02. Co-cultures were setup in clear 96 well plates (Costar 3370, Corning, Tewksbury, MA, USA) by adding 100 μL of each respective isolate per well (final absorbance at 600 nm of 0.01). Plates were sealed with parafilm and incubated in an Infinite F200 plate reader (Tecan, Morrisville, NC, USA) at 28°C with 2.5 mm orbital shaking. Every 10 min, absorbance at 600 nm, eYFP fluorescence intensity (excitation 510±5 nm, emission 535±10 nm), and mCherry fluorescence intensity (excitation 570±5 nm, emission 600±10 nm) were measured.

2.3.3 16Stimator

Sequence generation and processing

Using a modified protocol from Morgan *et al.* (2010), genomic DNA was isolated from 1 mL of bacterial overnight, KB culture. First, cells were pelleted by centrifugation at 14,000 RPM for 2 minutes. After removal of supernatant, cell pellets were resuspended in TES (10 mM Tris-HCl pH 7.5, 1 mM EDTA, 100 mM NaCl) supplemented with 50 U/μL lysozyme (Ready-lyse Lysozyme Solution, Epicentre, Madison, WI, USA), and incubated overnight at room temperature. Then, SDS and proteinase-k were added to final concentrations of 1% and 0.5 mg/mL, respectively, and incubated at 55°C for 4 hours. Three 2.3 mm stainless steel beads (BioSpec, Bartlesville, OK, USA) were added and lysates were mechanically disrupted at 1,750 RPM for 5 minutes (2010 Genogrinder, SPEX, Metuchen, NJ, USA). DNA was then extracted with the Gentra Puregene Yeast/Bacteria Kit (Qiagen, Valencia,

CA, USA), using the standard protocol.

DNA was quantified by fluorimetry (Quant-iT DNA Assay, Life Technologies, Carlsbad, CA, USA) and sent to the Institute for Genomics and Systems Biology Next Generation Sequencing Core at Argonne National Laboratory (Lemont, IL, USA). Short insert libraries (∼200 bp) were prepared with the standard protocol for the TruSeq DNA Sample Prep kit (Illumina, San Diego, CA, USA), and long insert libraries (insert ∼2500 bp) were constructed with the standard protocols for either the Mate Pair Library Prep Kit v2 or the Nextera Mate Pair Sample Prep Kit (Illumina, San Diego). Actual insert sizes are listed in Table 2.1. Paired end, 101 bp reads were generated using the Illumina HiSeq 2000 platform with base caller CASAVA v1.8.2.

All sequencing processing steps and parameters are included in custom shell scripts that are publicly available at https://bitbucket.org/perisin/16stimator. Briefly, raw reads were quality filtered with Trimmomatic v0.32 (Bolger *et al.*, 2014). The first and last three bases were trimmed due to low sequence quality. A sliding window of four bases was further used to trim reads where the average sequencing quality was below Q15. Any Illumina adapter sequences were removed and all the reads were cropped to the same length. Reads that aligned to PhiX were further removed with Bowtie 2 v2.1.0 (Langmead & Salzberg, 2012). To check for contamination in the DNA isolation and library generation, processed reads were taxonomically classified with Kraken v0.10.4 (Wood & Salzberg, 2014), and contaminating reads were removed. Finally, mate pair reads were reverse complemented with FASTX-Toolkit v0.0.13 (http://hannonlab.cshl.edu/fastx_toolkit/). After processing, some sequencing libraries did not meet the previous quality requirements and were not used in subsequent assemblies and analysis (Table 2.1). All processed reads were deposited in NCBI SRA (Table 2.1). For *Escherichia coli* TY-2482 (Rohde *et al.*, 2011), processed Illumina paired end reads generated from short and long insert libraries were downloaded from SRA (SRR292678, SRR292862).

Draft genome assembly

Interleaved reads were assembled into contigs with the VelvetOptimiser.pl script packaged with Velvet v1.2.10 (Zerbino & Birney, 2008). Final assemblies were generated with a kmer size that maximized N50 length, and deposited into NCBI WGS database (Table 2.1 for accession numbers).

16S rRNA gene copy number estimation for endophytes

16S copy numbers were estimated by a sequencing read depth approach as outlined in Figure 2.2. An annotated genome and sequencing reads are required for estimation. Draft genome assemblies were submitted to the Rapid Annotation using Subsystem Technology (RAST)(Overbeek *et al.*, 2013) server for annotation, and a custom R v3.1.0 (Team, 2015) script was used to extract the positions of 16S rRNA, 23S rRNA, and single-copy genes. To calculate read depths, bed files were created by applying a sliding window of two times the read length over the genome. Read depth can be biased for GC% of the genomic region; therefore, we corrected for GC%, where appropriate, by applying a modified method from Yoon *et al.* (2009). Briefly, processed reads were mapped back to the assembly with Bowtie 2 (Langmead & Salzberg, 2012), and the resulting sam file was converted to a bam file with SAMtools v0.1.19 (Li *et al.*, 2009). The GC% and read depth for each window was calculated via bedtools v2.17.0 (Quinlan & Hall, 2010) nuc and intersect functions, respectively. For single-copy gene windows, read depth was plotted against GC% to test for a linear relationship. If the linear model (lm function in R) explained a significant proportion of the variance ($R^2 > 0.1$, slope $p < 0.05$), then we used the model parameters to correct read depths for single-copy and 16S windows. Otherwise, we used the raw read depths for further calculations. Assemblies can result in partial resolution of 16S copies; therefore, each gene was collapsed to a single contig and read depths were summed for overlapping regions.

Estimates and confidence intervals were calculated by two different methods, Price-Bonett and Permuation. For Price-Bonett, the ratio of median read depths for 16S to single-copy

genes was calculated and the 95% confidence intervals were calculated as in Price & Bonett (2002). For Permutation, a random position of each gene was chosen and the read depth was calculated as above for the window surrounding each position. The copy number was calculated by dividing the 16S read depth by the single-copy read depth. This process was repeated 1000 times and 95% confidence intervals were calculated for the resulting distribution with the Hmisc R package (https://github.com/harrelfe/Hmisc).

16S rRNA gene copy number estimation for NCBI deposited genomes

The table of deposited complete and draft prokaryotic genomes was downloaded from NCBI on November 25, 2014. This table was parsed for assemblies that are annotated and have sequencing reads in SRA. For each sequencing read/assembly combination, copy numbers were computed using the Price-Bonett method outlined above, except annotations were downloaded directly from NCBI. The results were further parsed for sequencing coverage ($>$ 10,000 reads), proportion of mapped reads ($>$ 50%), and outliers of 16S copy number estimates ($<$ 20 copies). All estimates and metadata for NCBI genomes can be found in Supplementary Table S2 in Perisin *et al.* (2015).

Experimental validation of 16S rRNA gene copy number

We identified single-copy genes (*argS*, *atpD*, *ppK* or *valS*) for each bacterial isolate, designed and quality tested quantitative PCR primers (Table 2.4), and constructed gBlocks (Integrated DNA Technologies, Coralville, IA, USA) for reference. An individual gBlock reference contained one single-copy amplicon and one 16S amplicon, comparable to a reference plasmid (Lee *et al.*, 2008). The concentration of genomic DNA samples was estimated by fluorimetry (Quant-iT DNA Assay, Life Technologies, Carlsbad, CA, USA) and adjusted to 3ng/μL. gBlocks were adjusted to 0.1 pg/μL. A threefold dilution series with seven total dilutions was produced for each sample and gBlock. The final volume per PCR was 11 μL and contained 4.4 μL molecular grade water, 5 μL SYBR FAST ABI Prism qPCR Kit reagent

(KK4604, KAPA Biosystems, Wilmington, MA, USA), 0.3 μL of 10 μM forward and reverse primers, and 1 μL template. PCR amplification consisted of an initial denaturation step of 5 minutes at 95°C, followed by 2 cycles of denaturation at 95°C for 15 s, annealing at 67°C for 15 s, elongation at 72°C for 15 s, followed by 2 cycles of denaturation at 95°C for 15 s, annealing at 65°C for 15 s, elongation at 72°C for 15 s, followed by 2 cycles of denaturation at 95°C for 15 s, annealing at 60°C for 15 s, elongation at 72°C for 15 s, followed by 40 cycles of denaturation at 95°C for 10 s, annealing at 58°C for 15 s, elongation at 72°C for 15 s, followed by a final cycle of denaturation at 95°C for 15 s, annealing at 60°C for 15 s, elongation at 72°C for 15 s. qPCR runs were carried out in three technical replicates per primer and dilution on an ABI PRISM 7900HT (Applied Biosystems Instruments, Life Technologies, Carlsbad, CA, USA). Dissociation curves for each run were inspected for contamination and analysis performed by ABI PRISM software SDS 2.4 with the automatic C_t setting.

Copy numbers were determined by absolute quantification (Lee *et al.*, 2008). All calculations were performed in R v3.1.0 (Team, 2015). First, the weight of gBlock DNA for each dilution was converted into number of copies. Standard curves for 16S and single-copy genes were constructed for C_t vs. log(Copies) and fitted with a linear model (lm function). The regression parameters were used to calculate the number of gene copies for 16S and single-copy gene based on the genomic DNA C_t values. The 16S:single-copy gene copy ratio at each dilution was used to determine copy number per isolate.

16S rRNA copy number estimate comparison for 16Stimator and PICRUSt

To address whether 16Stimator represents an improvement in copy number prediction over ancestral reconstructions methods, we compared our estimates to PICRUSt predictions for Greengenes 13_5 operational taxonomic units (OTUs) (McDonald *et al.*, 2012). We first extracted all the 16S sequences for all genomes with 16Stimator copy number estimates. We aligned these sequences against

the Greengenes 99% OTUs (pick_closed_reference_otus.py in QIIME with uclust, enable_rev_strand_match True, and similarity 0.99) (Caporaso *et al.*, 2010b; Edgar, 2010). With the precalculated table found on the PICRUSt website (16S_13_5_precalculated.tab, http://picrust.github.io/picrust/picrust_precalculated_files.html), we were able to match PICRUSt copy number estimates for each OTU. Some genomes contained multiple 16S sequences that aligned to different 99% OTUs. We excluded these genomes from further analysis. As each genome had multiple 16Stimator estimates from multiple reads and sequencing libraries, we took the mean estimate and confidence interval for each genome. As several Greengenes OTUs encompassed multiple genomes, we then took the median estimate for each OTU. To calculate Pearson correlations between 16Stimator and PICRUSt estimates at different nearest sequenced taxon index (NSTI) cutoffs, we used the cor.test function in R (Team, 2015). The table of median 16Stimator copy number estimates for the Greengenes 99% OTUs for direct application in the PICRUSt workflow can be found in Supplementary Table S4 in Perisin *et al.* (2015).

2.3.4 Spatial visualization of bacterial communities

Bacterial culture fixation

Bacterial isolates were grown overnight in 5 mL at 28°C with 200 RPM shaking. Cells (1 mL) were pelleted by centrifugation at 13,000 RPM for 2 min. Pellets were resuspended in 500 μL PBS. Cells were fixed by addition of 500 μL fresh 4% paraformaldehyde pH 7.0 and incubated at room temperature for 2 h. Fixed cells were pelleted by centrifugation at 13,000 RPM for 2 min and resuspended in 1 mL PBS. After two more PBS washes, cells were resuspended 50% PBS 50% ethanol and stored at -20°C.

Combinatorial labeling and spectral imagining fluorescence *in situ* hybridization (CLASI-FISH) of bacterial cultures

CLASI-FISH was completed at the Marine Biological Laboratory in Woods Hole, MA, USA in collaboration with Blair Rossetti and Jessica Mark-Welch in the laboratory of Dr. Gary Borisy.

Fixed bacterial cells in 50% PBS 50% ethanol were pelleted by centrifugation at 13,200 RPM for 5 min. Cells were resuspended in 100 μL of hybridization solution (900 mM NaCl, 20 mM Tris-Cl pH 7.5, 0.01% SDS, 20% deionized Hi-Di formamide, each oligonucleotide probe at 2 pmol/μL). Each isolate was labeled with a different combination of fluorescently labeled EUB338 probes custom synthesized by Invitrogen (Thermo Scientific, Waltham, MA, USA) (5'-GCT GCC TCC CGT AGG AGT-3' with either 5' conjugated Alexa Fluor 488, Alexa Fluor 514, Alexa Fluor 555, Alexa Fluor 647, Rhodamine Red-X, or Texas Red-X; 200 pmol/μL stock in 10 mM Tris-Cl pH 7.5). The EUB338 probe targets a universal, conserved region of the 16S rRNA (Amann *et al.*, 1990; Valm *et al.*, 2011). Hybridization took place in the dark at 46°C for 2 h. Afterwards, cells were pelleted by centrifugation at 13,200 RPM for 10 min, and resuspended in 190 μL warm wash buffer (215 mM NaCl, 20 mM Tris-Cl pH 7.5, 5 mM EDTA). Cells were incubated in the dark at 48°C for 15 min and pelleted by centrifugation at 13,200 RPM for 10 min. Cells were resuspended in wash buffer (20-200 μL) and spotted onto UltraStick slides (Thermo Scientific). Slides were placed in a dark humid chamber at room temperature for 1 h to allow cells to settle. Slides were dipped in ice cold H_2O and 100% ethanol to remove salts. After air drying, one drop of ProLong Gold antifade with DAPI was added and samples were covered with a coverslip (#1.5 thickness, Thermo Scientific), and allowed to cure flat and in the dark at room temperature overnight.

Spectral images were acquired with a Zeiss LSM 780 confocal laser scanning microscope (Carl Zeiss Microscopy, Jena, Germany) with a 40x/1.4 Plan-Apochromat objective lens. For each field of view over the 212.4 μm X 212.4 μm scan area, spectral images were acquired for each of six lasers and in descending order to reduce photo bleaching (633 nm, 594 nm,

561 nm, 514 nm, 488 nm, 405 nm). Images were acquired as the average of two line scans. Image filtering, linear unmixing, cell assignment, and psuedocoloring was completed as in Valm *et al.* (2011).

Pseudomonas syringae FISH probe design and specificity confirmation

Using the probe design function of ARB software (Westram *et al.*, 2011) along with the SILVA database (Pruesse *et al.*, 2007), oligonucleotide probes targeted against 16S rRNA were designed to match to as many *P. syringae* sequences as possible while providing mismatches against closely related *Pseudomonas* species. Mismatches were designed to be near the ends of the probe for *P. syringae* sequences and to be centrally located for unwanted targets. Using the mathFISH program (Yilmaz *et al.*, 2011), probes were further screened for binding efficacy given the hybridization conditions listed in the previous section. A probe was selected based on these requirements (NP-H: 5'-GGC AGC ACG GGT ACT TGT ACC-3' with 5' conjugation to Alexa Fluor 555) and was custom synthesized by Invitrogen (Thermo Scientific).

To experimentally test the specificity of the NP-H probe, fixed cells for all *Pseudomonas* isolates in Table 2.3 were used in FISH reactions, as in the previous section, with the NP-H and EUB338-I (5' conjugated to Alexa Fluor 647) probes. Spectral images were acquired with a LSM 710 confocal laser scanning microscope (Carl Zeiss Microscopy) with a 40x/1.1 W LD C-Apochromat WD 0.62 mm objective. Samples were excited with the 633 nm and 561 nm lasers for Alexa Fluor 647 and Alexa Fluor 555 fluorophores, respectively. Lambda mode was used to capture the emission spectrums and images were colored using the composite of the lambda stack.

Gnotobiotic plant inoculation with *P. syringae*

Seeds for *A. thaliana* accession PT1.85 (ID 8057), originally collected from Hanna, IN, USA (41.3423, -86.7368), were surface sterilized by addition of 70% ethanol, incubation at room

temperature for 1 min, removal of ethanol, addition of 100% ethanol, incubation at room temperature for 1 min, and removal of ethanol. Seeds were air dried in a sterilized biological cabinet. Plants were grown by placing single seeds into each well of 24 well plates (Costar 3378, Corning, Tewksbury, MA, USA) containing 1 mL of Murashige and Skoog (MS) media (4.4 g MS (Duchefa, Haarlem, Netherlands), 10 g sucrose, 0.5 g MES, 6 g plant tissue culture agar (plantmedia, Dublin, OH, USA), pH 5.7-5.8, per 1 L). Seeds were stratified for 4 days at 4°C, and grown in long day conditions (16 h light at 21°C, 8 h dark at 18°C) with 50% relative humidity in a growth chamber (SANYO, MLR-351). Plants were grown for 15 days prior to inoculation.

For inoculation with *P. syringae* NP29.1a, an overnight KB culture was pelleted by centrifugation at 9000 RPM for 2 min. Cells were washed with 10 mM $MgSO_4$ and diluted in 10 mM $MgSO_4$ to an absorbance at 600 nm of $10^{-6} \approx 10^3$ CFU/mL. Plants were flood inoculated by adding 1 mL of inoculum to each well. After room temperature incubation for 2 min, the inoculum was removed, and plants were allowed to air dry in a sterilized biological cabinet. Plates were sealed with parafilm and placed back in the growth chamber. The inoculum was diluted and spread plated on KB to ensure bacterial viability and correct dosage.

Infected plant fixation and sectioning

At nine days post inoculation, plants were harvested and roots removed. The rosette was placed in a biopsy cassette and fixed in FAA fixative (50% ethanol, 10% formaldehyde, 5% acetic acid). Fixation was completed at room temperature for seven days to allow for full penetration of the fixative into the plant tissue. This fixation protocol was also confirmed to successfully fix *P. syringae* cultures for FISH.

Rosette leaves were sent to Nationwide Histology LLC (http://www.nationwidehistology.com) for sectioning. Sectioning was completed by leaf incubation in an ethanol series, 5 min each at 70%, 80%, 95% (two times), and 100%

(three times). Leaves were then incubated for 10 min each in xylene and clearing agent (two times). Leaves were embedded in paraffin overnight, and 4 μm sections were taken perpendicular to the midvein.

In planta FISH and spectral imaging

Paraffin was removed from leaf sections by incubation in xylene for 10 min at room temperature (3 times), followed by rehydration with an ethanol series of 10 min incubations in 100% (2 times), 95%, 80%, and 70% ethanol. Sections were finally incubated for 10 min in molecular grade H_2O and then air dried. FISH was completed directly on the slide by incubating sections in hybridization buffer (recipe described above) with EUB338-B (5' conjugated to Alexa Fluor 488) and NP-H probes. Hybridization took place overnight at 46°C in a dark, humid chamber. Washing and sample setting were completed as outlined above.

Spectral images were acquired with a LSM 710 confocal laser scanning microscope (Carl Zeiss Microscopy) with a 20x/0.8 Plan-Apochromat WD 0.55 mm objective for tile scans, and 40x/1.1 W LD C-Apochromat WD 0.62 mm objective for closer examination of bacterial localization. Samples were excited with the 561 nm and 488 nm lasers for Alexa Fluor 555 and Alexa Fluor 488 fluorophores, respectively. Lambda mode was used to capture the emission spectrums and images were colored using the composite of the lambda stack. The Zeiss ZEN2012 software was used for linear unmixing of selected areas.

2.4 Results

2.4.1 Fluorescently labeled bacteria

To distinguish between bacteria in liquid culture, we used the mini-Tn7 transposon system to chromosomally label seven endophytic species with fluorescent protein genes (Table 2.3). In this method, a mini-Tn7 plasmid encoding a fluorescent protein under control of a constitutive promoter, and a helper plasmid (pTNS3), encoding the TnsABCD transposase

machinery, are introduced into bacteria by electroporation or conjugation. Transposition machinery inserts the mini-Tn7 25 bp downstream of glucosamine-6-phosphate synthetase (*glmS*) at an *att*Tn7 site (Choi & Schweizer, 2006). These markers allow for high-throughput and high-density time series measurements of bacterial abundances using a plate reader assay.

With this assay, we competed fluorescently labeled strains in all pairwise combinations in liquid KB media, which we confirmed to properly support the growth of each isolate. In the plate reader, we monitored the fluorescence intensities of eYFP and mCherry as proxies for bacterial growth. In Figure 2.1, we show the outcomes of pairwise competitions between three isolates (*P. simiae* MEB105, *P. fulva* MEJ086, *P. syringae* NP29.1a). Each isolate was labeled with eYFP or mCherry and co-cultured with an isolate with the opposite fluorescent tag. We inoculated strains 1:1 at low densities, and, as controls, we included co-cultures of each isolate with itself (1:1, eYFP:mCherry labeled). To judge winners and losers in each competition, we qualitatively assessed growth (rate and carrying capacity) in the presence of each isolate as compared to co-culture with itself labeled with the opposite tag. For example, in the top panels of Figure 2.1, we observed that *P. fulva* grew to a higher level when co-cultured with *P. syringae* than when co-cultured with itself. *P. fulva* growth was decreased in the presence of *P. simiae*. For *P. simiae*, co-cultures with *P. fulva* and *P. syringae* both resulted in enhanced growth of *P. simiae*. For *P. syringae*, co-cultures with *P. simiae* and *P. fulva* both resulted in decreased growth of *P. syringae*. These competition outcomes indicate a hierarchy with *P. simiae* on top followed by *P. fulva* and then *P. syringae*. We completed these assays for all other pairwise combinations of isolates and observed a strict competitive hierarchy (Table 2.3).

2.4.2 16Stimator

To test the 16Stimator computational pipeline, we processed similarly generated sequencing reads for *Escherichia coli* TY-2482 (GCA_000217695.2, SRR292678, SRR292862), *Bacteroides fragilis* HMW 615 (GCA_000297735.1, SRR488169, SRR488170), *Pseudomonas*

aeruginosa PAO1 (GCA_000006765.1, SRR032420, SRR032832), and *Staphylococcus aureus* KPL1828 (GCA_000507725.1, SRR835799, SRR958927). These isolates have closed genomes or are closely related to isolates with closed genomes. As demonstrated in Figure 2.3, 16Stimator accurately estimates 16S copy numbers. The confidence of each estimate varies and depends on sequencing coverage variability within the 16S and single-copy gene regions. The permutation method for generating confidence intervals captures this variability to a greater extent than the Price & Bonett method, which assumes the ratio of coverage medians is asymptotically normal and thereby lessens the influence of outliers (Price & Bonett, 2002).

To extend the database of genomes and draft-genomes with known 16S copy numbers, we applied 16Stimator to all annotated genomes with corresponding SRA sequence reads that were published in NCBI by November 25, 2014. Of the 29,315 genome assemblies available, only 3,291, representing 1,519 unique bacterial species, were closed. After parsing our results for sequencing coverage, proportion of mapped reads, and outliers of 16S copy estimates, application of 16Stimator resulted in estimates for 26,353 different sequencing libraries representing 816 unique species, including 586 species without closed genomes. To test the accuracy of our method, we compared 16Stimator estimates to actual copy numbers for closed genomes and observed a strongly positive, linear correlation (Figure 2.4a, $R^2 = 0.82$ [0.75, 0.87], $p <$ 2e-16). The median absolute proportional deviation of our estimates to the actual copy numbers was 0.14. Furthermore, we found consistent reliability across sequencing technologies, except for estimates generated from the Illumina HiSeq 1000, Ion Torrent PGM, and PacBio RS; note that there are few published draft genomes with corresponding sequencing reads for these platforms (Figure 2.4b). The large spread of points for the Illumina MiSeq category corresponds to estimates for *Listeria monocytogenes*. Though the genomes analyzed were annotated to have six 16S copies, *rrn*DB lists completed genomes with five or six copies (Stoddard *et al.*, 2014). This variation in true copy number increases the variation in the differences between our estimates and the actual copy numbers. Further, we observed effects on the accuracy of our estimates due to interactions between low sequencing

coverage and GC correction of read depths (Figure 2.5). For low coverage sequencing libraries with GC correction, we underestimated copy numbers by 1-2 copies. For low coverage sequencing libraries without GC correction, we overestimated copy numbers by 1-5 copies. These libraries did not meet the thresholds for GC correction and provide future grounds for improvement of our method. Though this bias seems specific to *L. monocytogenes* MiSeq sequencing, we have included options in our scripts to calculate copy numbers with or without corrections so each researcher can better judge GC effects on 16S copy number estimation.

Ancestral reconstruction methods currently rely on a reference database of closed genomes, primarily human-associated microbes, to generate accurate metagenome predictions and abundance corrections from 16S amplicon surveys. The accuracy of these methods decreases with sequence divergence to reference, described by Langille *et al.* (2013) as nearest sequenced taxon index (NSTI). 16Stimator can fill holes in the reference database by generating 16S copy number estimates from draft assemblies, obviating the need to fully close each genome. Alternatively, ancestral reconstruction methods could predict 16S copy number for a single genome based on its 16S sequence relatedness to reference closed genomes. From cross-validation of 16S copy number estimation, Langille *et al.* (2013) showed PICRUSt accuracy decreased with increasing NSTI. Since our estimates are solely based on sequencing reads from a single genome, without comparison to references, they do not suffer from decreased performance with increasing NSTI. When we compared our method to PICRUSt for Greengenes 99% OTU copy number predictions (McDonald *et al.*, 2012), there was an overall positive correlation between estimates (Pearson correlation coefficient = 0.50, $p < 0.001$), but this correlation deteriorates with NSTI (Figure 2.6). We did not observe an effect of NSTI (Figure 2.7) or phylogeny (Figure 2.8) on confidence interval size for our estimates for the Greengenes OTUs. We therefore conclude that 16Stimator provides more accurate estimates than PICRUSt for 16S copy number prediction from single genomes.

2.4.3 Spatial visualization of bacterial communities

To visualize the spatial distribution of bacterial communities *in planta*, several hurdles needed to be cleared. First, we needed to simultaneously label and image multiple bacterial species *in vitro*. To confirm successful fixation and FISH protocols given the wide variety of endophytic species (Table 2.1), we fixed cultures of each isolate separately and performed FISH using a different combination of fluorophore conjugated EUB338 oligonucleotide probes for each species, as in Valm *et al.* (2011). After mixing labeled strains, we were able to differentiate isolates by spectral imaging. Each combination of fluorophores has a unique emission spectrum which we used to discriminate isolate identities (Figure 2.9). Next, we needed to specifically label isolates in a mixed species community. We designed a 16S rRNA targeted probe (NP-H) specific for *P. syringae* with central mismatches to closely related species. We confirmed specificity with FISH using fixed cultures of *P. syringae* and closely related *P. viridiflava*. We successfully labeled *P. syringae* with both EUB338 and NP-H probes and *P. viridiflava* with only the EUB338 probe (Figure 2.10). Finally, to visualize bacteria within the plant, we needed to discriminate probe fluorescence from plant autofluorescence. To this end, we infected gnotobiotic *A. thaliana* with *P. syringae* NP29.1a. In collaboration with Nationwide Histology LLC, we developed plant fixation and sectioning methods. After FISH of these sections with the NP-H probe, we successfully visualized *P. syringae* within the plant leaf (Figure 2.11a). Spectral imaging allowed for differentiation of the plant autofluorescent spectrum from the NP-H probe fluorescence (compare Figure 2.11b to Figure 2.10b), thus confirming the specific presence of *P. syringae*.

2.5 Discussion

In this chapter we developed three different approaches for characterization of bacterial communities, each with different benefits and pitfalls. With fluorescently labeled isolates, we were able to collect dense time series data of abundances and compare the outcomes

of pairwise competitions *in vitro*; however, we were limited to a group of strains, mostly *Pseudomonas*, that we could genetically manipulate. These labeled strains do hold value for more than just plate reader assays, though. Historically, CFU plating has been the method of choice for assessing bacterial growth *in planta*. This method becomes obsolete with co-infection of morphologically similar strains, but if one of these strains contains a fluorescent tag, then CFU plating can still be used with the aid of excitation flashlights to differentiate colonies. The use of 16S amplicon sequencing in this case may also be ineffective if strains contain identical 16S sequences. By linking fluorescence with organismal abundances, either through fluorescence intensity measured by the plate reader or CFU counts, these labeled strains provide a method to determine absolute abundances, as opposed to 16S amplicon sequencing which outputs relative abundances.

Another caveat of 16S amplicon sequencing is that due to differences in 16S copy numbers, amplicon abundances do not equal organismal abundances, thus confounding diversity estimates. When comparing β-diversity between samples with similar species profiles, these copy number complications cancel out, but α-diversity estimates remain confounded. 16S amplicon sequencing does provide an excellent method to obtain a whole community snapshot and can be made high-throughput by automated DNA extraction and PCR amplification protocols using liquid handling robotics. To account for 16S copy number differences between our endophytic isolates, we decided to whole genome sequence our isolates. Assembly of sequencing reads; however, resulted in collapse of 16S regions to one or few contigs. To resolve 16S copy numbers we developed 16Stimator, an approach that statistically estimates 16S copy number by comparison of read depths for 16S to single-copy, conserved genes. We further experimentally confirmed 16S copy number estimates with qPCR. We then discovered that copy number collapse was a common occurrence in deposited draft genomes. Application of our 16Stimator pipeline to these draft genomes has drastically increased the number of bacterial species and strains with estimated ribosomal copy numbers. This expanded database will increase the power of ancestral state reconstruction methods for metagenome predictions

and organismal abundance corrections for 16S amplicon surveys. We recommend that this method be applied to new genome submissions in NCBI to continually expand the database. As sequencing costs decline and read lengths increase, we envision higher quality genome assemblies with fully resolved repetitive regions. Nevertheless, our estimates greatly enhance the database now and provide an interim means to continue improving it.

Another caveat of 16S amplicon sequencing is that all spatial information is lost due to destructive sampling. Bacterial spatial distributions certainly influence capacities and opportunities to compete or cooperate. Fluorescence microscopy of leaf surface associated bacteria has thus far indicated non-random abundance distributions (Monier & Lindow, 2005; Remus-Emsermann *et al.*, 2014), indicating structural forces and niches that remain undetected in sequencing and CFU assays. Insights into spatial structure within the leaf have been inaccessible due to strong plant autofluorescence. To address these limitations, we developed FISH and spectral imaging techniques to visualize multiple species simultaneously and in the presence of autofluorescence.

2.6 Acknowledgments

Thanks to Madlen Vetter for performing the qPCR experiments, and for reviews of the accepted manuscript. Thanks to Sarah Owens for whole genome sequencing resources. Thanks to Jack Gilbert for manuscript reviews and insights into figure design. Thanks to John Wilmes and Stefano Allesina for helpful discussions, and the Center for Research Informatics at the University of Chicago for computational resources. Thanks to the Confocal Digital Imaging Facility in the Department of Organismal Biology and Anatomy at the University of Chicago for microscopy resources and training. Thanks to Blair Rossetti and Jessica Mark-Welch for CLASI-FISH training. Finally, thanks to Heather Marlatt at Nationwide Histology for development of leaf sectioning protocols.

2.7 Tables and figures

Isolate	Organism	host	isolation_source	collection_date	geo_loc_name	sample_type	collected_by	env_biome	host_tissue_sampled	lat_lon
LMCA8	*Pseudomonas viridiflava*	*Arabidopsis thaliana*	leaf	4/22/08	Lake Michigan College Berrien County, MI, USA	cell culture	Luke Barrett	plant-associated habitat	leaf	42.090114 N 86.393408 W
LMCP11	*Xanthomonas campestris*	*Arabidopsis thaliana*	leaf	4/22/08	Lake Michigan College Berrien County, MI, USA	cell culture	Luke Barrett	plant-associated habitat	leaf	42.090114 N 86.393408 W
MEDvA23	*Variovorax paradoxus*	*Draba verna*	leaf	4/22/08	Michigan State University Extension Berrien County, MI, USA	cell culture	Luke Barrett	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEB024	*Frigoribacterium sp.*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Brian Traw	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEB061	*Flavobacterium sp.*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Brian Traw	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEB064	*Rhodococcus sp.*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Brian Traw	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEB105	*Pseudomonas simiae*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Brian Traw	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEB126	*Curtobacterium flaccumfaciens*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Brian Traw	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEJ076	*Agrobacterium tumefaciens*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Joel Kniskern	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEJ086	*Pseudomonas fulva*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Joel Kniskern	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEJ108Y	*Microbacterium sp.*	*Arabidopsis thaliana*	leaf	5/4/04	Michigan State University Extension Berrien County, MI, USA	cell culture	Joel Kniskern	plant-associated habitat	leaf	42.0927 N 86.356322 W
MEP34	*Pseudomonas fluorescens*	*Arabidopsis thaliana*	leaf	4/22/08	Michigan State University Extension Berrien County, MI, USA	cell culture	Luke Barrett	plant-associated habitat	leaf	42.0927 N 86.356322 W

Isolate	BioProject	BioSample	WGS	SRA_pe	SRA_mp	insert_pe	insertSD_pe	insert_mp	insertSD_mp	Coverage
LMCA8	PRJNA270634	SAMN03267840	JXQO00000000	SRR1721669	SRR1736718	206	27	3167	1824	34
LMCP11	PRJNA270636	SAMN03267868	JXQP00000000	SRR1736253	SRR1736254	243	35	2333	1405	61
MEDvA23	PRJNA270638	SAMN03267909	JXQQ00000000	SRR1736370	NA	260	43	NA	NA	60
MEB024	PRJNA273162	SAMN03267928	JXQZ00000000	SRR1798561	SRR1798572	306	64	71	48	152
MEB061	PRJNA270642	SAMN03267948	JXQR00000000	NA	SRR1736377	NA	NA	2692	1004	30
MEB064	PRJNA270644	SAMN03267949	JXQS00000000	SRR1736481	SRR1736514	249	48	49	26	61
MEB105	PRJNA270647	SAMN03267954	JXQT00000000	SRR1736635	SRR1736646	223	38	2666	1132	98
MEB126	PRJNA270649	SAMN03267955	JXQU00000000	SRR1736662	SRR1736696	206	31	1359	1318	87
MEJ076	PRJNA270650	SAMN03267957	JXQV00000000	SRR1736719	SRR1736720	238	32	2470	757	70
MEJ086	PRJNA270651	SAMN03267958	JXQW00000000	SRR1736721	SRR1736725	236	33	39	23	72
MEJ108Y	PRJNA270652	SAMN03267959	JXQX00000000	SRR1736789	SRR1736790	244	43	74	46	78
MEP34	PRJNA270653	SAMN03267960	JXQY00000000	SRR1736791	SRR1736792	300	50	2657	924	50

Isolate	sum	num_contigs	min_contig_len	med_contig_len	mean_contig_len	max_contig_len	n50	n50_len	n90	n90_len	kmer
LMCA8	5992488	73	220	51480	82088	397726	11	178639	35	55211	65
LMCP11	5063135	47	204	46448	107726	739539	5	270373	19	78071	69
MEDvA23	6906509	138	201	25816	50047	358288	21	103570	60	39590	63
MEB024	3406371	136	202	11749	25046	281555	16	61600	64	13934	65
MEB061	5747109	80	213	1695	71838	613618	7	288355	21	90807	67
MEB064	4664255	81	207	24547	57583	368339	11	133542	33	53201	71
MEB105	6105492	44	202	40681	138761	942084	5	461104	17	107738	71
MEB126	3684420	92	227	17151	40048	272977	12	91243	39	25423	71
MEJ076	5122354	61	207	16820	83973	687831	6	328370	20	63687	75
MEJ086	5891415	125	201	9126	47131	360058	13	163091	42	42863	73
MEJ108Y	3669952	43	602	52281	85347	596739	8	143866	24	43452	71
MEP34	5914240	58	205	19069	101969	482063	9	272244	22	96202	75

Table 2.1: Endophytic isolate collection data and genome assembly summary statistics.

Name	Strain	Vector contained	Antibiotic resistance
E3275	HPS1	pUC18T-mini-Tn7T-Gm-PrrnBP1::gfpmut3T0T1	100 μg/mL ampicillin, 15 μg/mL gentamicin
E3322	HPS1	pUC18T-mini-Tn7T-Gm-PA1/03/04DsRedExpressT0T1	100 μg/mL ampicillin, 15 μg/mL gentamicin
E3348	HPS1	pUC18T-mini-Tn7T-Gm-PA1/03/04ecfpT0T1	100 μg/mL ampicillin, 15 μg/mL gentamicin
E3350	HPS1	pUC18T-mini-Tn7T-Gm-PA1/03/04eyfpT0T1	100 μg/mL ampicillin, 15 μg/mL gentamicin
E4728	DH5α	pUC18T-mini-Tn7-Gm-mcherry	15 μg/mL gentamicin
E4744	DH5α	pUC18T-mini-Tn7-Gm-P1-egfp	15 μg/mL gentamicin
E3972	DH5α(λ pir)	pTNS3	100 μg/mL ampicillin

Table 2.2: *Escherichia coli* strains containing mini-Tn7 vectors (Choi & Schweizer, 2006)

Isolate	Putative species	Hierarchy order	mini-Tn7 gentamicin selection	Fluorescent protein genes chromosomally integrated
MEB105	*Pseudomonas simiae*	1	25 μg/mL	eCFP, eYFP, mCherry
MEP34	*Pseudomonas fluorescens*	2	25 μg/mL	eCFP, eYFP, mCherry
MEJ086	*Pseudomonas fulva*	3	5 μg/mL	eCFP, eYFP, mCherry
RMX3.1b	*Pseudomonas viridiflava*	4	2.5 μg/mL	eCFP, eYFP, mCherry
NP29.1a	*Pseudomonas syringae*	5	2.5 μg/mL	eCFP, eYFP, mCherry
LMCP11	*Xanthomonas campestris*	6	2.5 μg/mL	eCFP, eYFP, mCherry
MEDvA23	*Variovorax paradoxus*	7	7.5 μg/mL	eCFP, eYFP, mCherry

Table 2.3: Endophytic isolates chromosomally labeled with fluorescent protein genes.

Isolate	Single copy gene primers 5'-3'			16S rRNA gene primers 5'-3'		
	Amplicon name	Forward	Reverse	Amplicon name	Forward	Reverse
LMCA8	valSg	CCAGAAGATGTCCAAGTCCA	CCCAKRTCGAACTTGATGTC	1254_1492	CGTGTCGTGAGATGTTGGGTT	TACGGCTACCTTGTTACGACTT
LMCP11	ppKSs	GGCAAGGATTCGCTGATTGT	CCAGATTGGCTTCTTCGTCG	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEDvA23	atpDs	ATTGTCCAATGTATCGGCGC	TGAACTTCGAGGGTCAGTCC	16Spec	TACGTAGGGTGCAAGCGTTA	AGTCACAGATGCAGTTCCCA
MEB024	ppKs	TCCACATCGAGAACTGGTCC	GAGATGTAGGGGAAGGGGTG	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEB061	atpDs	TGGATTGTTTGGTGGTGCAG	TGAAAGTCCACCGTGACCTT	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEB064	valSs	GATGATCAAACCGGCGTACC	TCCGGAGTTCTGTCAGCAAT	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEB105	atpDg	CATCTCGTGGTAGAAGTCGTT	GCATCAAGGTNATCGACCTG	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEB126	atpDs	AGAGGACAAGATCACCGTCG	ACTCGATGGTGTCCTTGAGG	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEJ076	argSs	TGATTAAGCGCCGCAGATTC	GTTGGTTGATGGTGCGTCTT	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEJ086	ppKg	CGYAACGGCAAGGAAGTCA	AGGATSAGCATCATCTTGG	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEJ108Y	atpDs	GGTCAAGCAGATCCTCCAGA	TGTAGGTGTTCTGCGAGAGG	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT
MEP34	valSg	CCAGAAGATGTCCAAGTCCA	CCCAKRTCGAACTTGATGTC	1361_1492	GGAAGGTGGGGATGACGT	TACGGCTACCTTGTTACGACTT

Table 2.4: Quantitative PCR primers for 16S rRNA gene and single copy gene amplifications.

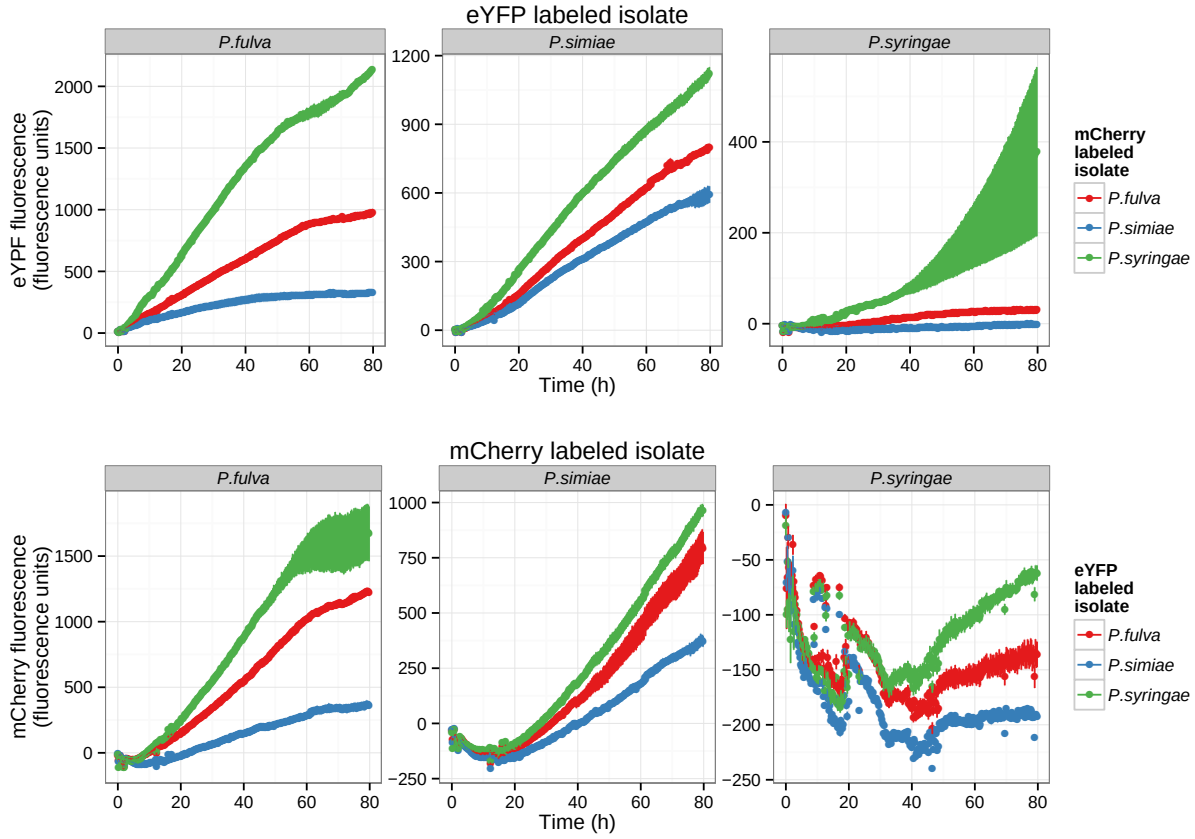


Figure 2.1: Pairwise competition in liquid KB monitored by fluorescence plate reader assays.

Isolates (*P. simiae* MEB105, *P. fulva* MEJ086, and *P. syringae* NP29.1a) were inoculated in all pairwise combinations into KB media. Each isolate's abundance was monitored over time by eYFP (top panels) or mCherry (bottom panels) fluorescence in a plate reader assay. Each of the top panels represent the eYFP fluorescence for each isolate in the presence of mCherry labeled isolates, including itself. Each of the bottom panels represent the mCherry fluorescence for each isolate in the presence of eYFP labeled isolates, including itself.

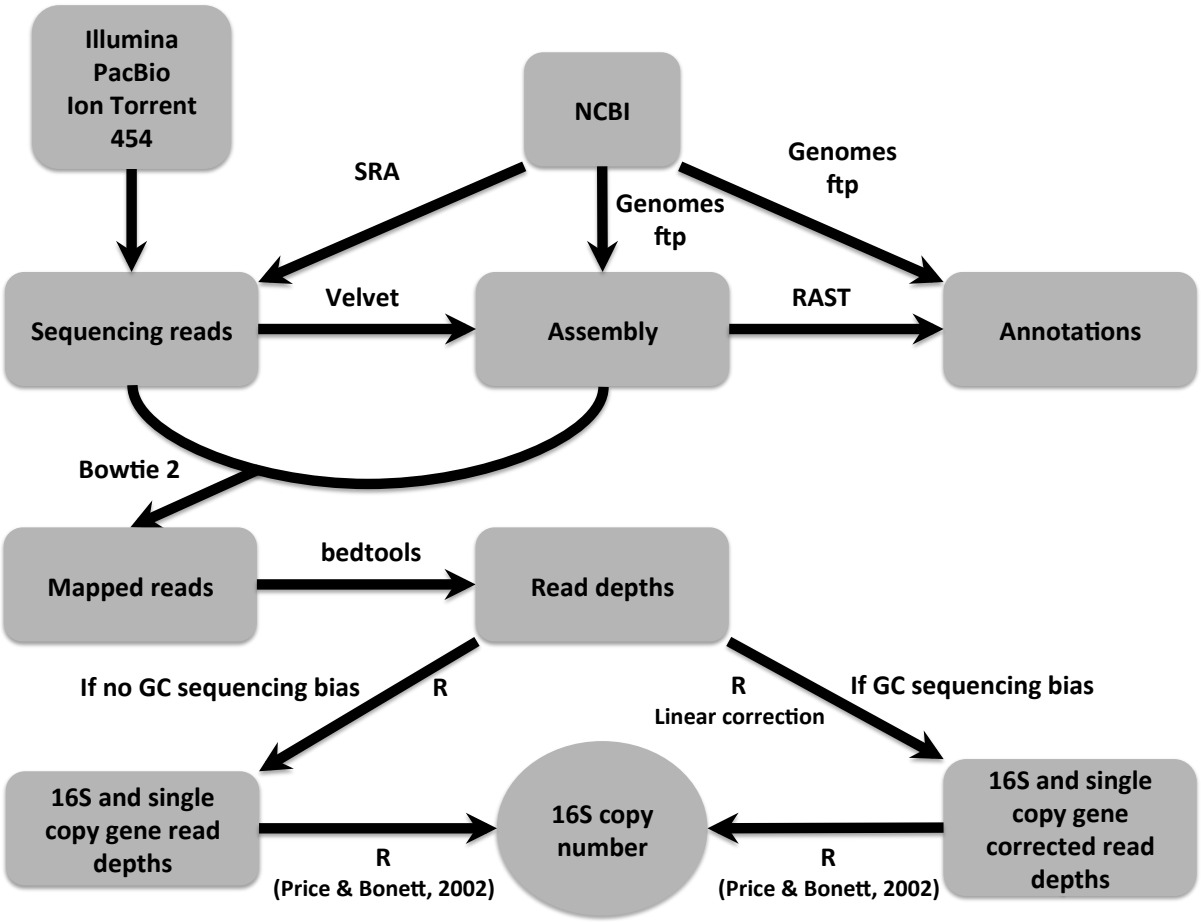


Figure 2.2: 16Stimator pipeline outline.

This pipeline requires three inputs: sequencing reads, assembly fasta, and annotations. These inputs can be downloaded from NCBI Short Read Archive (SRA) and Genomes ftp sites. For analysis of *de novo* assemblies, reads are generated on various sequencing platforms, assembled into contiguous sequence fragments (contigs), here using the Velvet assembler (Zerbino & Birney, 2008), and annotated, here using RAST (Aziz *et al.*, 2008; Overbeek *et al.*, 2013). Sequencing reads are then aligned to the assembly with Bowtie 2 (Langmead & Salzberg, 2012), and read depth at each genomic position is calculated with bedtools (Quinlan & Hall, 2010). If a significant GC% effect is detected, read depths are corrected using a linear model. Custom R scripts are used to extract read depths for 16S and single-copy genes and to estimate 16S copy number with Price & Bonett (2002) confidence intervals. Custom shell and R scripts can be found at https://bitbucket.org/perisin/16stimator.

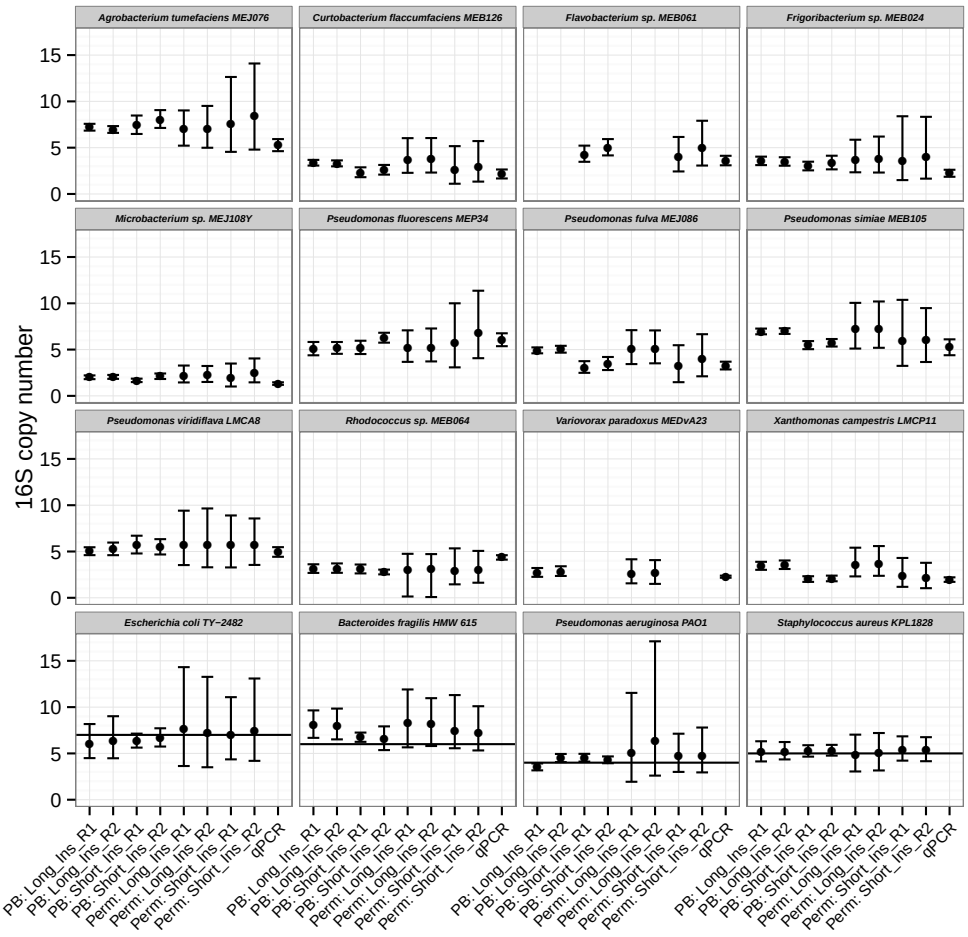


Figure 2.3: 16S copy number estimates from *de novo* assemblies.

For each endophytic isolate, paired-end sequencing reads (R1, R2) were generated on the Illumina HiSeq 2000 from short (∼250 bp) and long (∼2500) insert libraries (Short_Ins and Long_Ins, respectively). For closed genome controls, similarly generated sequencing reads were downloaded from SRA: *Escherichia coli* TY-2482 (GCA_000217695.2, SRR292678, SRR292862), *Bacteroides fragilis* HMW 615 (GCA_000297735.1, SRR488169, SRR488170), *Pseudomonas aeruginosa* PAO1 (GCA_000006765.1, SRR032420, SRR032832), and *Staphylococcus aureus* KPL1828 (GCA_000507725.1, SRR835799, SRR958927). The 16Stimator pipeline was used to estimate 16S copy number as the ratio of median coverage for 16S and single-copy genes. Confidence intervals (95%) were either calculated as in Price & Bonett (2002) (PB), or via permutations (Perm). For endophytic isolates, 16S copy numbers were independently verified by absolute quantification via qPCR with the mean and standard deviation of technical replicates shown. For closed genome controls, each horizontal line marks the *rrn*DB (Stoddard *et al.*, 2014) consensus 16S copy number for each species. Note: the short insert library for MEDvA23 and the long insert library for MEB061 did not meet quality thresholds. 16S copy number was not experimentally determined by qPCR for *E. coli* TY-2482, *B. fragilis* HMW 615, *P. aeruginosa* PAO1, and *S. aureus* KPL1828.

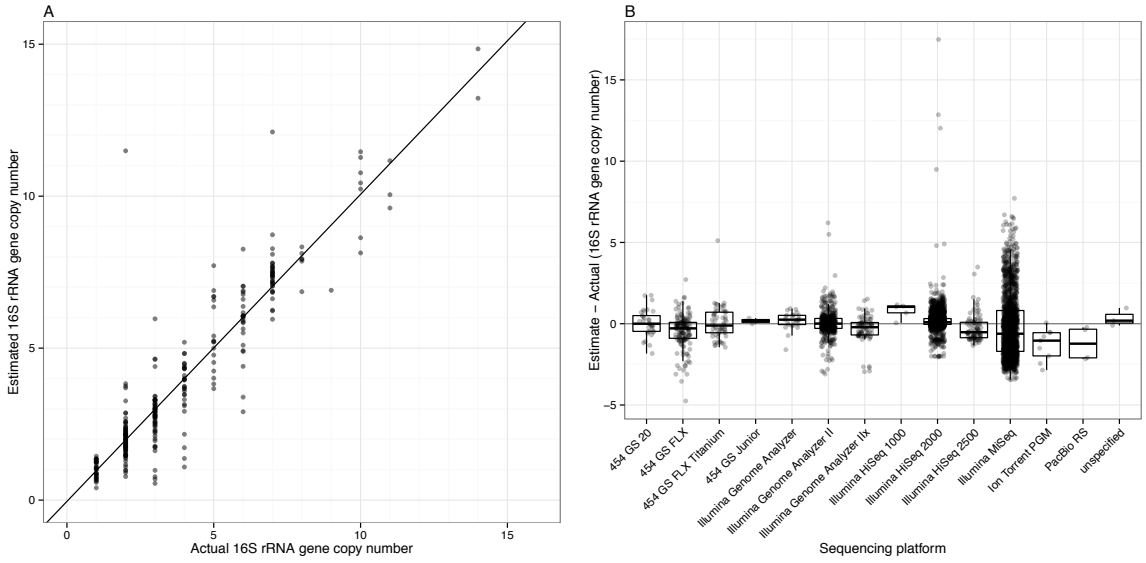


Figure 2.4: 16Stimator pipeline accuracy and biases.

(A) Accuracy of 16Stimator pipeline. 16Stimator was applied to sequencing libraries generated from isolates with closed genomes. Plotting estimate versus actual 16S copy numbers assessed accuracy. Each point corresponds to the point estimate for 16S copy number generated from a separate sequencing library. As the number of sequencing libraries per isolate varies, we randomly chose one library for each Isolate:Bioproject combination (283 total) and then fit a linear model (lm function in R). This process was repeated 1000 times to obtain 95% confidence intervals for the mean intercept (-0.05 [-0.13, 0.04] p = 0.65 [0.24, 0.98]), slope (1.01 [0.98, 1.03] $p <$ 2e-16), and adjusted R^2 values (0.82 [0.75, 0.87]). The above plot represents one such iteration. (B) Bias of 16Stimator by sequencing technology. Each point corresponds to results from different sequencing libraries mapped back to their respective closed genomes. Actual 16S copy number values were subtracted from estimates and grouped by sequencing platform. (454 GS 20, n = 33; 454 GS FLX, n = 145; 454 GS FLX Titanium, n = 57; 454 GS Junior, n = 2; Illumina Genome Analyzer, n = 26; Illumina Genome Analyzer II, n = 521; Illumina Genome Analyzer IIx, n = 78; Illumina HiSeq 1000, n = 5; Illumina HiSeq 2000, n = 2002; Illumina HiSeq 2500, n = 92; Illumina MiSeq, n = 1671; Ion Torrent PGM, n = 9; PacBio RS, n = 4; Unspecified, n = 3)

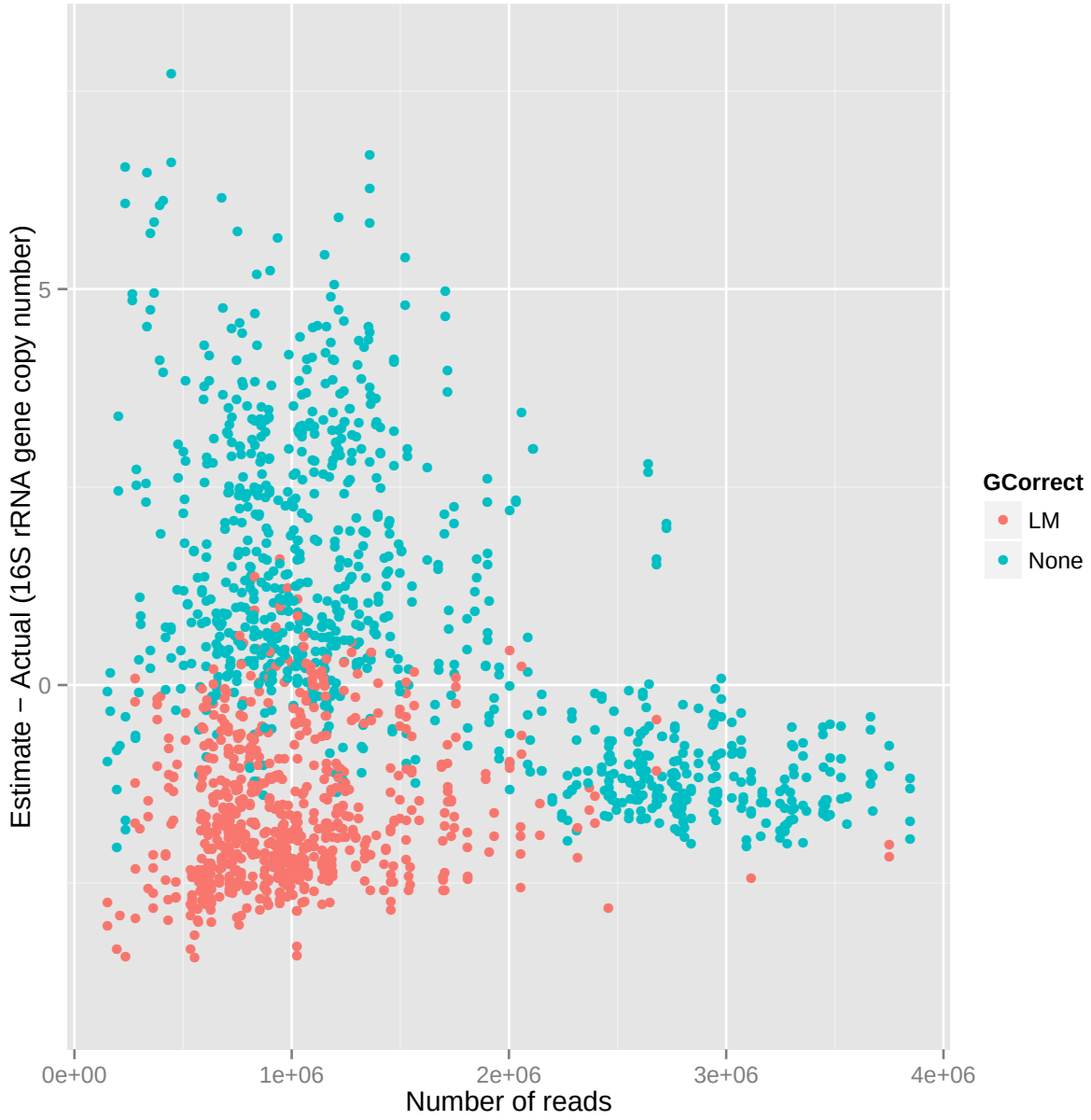


Figure 2.5: Interaction between sequencing depth and GC correction affects 16Stimator accuracy for *Listeria monocytogenes* genomes sequenced on the Illumina MiSeq.

Sequencing libraries generated on the MiSeq (n = 1616) were used to estimate 16S copy number for *L. monocytogenes* genomes. The sequencing read depth is plotted against the difference between 16Stimator estimate and actual copy number. Points are colored red if linear (LM) GC% correction was applied and blue if no GC% correction was applied to the read depths.

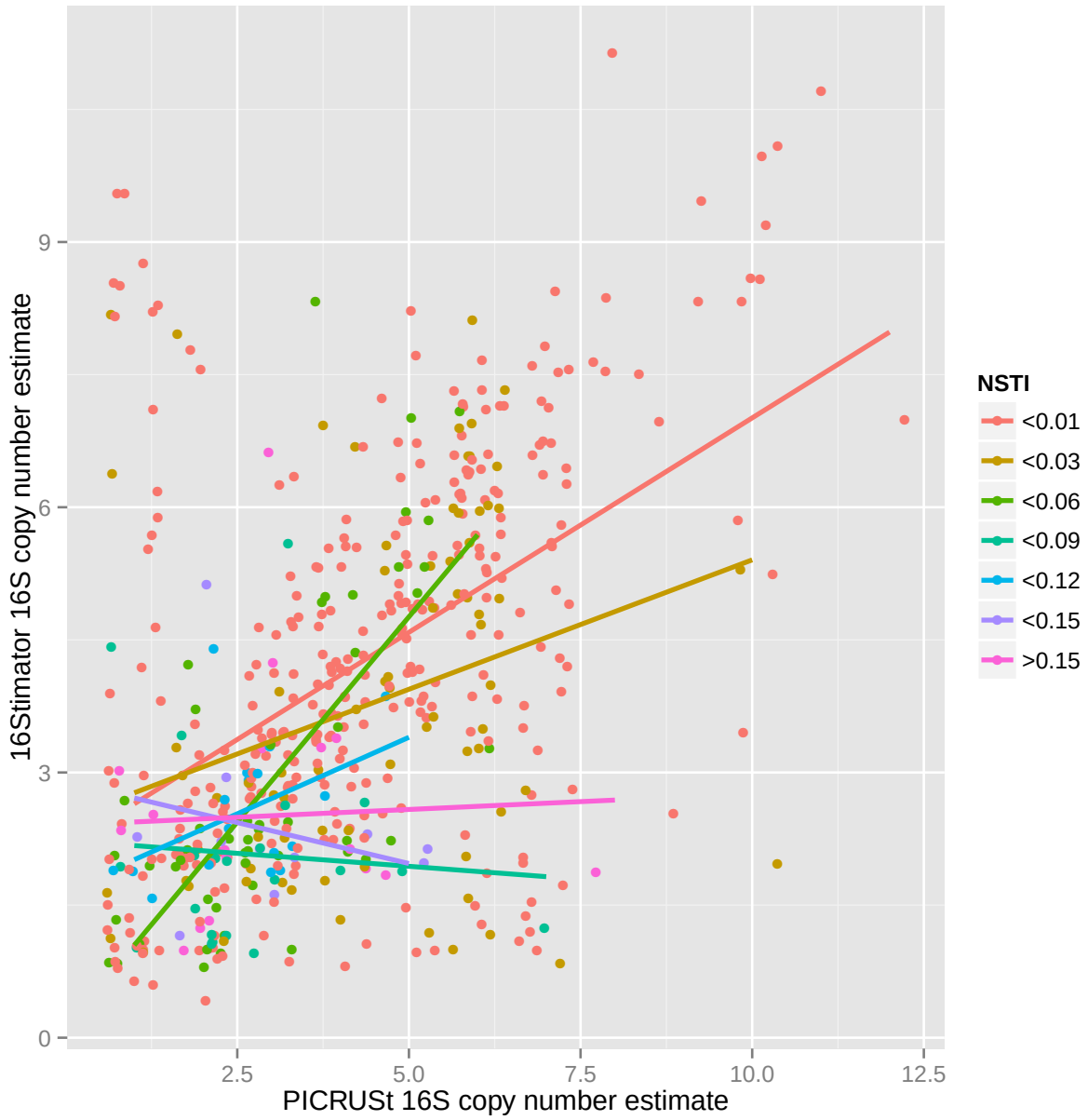


Figure 2.6: Correlations between 16Stimator and PICRUSt 16S copy number estimates deteriorate with increasing NSTI.

For genomes with 16Stimator estimates, 16S sequences were extracted and aligned to Greengenes 99% OTUs (McDonald *et al.*, 2012). For matching OTUs (535 total), the median 16Stimator estimates were plotted against PICRUSt estimates (Langille *et al.*, 2013), and colored according to their nearest sequenced taxon index (NSTI) cutoff (<0.01, n = 341; <0.03, n = 77; <0.06, n = 50; <0.09, n = 21; <0.12, n = 17; <0.15, n = 11; >0.15, n = 18). Lines represent linear relationships between estimates colored by NSTI, with Pearson correlations: (<0.01, PCC = 0.50, $p < 0.001$; <0.03, PCC = 0.28, $p = 0.014$; <0.06, PCC = 0.71, p <0.001; <0.09, PCC = -0.07, $p = 0.76$; <0.12, PCC = 0.47, $p = 0.056$; <0.15, PCC = -0.24, $p = 0.48$; >0.15, PCC = 0.05, $p = 0.86$).

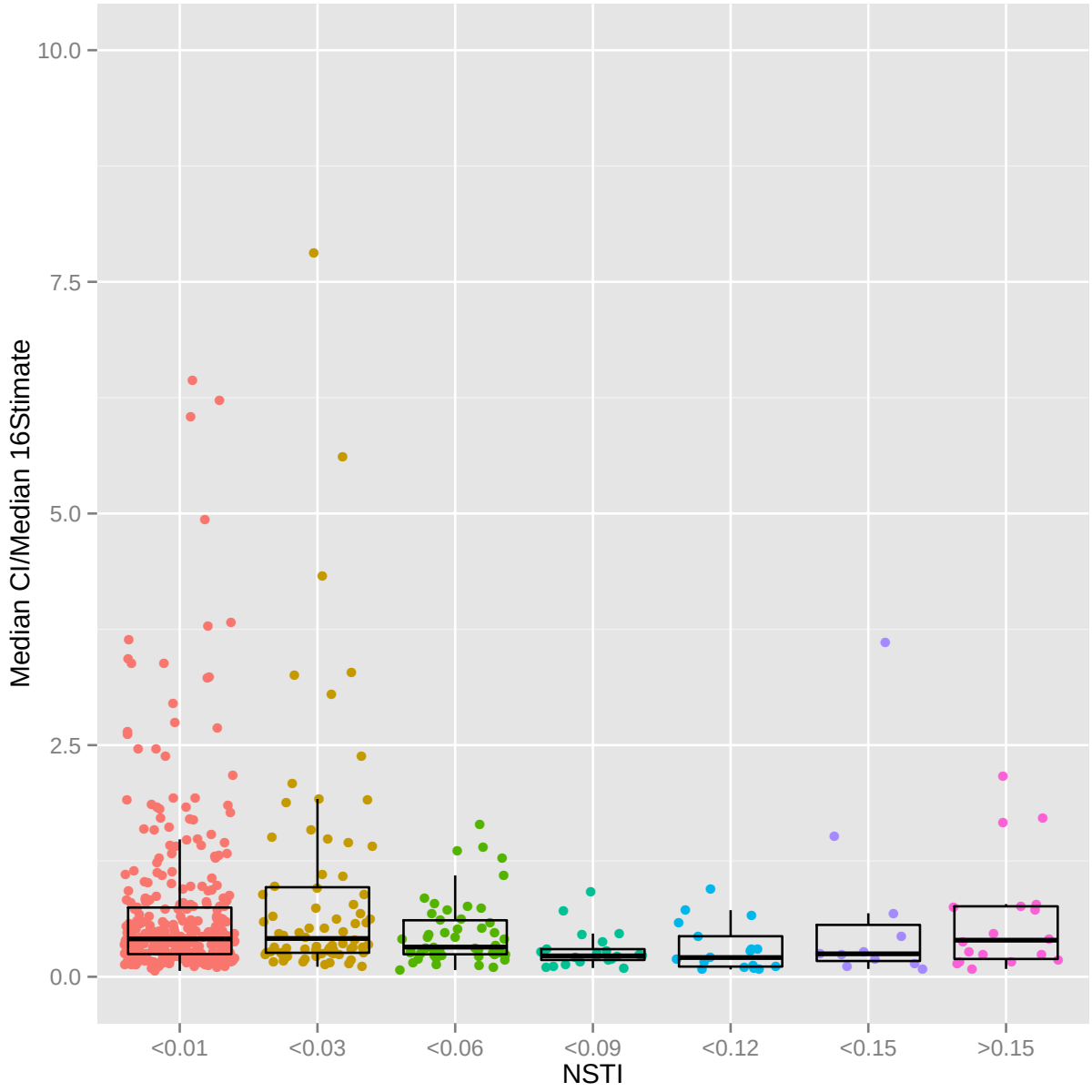


Figure 2.7: 16Stimator confidence interval size is unaffected by NSTI.

We calculated 16Stimator copy number estimates for the Greengenes 99% OTUs and plotted the proportion of median confidence interval (CI) to median copy number estimate over increasing NSTI cutoffs. The proportion of CI to median estimate allows for comparison of estimates with different copy numbers, as the CI size alone will necessarily increase with higher copy numbers. To better visualize and compare the boxplots, outlier points greater than 10 are not shown.

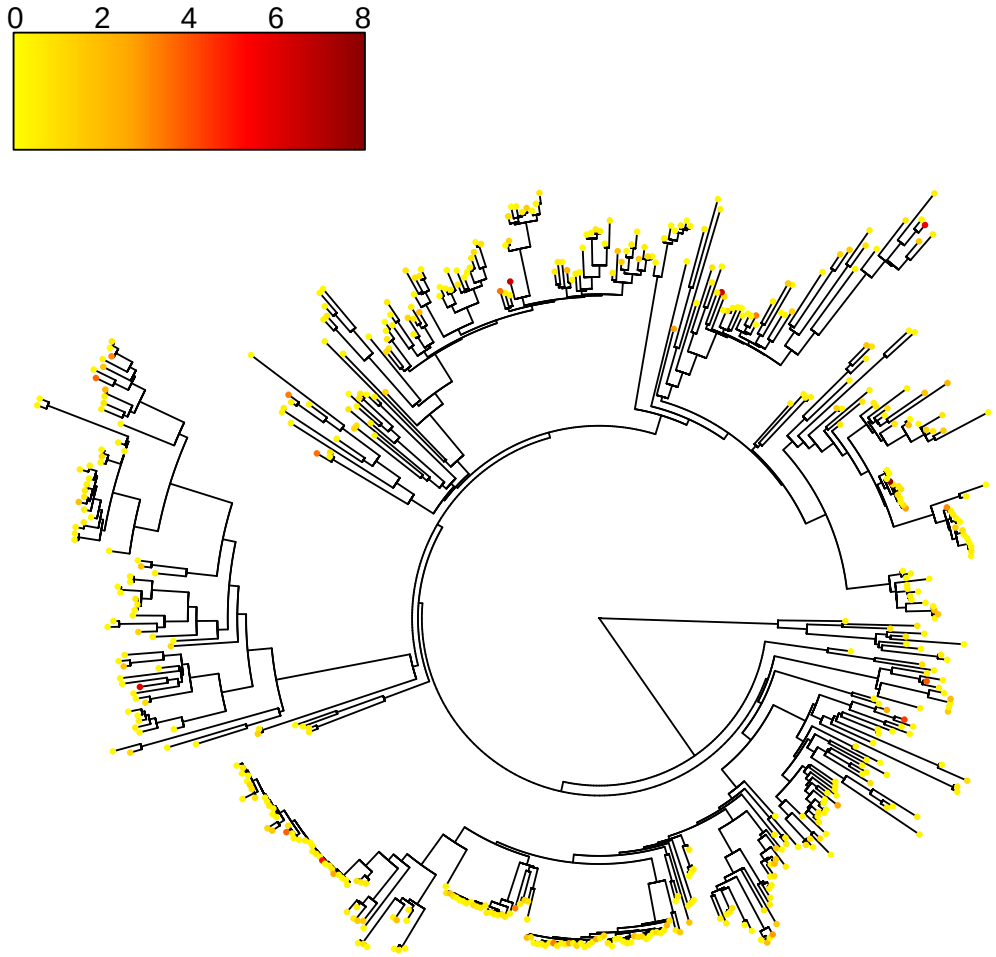


Figure 2.8: Bacterial phylogeny does not impact 16Stimator confidence intervals.

For the Greengenes 99% OTUs, the proportion of median confidence interval to median 16Stimator copy number estimate is shown as a heatmap plotted on the Greengenes phylogeny tree.

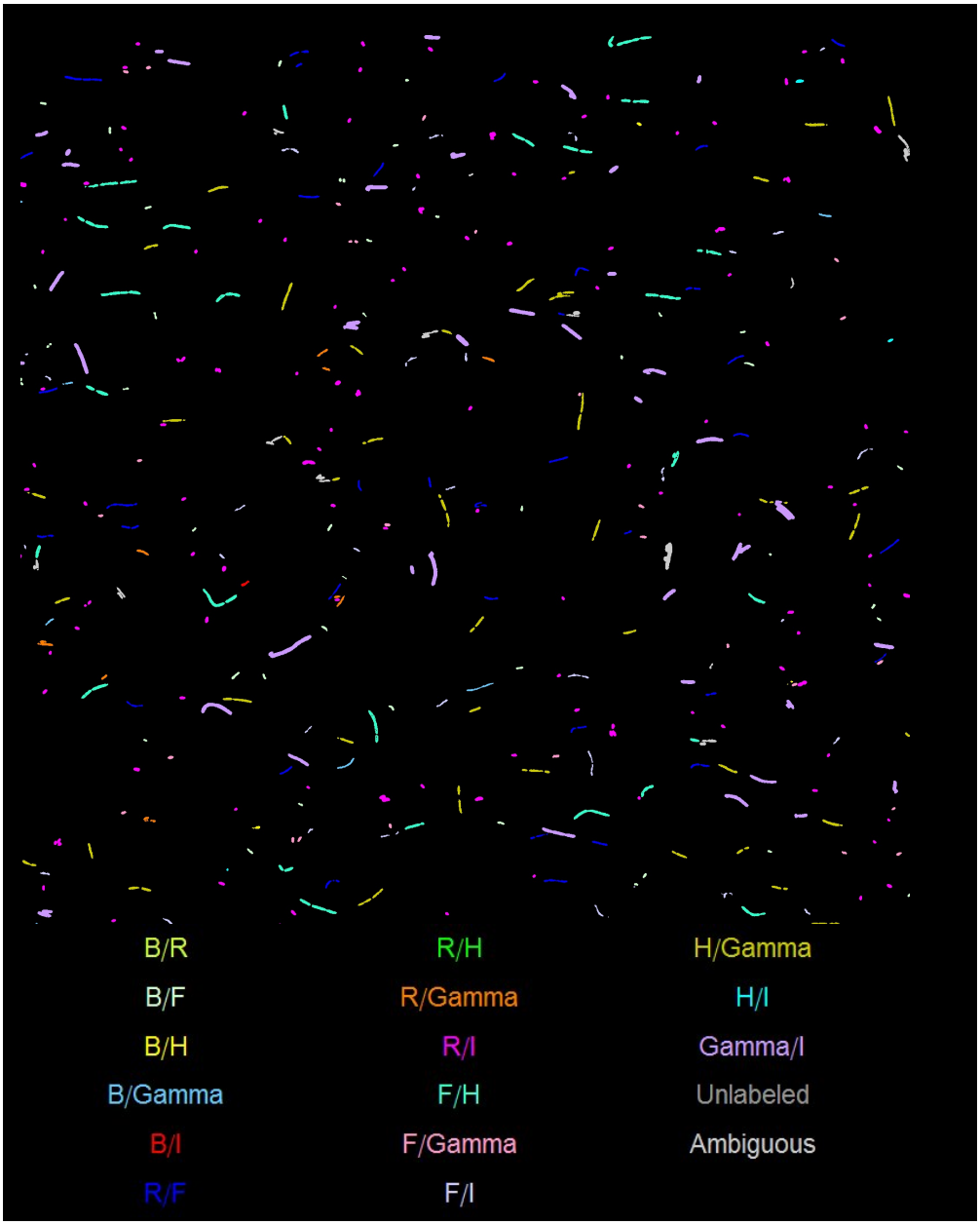
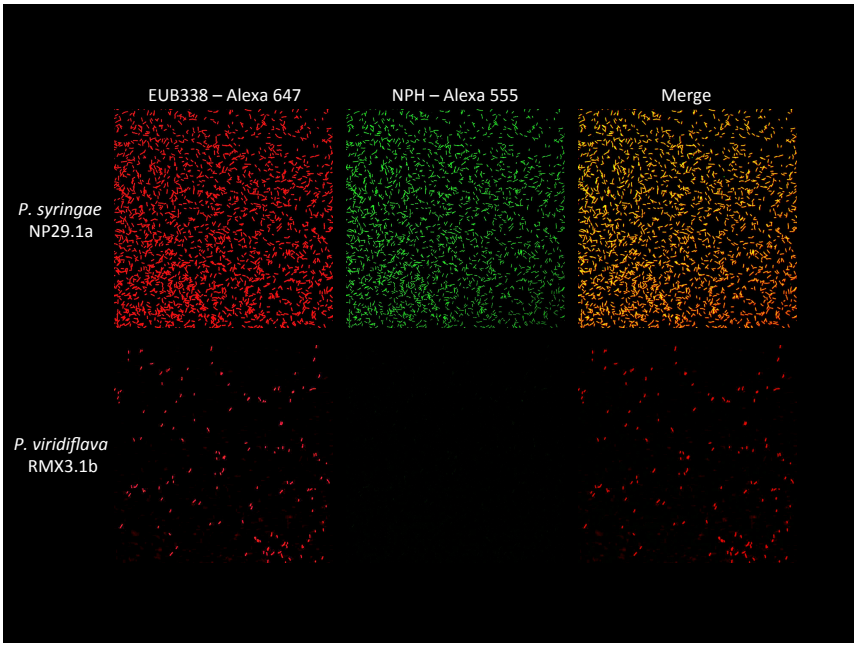
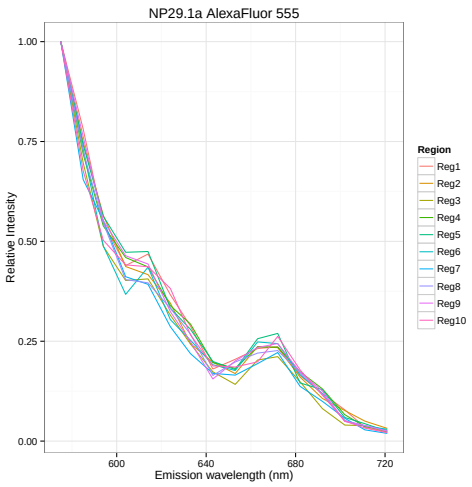


Figure 2.9: CLASI-FISH of endophytic bacterial isolates.

Cultures for each endophytic isolate were separately fixed and hybridized with different combinations of fluorophore conjugated EUB338 oligonucleotide probes. Hybridized cells were then mixed and spectral imaged. Fluorophores: B = Alexa Fluor 488, R = Alexa Fluor 514, H = Alexa Fluor 555, F = Rhodamine Red-X, I = Alexa Fluor 647, Gamma = Texas Red-X. Isolates: F/Gamma = *X. campestris* LMCP11, R/Gamma = *P. fluorescens* MEP34, H/I = *V. paradoxus* MEDvA23, B/I = *Frigoribacterium sp.* MEB024, B/Gamma = *Flavobacterium sp.* MEB061, R/H = *Rhodococcus sp.* MEB064, R/I = *P. simiae* MEB105, B/F = *C. flaccumfaciens* MEB126, B/H = *A. tumefaciens* MEJ076, R/F = *P. fulva* MEJ086, F/I = *Microbacterium sp.* MEJ108Y, H/F = *P. syringae* NP29.1a, H/Gamma = *P. viridiflava* RMX3.1b, Gamma/I = *Shewanella sp.* as a control for hybridization and imaging (Valm *et al.*, 2012).



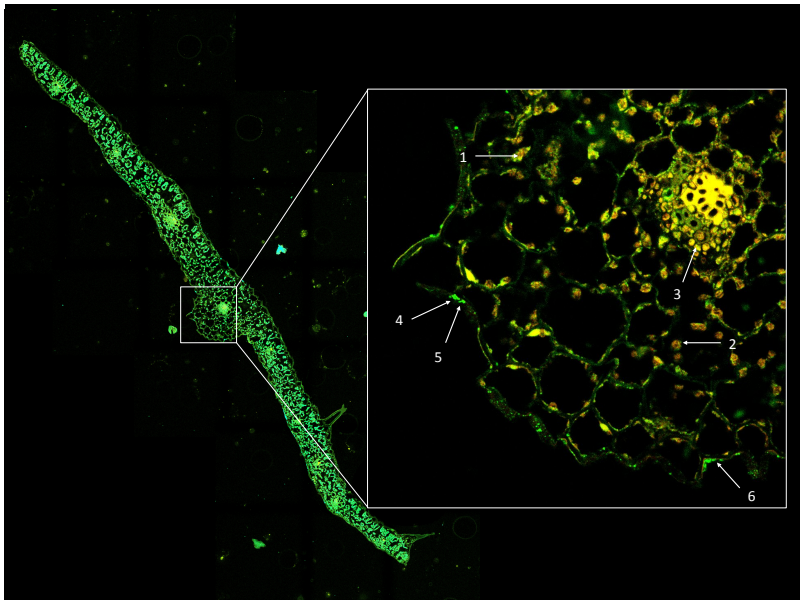
(a)



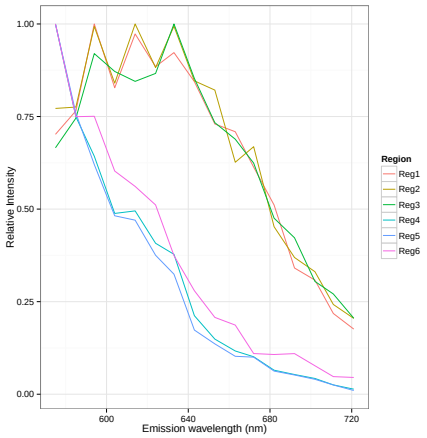
(b)

Figure 2.10: *P. syringae* specific FISH probe confirmation.

Fixed *P. syringae* NP29.1a and closely related *P. viridiflava* RMX3.1b cells were hybridized with the EUB338 probe conjugated to Alexa Fluor 647 and NP-H probe conjugated to Alexa Fluor 555. (a) For each isolate, panels show either the composite spectral image using the 633 nm laser for excitation of Alexa Fluor 647 and, the composite spectral image using the 561 nm laser for excitation of Alexa Fluor 555, or the merge of the two composite spectral images. (b) Ten regions from ten random cells were chosen from the *P. syringae* NP29.1a with NP-H panel and linear unmixing was used to extract the emission spectrums.



(a)



(b)

Figure 2.11: *In planta* FISH and spectral imaging of *P. syringae*

Gnotobiotic *A. thaliana* were inoculated with *P. syringae* NP29.1a and plants 15 days post inoculation were harvested, fixed, and sectioned. Leaf cross-sections were used in FISH with the EUB338 probe conjugated to Alexa Fluor 488 and NP-H probe conjugated to Alexa Fluor 555. (a) The entire leaf cross-section (perpendicular to the midvein) was imaged with a tile scan using the 20x objective with and 488 nm laser to take spectral images. The composite image showing plant autofluorescence and probe fluorescence is shown. The inset box shows the 40x image of the area around the midvein. The spectral image was taken using the 561 nm laser to excite Alexa Fluor 555 and plant autofluorescence. Arrows 1-3 point to areas of autofluorescence. Arrows 4-6 point to *P. syringae* cells. (b) Linear unmixing was used to extract emission spectra for the regions indicated by the arrows in (a), confirming spectral differences between bacterial cells labeled with Alexa Fluor 555 and plant autofluorescence.

CHAPTER 3

ENVIRONMENTAL COMPLEXITY AND SPECIES RICHNESS ALTER BACTERIAL COMMUNITY COMPOSITION AND PREDICTABILITY UNDER CONTROLLED CONDITIONS

3.1 Abstract

The recent explosion of microbiome studies has uncovered important correlations of bacterial community structures with temporal and environmental gradients; however, our ability to predict species' abundances and community succession trajectories under such complex conditions remains poor. In this study, we characterized bacterial succession under controlled conditions and across two axes of complexity, environmental heterogeneity and species richness. We took advantage of cultivatable bacterial species isolated from the leaf endophytic environment of *Arabidopsis thaliana*. We inoculated random subsets of twelve species into three increasingly complex environments (liquid media, solid media, gnotobiotic *A. thaliana*) and quantified community composition by 16S rRNA gene amplicon sequencing. We observed strong environmental and species richness effects on resultant community structures. By comparison of replicate community structures, we observed the highest divergence, and therefore, least predictability, in plant communities followed by communities in liquid and then solid media. Species abundances also showed hierarchical patterns, and these hierarchies differed by environment. Interestingly, *Pseudomonas viridiflava* and *Curtobacterium flaccumfaciens*, which have different effects on plant health, were barely detected in media but were highly abundant *in planta*.

3.2 Introduction

The diversity of plant-associated microbial communities is staggering. The phyllosphere is estimated to host 10^{26} total bacterial cells (Lindow & Brandl, 2003), and within a single

plant, hundreds of species (Bodenhausen *et al.*, 2013). These levels of diversity and coexistence are perplexing given the principle of competitive exclusion (Gause, 1934). How species interact and organize to build a productive community has been a fundamental, ecological question with niche (Hutchinson, 1961) and neutral (Hubbell, 2001) theories proposed to explain these assembly processes.

Greater understanding of microbial coexistence is of particular interest to predict, and possibly control, microbial communities for promotion of host health, crop yield augmentation, increased bioreactor function, bioremediation, and environmental forecasting. Much work has gone into predicting community compositions based on environmental factors, with clear successional and seasonal patterns observed (Fuhrman *et al.*, 2006; Redford & Fierer, 2009; Gilbert *et al.*, 2011; Koenig *et al.*, 2011; Eiler *et al.*, 2012). The type of environment clearly influences predictability. As shown in a meta analysis of temporal patterns of bacterial marker gene surveys, soil and brewery wastewater treatment water were least variable and stream, infant, gut, flower, and human palm communities were the most variable (Shade *et al.*, 2013). In less than half of the studies, variability correlated with time, indicating a chance to predict communities (Shade *et al.*, 2013). However, the ability to infer community structure from environmental factors, implies that the influential factors to measure are known. A powerful approach would be to predict future community states based on the current composition. This is no easy task, but has a large upside as Gilbert *et al.* (2011) noted that correlations between bacterial operational taxonomic units (OTUs) were stronger than between bacteria and environmental factors.

Further evidence for diversity-predictability relationships was shown in Shade *et al.* (2013), in which alpha-diversity metrics correlated well with beta-diversity metrics in most studies. These patterns have also been observed in human surveys in which temporal variability of bacterial communities was shown to be a personalized trait (Flores *et al.*, 2014). More diverse tongue and gut communities, as evidenced by higher Shannon entropy, were associated with less temporal variability, as evidenced by decreased beta-diversity (Flores

et al., 2014).

Studies of natural communities, while directly relevant, are hard to replicate, hence, the need for studies of diversity and predictability under controlled conditions. Bioreactors offer convenient and industrially important environments to inoculate replicate communities and observe changes in community structure and productivity over time. Both predictable (Falk *et al.*, 2009; Vanwonterghem *et al.*, 2014) and unpredictable (Zhou *et al.*, 2013; Pagaling *et al.*, 2014) community structures have been observed when replicate bioreactors have been inoculated with the same mix of microbes. Similarity of environments from which the inoculum was taken to the new environment may influence the variability in resultant structures as Pagaling *et al.* (2014) observed that preconditioned communities inoculated into the same sterile environment were highly similar. The desired function of the bioreactor in addition to the functional redundancy in the inoculated community will also influence predictability. Interestingly, given the composition of the initial inoculum, one cannot predict the communities that will form. In some cases, rare to undetectable species in the inoculum can become the most abundant after introduction to the new environment (Pagaling *et al.*, 2014; Vanwonterghem *et al.*, 2014). As function is generally selected in bioreactors, variable structures may not be problematic if they accomplish the desired task.

To examine the variability in structures that arise from a common inoculum, more controlled experiments are needed in which the initial structure is completely controlled. Recently, Celiker & Gore (2014) have investigated how community structure evolves by inoculating 96 wells with the same six bacterial isolates and then serially passaging the replicates for 400 generations. Short-term community structures were highly conserved, but after experimental evolution, the communities diverged into four types of structures, demonstrating the unpredictability of communities even under highly controlled conditions (Celiker & Gore, 2014).

Previous work on bacterial community assembly with replicate greenhouse plants, described convergent epiphytic community structures in terms of membership but high vari-

ability in terms of relative abundance (Maignien *et al.*, 2014). Replicable communities have been observed using gnotobiotic *A. thaliana* sprayed with a synthetic community of seven bacterial isolates, though these isolates did not include common pathogens (Bodenhausen *et al.*, 2014).

Here we build upon these studies by first examining the repeatability, and hence potential for predictability, of bacterial community succession under controlled conditions. We observed community succession across two axes of complexity, species richness and environmental heterogeneity. The environments ranged from constantly shaken, homogenous liquid media, to the same media in solid form, to gnotobiotic plants. Richness ranged from two to eleven species. We took advantage of bacterial cultures originally isolated from the leaf endophytic niche of *A. thaliana* located in the Midwestern United States. These isolates are among the most abundant genera as determined by culture-dependent (Kniskern *et al.*, 2007; Traw *et al.*, 2007; Barrett *et al.*, 2011) and independent (Delmotte *et al.*, 2009; Bodenhausen *et al.*, 2013; Horton *et al.*, 2014) means. Included in these assays are species thought to be pathogenic (Buell, 2002; Jakob *et al.*, 2002), beneficial (Haas & Défago, 2005), and unknown (Table 3.1).

To characterize the bacterial community structure at each time point post inoculation, we destructively sampled communities, extracted DNA, amplified variable regions of the 16S rRNA gene, and sequenced the amplicons. From this compositional data, we observed that environment and initial species richness affected community structure. Based on comparison of replicates, community repeatability did not necessarily decrease with environmental complexity but did decrease with increasing number of species. However, by averaging across replicates, we observed hierarchical patterns of species' abundances, with a different hierarchy order for each environment. Interestingly, two species that were lowly abundant in liquid or solid media, rose to high abundance *in planta*, and had different relationships with plant fresh mass.

3.3 Methods

3.3.1 Bacterial strains

Endophytic bacteria were isolated from leaves of *Arabidopsis thaliana* or *Draba verna* located in the Midwestern United States (Table 3.1). Isolates were cultured and maintained in King's B (KB) media (King *et al.*, 1954). These isolates have had their genomes sequenced, assembled, and deposited in NCBI (Table 3.1), and 16S rRNA gene copy numbers determined by sequencing read depth and qPCR approaches (Perisin *et al.*, 2015).

3.3.2 Gnotobiotic plant growth conditions

All seeds used in this study were from *Arabidopsis thaliana* accession PT1.85 (ID 8057), originally collected from Hanna, IN USA (41.3423, -86.7368). Seeds were surface sterilized by addition of 70% ethanol, incubation at room temperature for 1 minute, removal of ethanol, addition of 100% ethanol, incubation at room temperature for 1 minute, and removal of ethanol. Seeds were air dried in a sterilized biological cabinet. Plants were grown by placing single seeds into each well of 24 well plates (Costar 3738, Corning, Tewksbury, MA, USA) containing 1 mL of sterile Murashige and Skoog (MS) media (4.4 g MS (Duchefa, Haarlem, Netherlands), 10 g sucrose, 0.5 g MES, 6 g plant tissue culture agar (plantmedia, Dublin, OH, USA), pH 5.7-5.8, per 1 L). Seeds were stratified for 3 days at 4°C, and grown in long day conditions (16 h light at 21°C, 8 h dark at 18°C) with 50% relative humidity in a growth chamber (SANYO, MLR-351). Plants were grown for 14 days prior to inoculation.

3.3.3 Community inoculation

Community compositions were randomly chosen from the pool of twelve isolates, with five combinations at each species richness level (Table 3.2). Negative blank buffer and positive single isolate controls were also included. Experiments were carried out over the course of six temporal blocks. One replicate of each combination (25 total) along with controls were

included in each block. All inoculations were completed in 24 well plates with either liquid media, solid media, or gnotobiotic plants. For each block and environment, the inoculation order of 25 combinations and controls was randomized among wells in two plates per time point (8 plates total per block:environment).

For each temporal block, isolates were streaked from frozen stocks onto KB agar plates, and grown at 28°C until colonies appeared (24-48 h). Single colonies were inoculated into liquid KB and grown overnight at 28°C with 175 RPM shaking. Cells were pelleted by centrifugation at 9000 RPM for 2 min, washed once with 10 mM $MgSO_4$, and resuspended in 10 mM $MgSO_4$ to $OD_{600} = 0.1$ ($\approx 10^8$ CFU/mL). All further steps were conducted within a sterile biological cabinet.

For liquid and solid media inoculations, each isolate was diluted to 10^7 CFU/mL in Super Optimal Broth with glucose (SOC, 5 g yeast extract, 20 g tryptone, 0.5 g NaCl, 0.186 g KCl, 10 mL 1 M $MgCl_2$, 10 mL 1 M $MgSO_4$, 20 mL 1 M glucose, 7.5 g agar for solid, per 1 L). Isolates were then combined in equal proportions for a final total concentration of 10^5 CFU/mL. For liquid inoculations, 1 mL of inoculum was added to each well. Plates were sealed with parafilm and incubated at 28°C with 175 RPM shaking. For solid media, 1 mL of inoculum was added to 1 mL of solid SOC media in each well. After room temperature incubation for 2 minutes, the inoculum was removed, and plates were allowed to air dry. Plates were sealed with parafilm and incubated at 28°C.

To measure single isolate growth in SOC broth, isolates were prepared and diluted as above, except the final concentration was 10^5 CFU/mL. Seven replicates of each isolate (200 μL per well) were added to a clear 96 well plate (Costar, 3370, Corning, Tewksbury, MA, USA) in a randomized pattern across the plate. The plate was incubated in an Infinite F200 plate reader (Tecan, Morrisville, NC, USA) at 28°C with 182 rpm orbital shaking. The absorbance at 600 nm was measured every 10 min for 7 days.

For plant inoculations, each isolate was diluted to 10^7 CFU/mL in 10 mM $MgSO_4$. Isolates were then combined in equal proportions for a final total concentration of 10^4 CFU/mL.

Each plant was flood inoculated by adding 1 mL of inoculum to each well. After room temperature incubation for 2 minutes, the inoculum was removed, and plants were allowed to air dry. Plates were sealed with parafilm and placed back in the growth chamber. After all inoculations, inoculums were diluted and spread on KB plates to ensure bacterial viability and correct dosage.

3.3.4 Sample collection

At each time point (2, 4, 6, 8 days) post inoculation, liquid media, solid media, and plant samples were all harvested in a random order and added to Nunc deepwell 96 well plates (Thermo Scientific, Waltham, MA, USA). For liquid samples, all liquid was removed from the well and added to the appropriate harvest plate well. For solid media samples, 1 mL of 10 mM $MgSO_4$ was added to each well, and bacteria were removed from the surface by scraping with a flame sterilized spreader. Cells were resuspended in the buffer and added to the appropriate harvest plate well. All bacteria were then pelleted by centrifugation at 6600xg for 5 minutes, and supernatants were removed.

For plant samples, rosettes were removed from roots with flame sterilized forceps, and fresh mass was recorded. Rosettes were washed by dipping ten times in 70% ethanol, TE (10 mM Tris-HCl, 1 mM EDTA, pH 7.5), 70% ethanol, and TE. Plants were placed in appropriate wells of the harvest plate. Harvest plates were immediately placed and kept at -80°C until DNA extraction.

3.3.5 DNA extraction

Genomic DNA was extracted using a modified protocol from Morgan *et al.* (2010), that was confirmed to successfully extract DNA from each of the twelve bacterial isolates used in this study. All centrifugation steps were completed using Beckman Coulter Avanti J-25 centrifuge (Beckman Instruments, Munich, Germany) at 6600xg at 4°C. Harvest plates were thawed in a 37°C water bath for 10 minutes. To each well, 200 μL of TES (10 mM Tris-HCl,

1 mM EDTA, 100 mM NaCl, pH 7.5) supplemented with 50 U/μL Ready-lyse Lysozyme (Epicentre, Madison, WI, USA) was added along with two 2.3 mm stainless steel beads (BioSpec, Bartlesville, OK, USA). Samples were mechanically disrupted at 1750 RPM for 5 minutes using a 2010 Genogrinder (SPEX, Metuchen, NJ, USA), and incubated overnight at room temperature. The following day, 400 μL of Cell Lysis Solution (Gentra Puregene Yeast/Bacteria Kit, Qiagen, Valencia, CA, USA) was added to each well, and the plate was incubated at 80°C for 5 minutes. After cooling to room temperature, plates were flash centrifuged, and Protein Precipitation Solution (Gentra Puregene Yeast/Bacteria Kit, Qiagen) was added to each well (200 μL). Plates were vortexed vigorously for 30 s, and incubated at 4°C for 3 h. Proteins were pelleted by centrifugation for 15 minutes. All following pipetting and shaking steps were completed using custom liquid handling robotics scripts on the Freedom Evo 200 (Tecan, Morrisville, NC, USA). DNA was precipitated by pipetting 450 μL of supernatant and 450 μL of isopropanol into a new Nunc deepwell plate. Plates were inverted 50 times to mix before centrifugation for 15 minutes. Isopropanol was removed, and 450 μL of 70% ethanol was added to wash the DNA pellets. Plates were centrifuged for 2 minutes and ethanol was removed. DNA pellets were air dried in a chemical hood and 200 μL of TE was added to each well. Pellets were resuspended by shaking for 4 minutes. After incubation on ice for 5 minutes, plates were centrifuged for 15 minutes to pellet impurities. Supernatants were pipetted into new Nunc 0.5 mL 96 well plates and kept at -20°C.

3.3.6 16S rRNA gene amplification

The V3-V4 region of the 16S rRNA gene was amplified from each sample using custom primers designed to be flexible enough to amplify the twelve isolates used in this study while excluding amplification of plant chloroplast and mitochondrial sequences (MP574F 5'-ACA ATG GGC GMA AGC CTG AT-3'; MP984R 5'-ACC AGG GTA TCT AAK CCT G-3'). Primers were designed as in Caporaso *et al.* (2012) and also contained Illumina MiSeq

adapters, custom pads and linkers, and barcode sequences (Appendix A). All following pipetting steps were completed using custom scripts on the Freedom Evo 200 (Tecan, Morrisville, NC, USA). Each PCR was completed in triplicate. The total PCR volume was 25 μL, which contained 2.5 μL of template, 0.2 μM of each primer and 1X 5PRIME HotMasterMix (5PRIME, Gaithersburg, MD, USA). PCR amplification consisted of an initial denaturation step of 2 minutes at 94°C, followed by 35 cycles of denaturation at 94°C for 45 s, annealing at 65°C for 30 s, elongation at 72°C for 45 s, followed by a final elongation of 5 minutes at 72°C.

Replicate reactions were pooled and amplicons purified with 1.2X Sera-Mag Speedbeads (Thermo Scientific, 0.1% beads, 18% PEG-8000, 1 M NaCl, 10 mM Tris-HCl, 1 mM EDTA, pH 8). Purification steps were the same as the standard protocol for Agencourt AMPure XP PCR purification (Beckman Coulter). Amplicon concentrations were quantified by fluorimetry (QUANT-iT PicoGreen dsDNA Assay Kit, Life Technologies, Carlsbad, CA, USA), and 30 ng or a maximum of 30 μL per sample were pooled for two blocks per sequencing run. The length distribution and purity of the final pools were visualized with agarose gels. The final DNA concentrations were determined, and the amplicon pools were sequenced using the Illumina MiSeq platform and MiSeq V2 Reagent Kits (Illumina, San Diego, CA, USA). Runs consisted of paired-end 250 bp reads (MiSeq Control Software v2.2.0.2). To control for variations between sequencing runs, amplicons generated separately from each bacterial isolate were included in each run.

3.3.7 Sequence processing

Raw sequencing reads that aligned to PhiX using Bowtie 2 v2.1.0 (Langmead & Salzberg, 2012) were removed. Paired-end reads were merged using USEARCH v6.1.544 (Edgar, 2010) with fastq_truncqual 10, fastq_minmergelen 360, and fastq_maxmergelen 400. Merged reads were demultiplexed using QIIME v1.7.0 (Caporaso *et al.*, 2010b) split_libraries_fastq.py with q20 and default settings. Demultiplexed sequences were aligned against the known isolate

16S rRNA gene sequences using pick_closed_reference_otus.py with UCLUST v1.2.22 (Edgar, 2010) percentage alignment set at 99%. To account for differences in sequencing depth, reads for each sample were rarefied to 1000 sequences. Rarefaction was completed 1000 times, and the averages for isolate abundances were determined and used for all downstream analysis except where indicated differently.

3.3.8 Statistical analysis

The R statistical environment (Team, 2015) was used for all analyses along with the ggplot2 package for plotting (Wickham, 2009).The vegan package was used for multivariate dissimilarity calculations and tests (Oksanen *et al.*, 2013). To test for treatment effects on community structure, permutational multivariate analysis of variance (PERMANOVA, *adonis* function) was used to partition Bray-Curtis distance matrices (*vegdist* function) by treatments. Significance was assessed by 999 permutations.

The following model formula was used for PERMANOVA:

Bray-Curtis distances $\sim$ Environment + Richness + Time + Block + Environment:Richness + Inoculum/Richness + Environment:(Inoculum/Richness)

The *pco* function in the labdsv package was used for principal coordinates analysis (PCoA) (https://cran.r-project.org/web/packages/labdsv/index.html) and the *chull* function in grDevices was used to calculate convex hulls for PCoA plots. To test for treatment effects on repeatability, only pairwise dissimilarities between replicates were calculated. ANOVA (*aov* function) was used with the following model formula:

sqrt(Bray-Curtis distances) $\sim$ Environment + Richness + Time + Environment:Richness + Inoculum/Richness + Environment:(Inoculum/Richness)

To test for underlying hierarchical structures in species' abundances, the matrix of average species abundance by inoculum for each environment at T4 was sorted by the leading eigenvector of this matrix, thereby maximizing nestedness (Staniczenko *et al.*, 2013). To calculate

single isolate growth rates in SOC liquid media, the grofit package (Kahm & Hasenbrink, 2010) was used to fit smoothed splines to each replicate absorbance curve for the first 50 h (*gcFitSpline* function). The maximum slope (growth rate in exponential phase) of each growth curve was extracted from this function. To test the relationship between species' relative abundances and plant fresh mass, mass was first transformed into z-scores and a multinomial log-linear model (*multinom* function in the nnet package (Venables & Ripley, 2013)) was fit to the mass and abundance data. For comparison, inoculated plants were compared to buffer control treated plants to test the effects of each species. Based on fresh mass, the log odds of each species' abundance versus buffer treatment alone was given by the model coefficients. A two tailed Wald z-test was used to assess significance of coefficients.

3.4 Results

We tested the effects of environmental heterogeneity and initial species richness on succession of replicate communities under controlled conditions. At richness levels of 2, 3, 5, 8, and 11 species, we randomly chose five different combinations of isolates from a pool of twelve species (Table 3.1). We inoculated these twenty-five combinations (Table 3.2) into three different environments (liquid media, solid media, gnotobiotic plant), and followed community structures over time by 16S rRNA gene amplicon sequencing. To account for differences in sequencing depth, reads for each sample were rarefied to 1000 sequences. Rarefaction was completed 1000 times, and the averages for isolate abundances were determined and used for all downstream analysis except where indicated differently. As bacteria differ in 16S rRNA gene copy numbers, we corrected for these differences using the 16S rRNA gene copy numbers determined with 16Stimator (Perisin *et al.*, 2015).

3.4.1 Environment type and initial species richness affect community structure

There were two axes of complexity along which we mixed and inoculated subgroups of bacteria from our isolate pool: environmental heterogeneity and initial number of species. We first examined whether communities from each environment and species richness level had different community structures. To test for differences between multivariate centroids, we computed pairwise Bray-Curtis dissimilarities for all samples at all time points and performed a PERMANOVA test (see methods). Not surprisingly, the largest effect was due to inoculum combination nested in species richness ($F_{20} = 80.00$, $R^2 = 0.35$, $P = 0.001$). Even given the obvious effect of initial composition, there were additional significant effects of environment ($F_2 = 188.95$, $R^2 = 0.08$, $P = 0.001$) and species richness ($F_4 = 121.68$, $R^2 = 0.11$, $P = 0.001$). Environmental effects are evident in PCoA plots in which, particularly in the case of 11 species combinations, communities from each environment initially cluster together and then cluster apart over time (Figure 3.1). The magnitude of environmental effects varied with initial species richness, but were still significant by PERMANOVA at each richness level (2 sp: $F_2 = 20.24$, $R^2 = 0.05$, $P = 0.001$; 3 sp: $F_2 = 52.36$, $R^2 = 0.07$, $P = 0.001$; 5 sp: $F_2 = 46.30$, $R^2 = 0.13$, $P = 0.001$; 8 sp: $F_2 = 58.23$, $R^2 = 0.19$, $P = 0.001$; 11 sp: $F_2 = 67.69$, $R^2 = 0.28$, $P = 0.001$). Species richness effects are harder to visualize with PCoA but differences in clustering by richness are still evident by comparison of PCoA scores at each environmental level, and these effects are significant by PERMANOVA (Figures 3.2 - 3.4; Liquid: $F_4 = 41.26$, $R^2 = 0.12$, $P = 0.001$; Solid: $F_4 = 152.36$, $R^2 = 0.22$, $P = 0.001$; Plant: $F_4 = 25.54$, $R^2 = 0.13$, $P = 0.001$).

3.4.2 Environment and species richness affect replicate community structure repeatability

We next asked whether the environment and species richness affect the repeatability of community succession. To test repeatability, we calculated pairwise Bray-Curtis distances for replicate communities only and tested for environment and species richness effects by ANOVA (see methods). Replicate dissimilarity and repeatability are inversely related. We observed significant effects of environment ($F_2 = 388.89$, $P < 0.001$) and species richness ($F_4 = 113.40$, $P < 0.001$). Not surprisingly, replicate dissimilarity increased with increasing species richness (Figure 3.5a). Interestingly, replicate dissimilarity did not necessarily increase with increasing environmental complexity. Solid communities generally displayed the lowest replicate dissimilarity followed by liquid then plant communities (Figure 3.5b), and this trend is evident in the majority of species combinations (Figure 3.6). There was also a significant effect of time on repeatability ($F_3 = 9.74$, $P < 0.001$). To explore whether communities were converging or diverging in structure over time, we plotted dissimilarities by time for each environment. Surprisingly, replicate communities in liquid media increase in dissimilarity over time, solid media communities show no change over time, and plant dissimilarities increase and then decrease over time (Figure 3.7).

3.4.3 Each environment has a different species abundance hierarchy

Given that we observed strong environmental effects on community composition, we further tested which species differed in abundance by environment. To test for underlying patterns in abundances, we sorted the matrix of mean species abundances by inoculum and environment at T4 by the leading eigenvector, thereby testing for nestedness (Staniczenko *et al.*, 2013). By plotting abundances in this way, it became apparent that each environment had its own species abundance hierarchy (Figure 3.8). If the most dominant species was inoculated into a particular environment, that species was the most abundant. If that dominant species was

not inoculated, then the next most dominant species was the most abundant and so forth down the hierarchy. In the case of liquid media communities, we expected that the hierarchy would correspond with single isolate growth rates; however, this was not the case (Figure 3.9). We did not test growth rates in the solid and plant environments. Interestingly, two species, *P. viridiflava* and *C. flaccumfaciens*, were barely detected in liquid and solid media, but both rose to high abundances *in planta*. These two species had different relationships with plant fresh mass (Figure 3.10). A multinomial log-linear test revealed that small plants were more likely to contain *P. viridiflava* rather than buffer treatment, and large plants were equally likely to have been buffer treated or have an abundance of *C. flaccumfaciens*.

3.5 Discussion

In this study, we characterized bacterial succession across two axes of complexity, environmental heterogeneity and species richness. In every combination of inoculated isolates, the type of environment affected community structure. In the vast majority of richness and environment combinations, we did not observe convergent community assembly, as evidenced by high replicate dissimilarities that increased over time. Surprisingly, replicate liquid media communities demonstrated the highest dissimilarities even given the homogenous nature of this environment. We observed these liquid communities diverge further from each other over time, as opposed to solid media or plant communities that displayed relatively constant or decreasing dissimilarity over time.

Our liquid media results are similar to the community succession observed in Celiker & Gore (2014) in which six bacterial isolates were inoculated into 96 wells and experimentally evolved. We did not include enough replicates to fully assess whether communities converged on different states for each inoculum combination. In contrast to Celiker & Gore (2014), we did not observe consistent community structures over short time spans. We did not screen bacterial isolates for the ability to coexist prior to inoculation and hence, we may observe stronger competitive interactions, which are thought to be the prevailing interaction type in

culture (Foster & Bell, 2012).

Overall, we demonstrated that community assembly in rich liquid media is highly variable. These results could have implications for bioreactor performance. There already exists conflicting reports of community structure reproducibility (Falk *et al.*, 2009; Vanwonterghem *et al.*, 2014) and unpredictability (Zhou *et al.*, 2013; Pagaling *et al.*, 2014). As we observed the greatest variability under rich liquid media conditions with the least selective environmental pressures, it may be important for bioreactor operators to impose continuous pressure to the community to obtain the desired community structure and function. If there is high redundancy between community members, than the exact structure may not matter.

The highest repeatability in community structure occurred in solid media. We believe this is due to strong initial competition for resources. The solid media environment provides a spatial dimension that we originally thought would add more variability. As communities were plated to confluency, perhaps there was not enough space for isolates to segregate, and therefore they were forced to strongly compete for nutrients. The near constant dissimilarity indicates that community structure was set early and though slightly different between replicates, this dissimilarity was not amplified over time.

Based on sorting the species abundance matrices by the leading eigenvectors, we observed abundance hierarchies in each environment. Interestingly, the order of dominance is not the same for each environment. *P. simiae* is highly dominant on solid media, but the same media in liquid form allows other species to rise in abundance and outcompete *P. simiae*. Initially, we expected the liquid competition order to mirror the order in growth rates, but this was not the case. *P. syringae* and *P. viridiflava* had high growth rates and were outcompeted by almost all other isolates. *In planta*, the dominance order is striking. Two isolates that have different effects on plant fresh mass (*P. viridiflava* negative, *C. flaccumfaciens* non-significant compared to buffer treatment) were outcompeted in liquid and solid media, but were strong competitors *in planta*. Also of interest, are isolates that can outcompete pathogens in plant colonization, such as *P. simiae*, *C. flaccumfaciens*, *P. fluorescens*, and *Microbacterium sp.*

These isolates have potential for biocontrol.

3.6 Acknowledgments

Thanks to Stefano Allesina who co-advised this work. Thanks to Leah Johnson for statistical modeling advice.

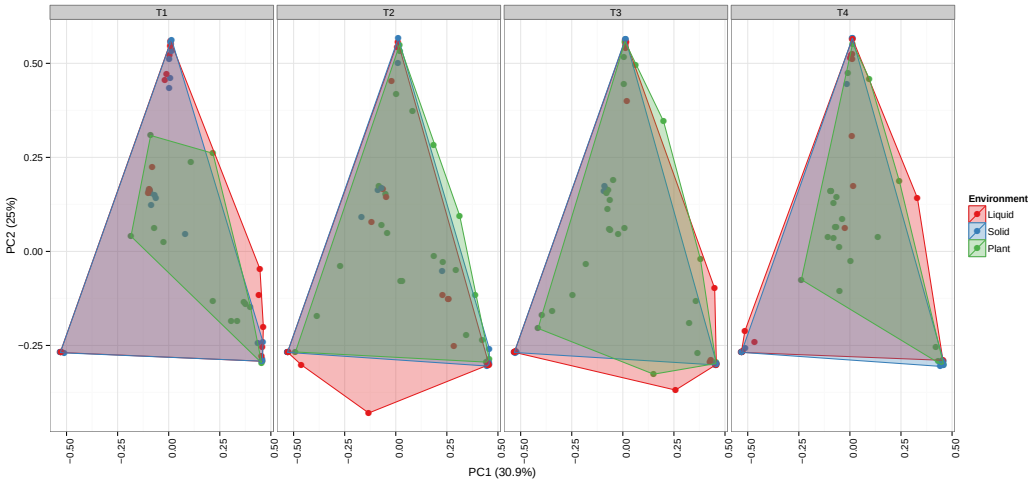
3.7 Tables and figures

Strain	Class	Putative species	WGS accession	Plant interaction	Isolation reference
RMX3.1b	Gammaproteobacteria	*Pseudomonas viridiflava*	Not deposited	Pathogen	Jakob *et al.* (2002)
NP29.1a	Gammaproteobacteria	*Pseudomonas syringae*	Not deposited	Pathogen	Jakob *et al.* (2002)
MEB105	Gammaproteobacteria	*Pseudomonas simiae*	JXQT00000000	Biocontrol	Traw *et al.* (2007)
MEP34	Gammaproteobacteria	*Pseudomonas fluorescens*	JXQY00000000	Biocontrol	Barrett *et al.* (2011)
MEJ086	Gammaproteobacteria	*Pseudomonas fulva*	JXQW00000000	Unknown	Kniskern *et al.* (2007)
LMCP11	Gammaproteobacteria	*Xanthomonas campestris*	JXQP00000000	Pathogen	Barrett *et al.* (2011)
MEJ076	Alphaproteobacteria	*Agrobacterium tumefaciens*	JXQV00000000	Pathogen	Kniskern *et al.* (2007)
MEDvA23	Betaproteobacteria	*Variovorax paradoxus*	JXQQ00000000	Unknown	Barrett *et al.* (2011)
MEB061	Flavobacteria	*Flavobacterium sp.*	JXQR00000000	Unknown	Traw *et al.* (2007)
MEB126	Actinobacteria	*Curtobacterium flaccumfaciens*	JXQU00000000	Unknown	Traw *et al.* (2007)
MEB024	Actinobacteria	*Frigoribacterium sp.*	JXQZ00000000	Unknown	Traw *et al.* (2007)
MEJ108Y	Actinobacteria	*Microbacterium sp.*	JXQX00000000	Unknown	Kniskern *et al.* (2007)

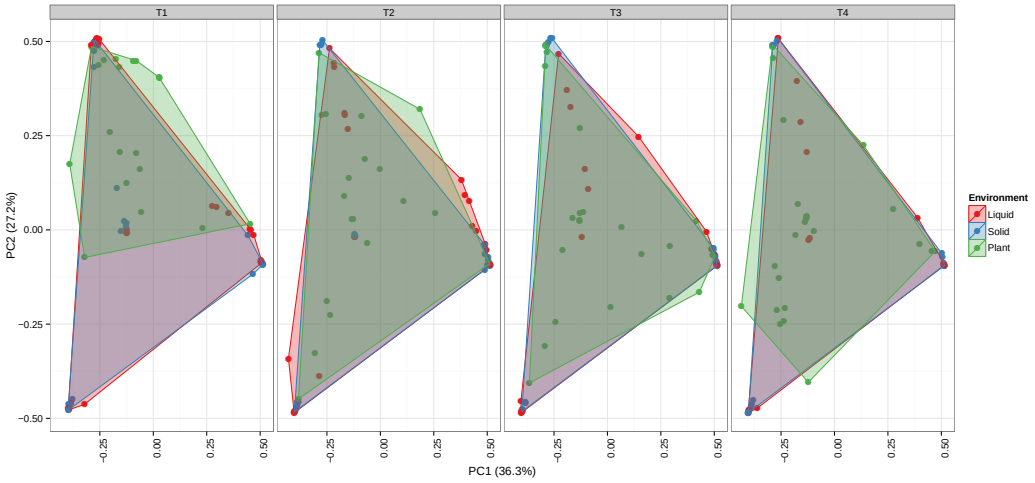
Table 3.1: Bacterial strains used in this study

Number	Number of Isolates	*P. viridiflava*	*P. syringae*	*P. simiae*	*P. fluorescens*	*P. fulva*	*X. campestris*	*A. tumefaciens*	*V. paradoxus*	*Flavobacterium sp.*	*C. flaccumfaciens*	*Frigoribacterium sp.*	*Microbacterium sp.*
2sp.A	2	X	X										
2sp.B	2			X		X							
2sp.C	2		X									X	
2sp.D	2									X	X		
2sp.E	2					X							X
3sp.A	3				X			X					X
3sp.B	3	X		X									X
3sp.C	3						X		X	X			
3sp.D	3	X			X						X		
3sp.E	3	X				X		X					
5sp.A	5			X	X		X		X			X	
5sp.B	5			X		X	X	X	X				
5sp.C	5		X	X	X				X				X
5sp.D	5	X	X	X						X	X		
5sp.E	5			X	X	X		X					X
8sp.A	8	X	X	X			X	X	X			X	X
8sp.B	8	X	X	X	X	X	X			X			X
8sp.C	8	X		X	X	X	X			X		X	X
8sp.D	8	X		X	X		X	X			X	X	X
8sp.E	8	X				X	X	X		X	X	X	X
11sp.A	11	X	X	X	X		X	X	X	X	X	X	X
11sp.B	11	X	X	X	X	X		X	X	X	X	X	X
11sp.C	11	X	X	X	X	X	X	X	X	X	X		X
11sp.D	11	X	X	X	X	X	X	X	X	X	X	X	
11sp.E	11	X	X	X	X	X	X		X	X	X	X	X

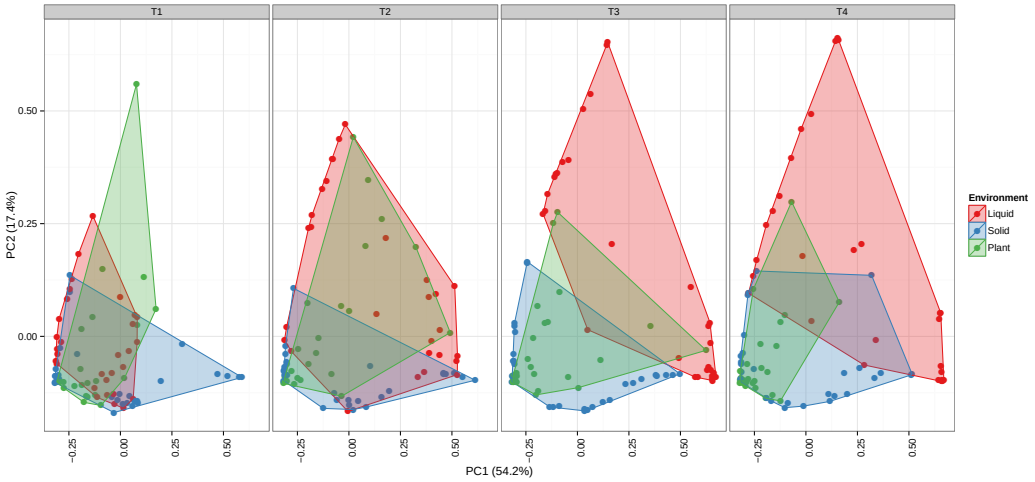
Table 3.2: Isolate combinations used in this study



(a)

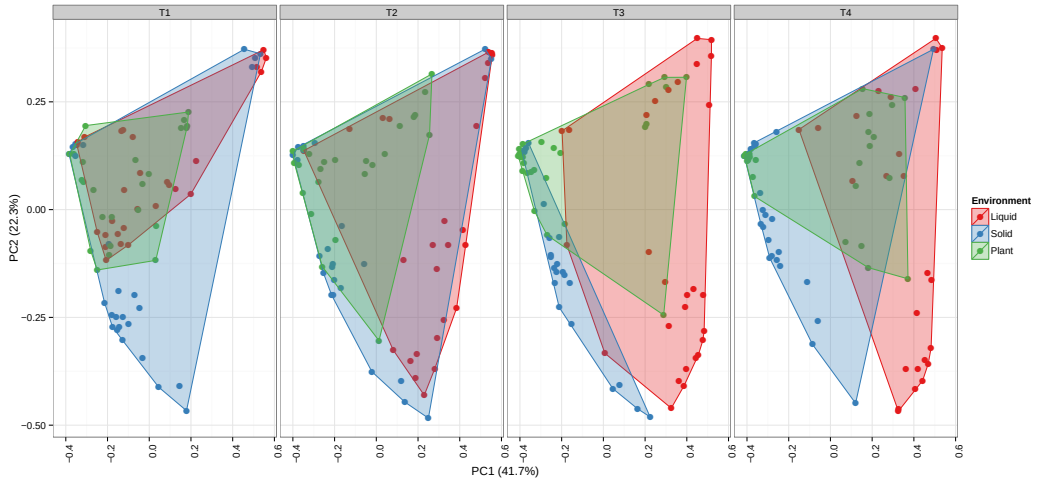


(b)

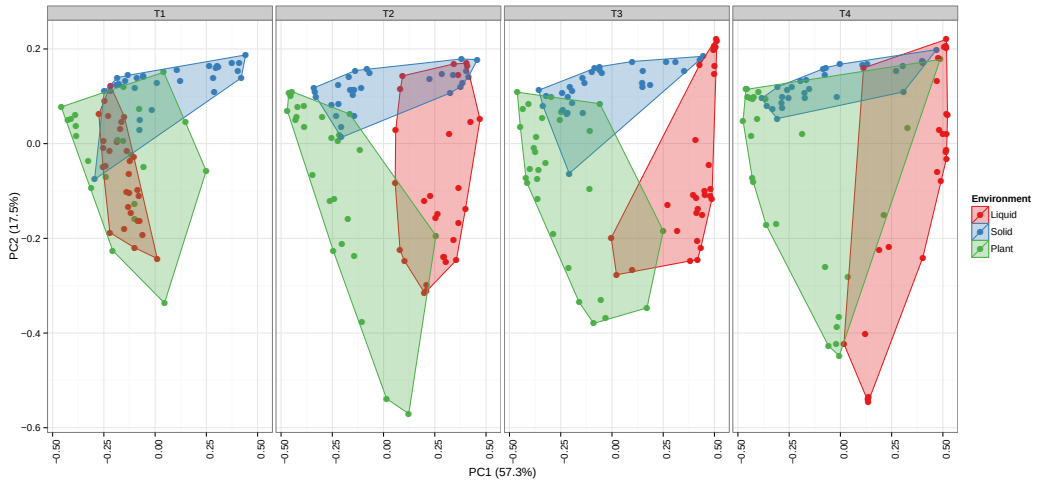


(c)

Figure 3.1: Bacterial community structures cluster by environment.



(d)



(e)

Figure 3.1, continued

Pairwise Bray-Curtis distances were calculated for all samples inoculated with (a) 2 species, (b) 3 species, (c) 5 species, (d) 8 species, or (e) 11 species. The distance matrix was subjected to principal coordinates analysis with sample scores for top two principal coordinates shown. Percentages indicate the amount of variation explained by each axis. All samples for each species richness level were analyzed together, faceted by time point, and colored by environment. Polygons represent the convex hull for samples originating from each environment at each time point.

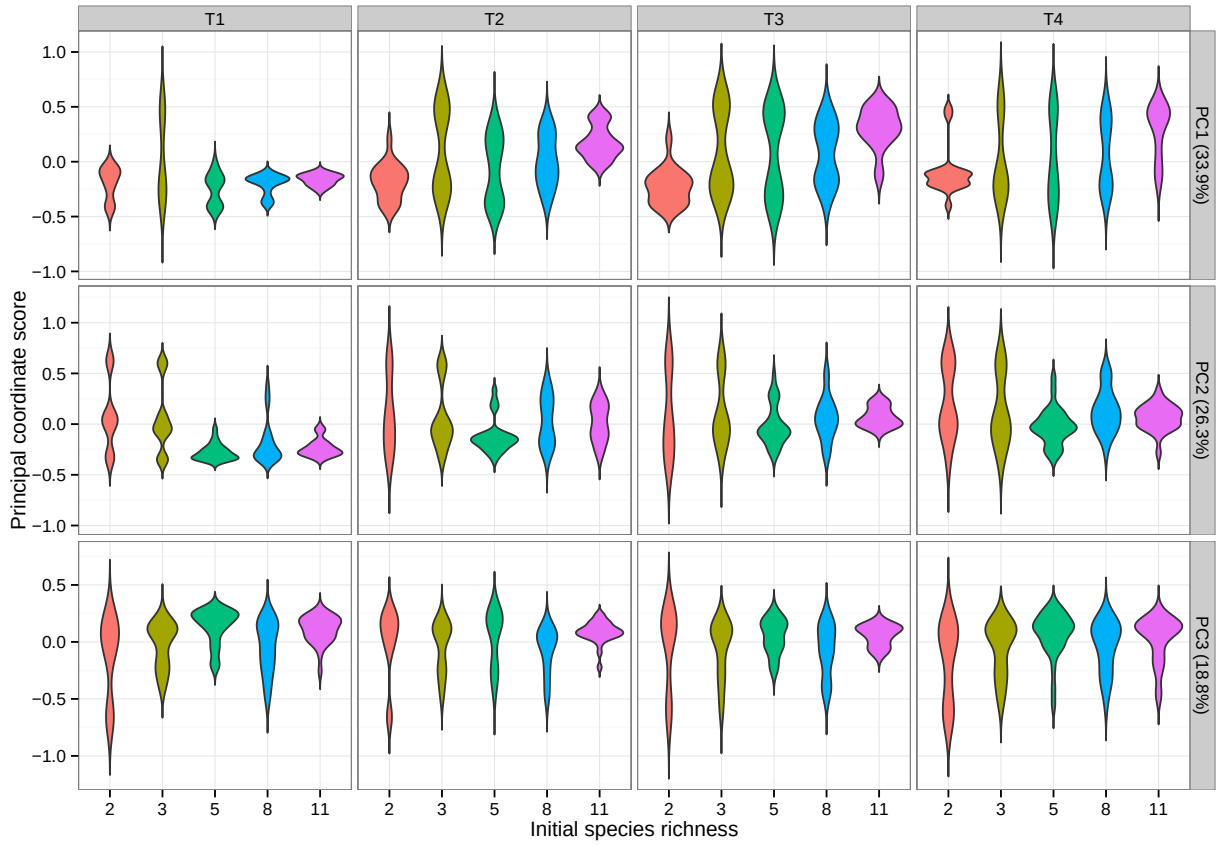


Figure 3.2: Liquid media bacterial community structures vary by initial species richness.

Pairwise Bray-Curtis distances were calculated for all liquid media samples. The distance matrix was subjected to principal coordinates analysis, and sample scores for top three principal coordinates were violin plotted by initial species richness and faceted by principal coordinate and time point. Percentages indicate the amount of variation explained by each principal coordinate.

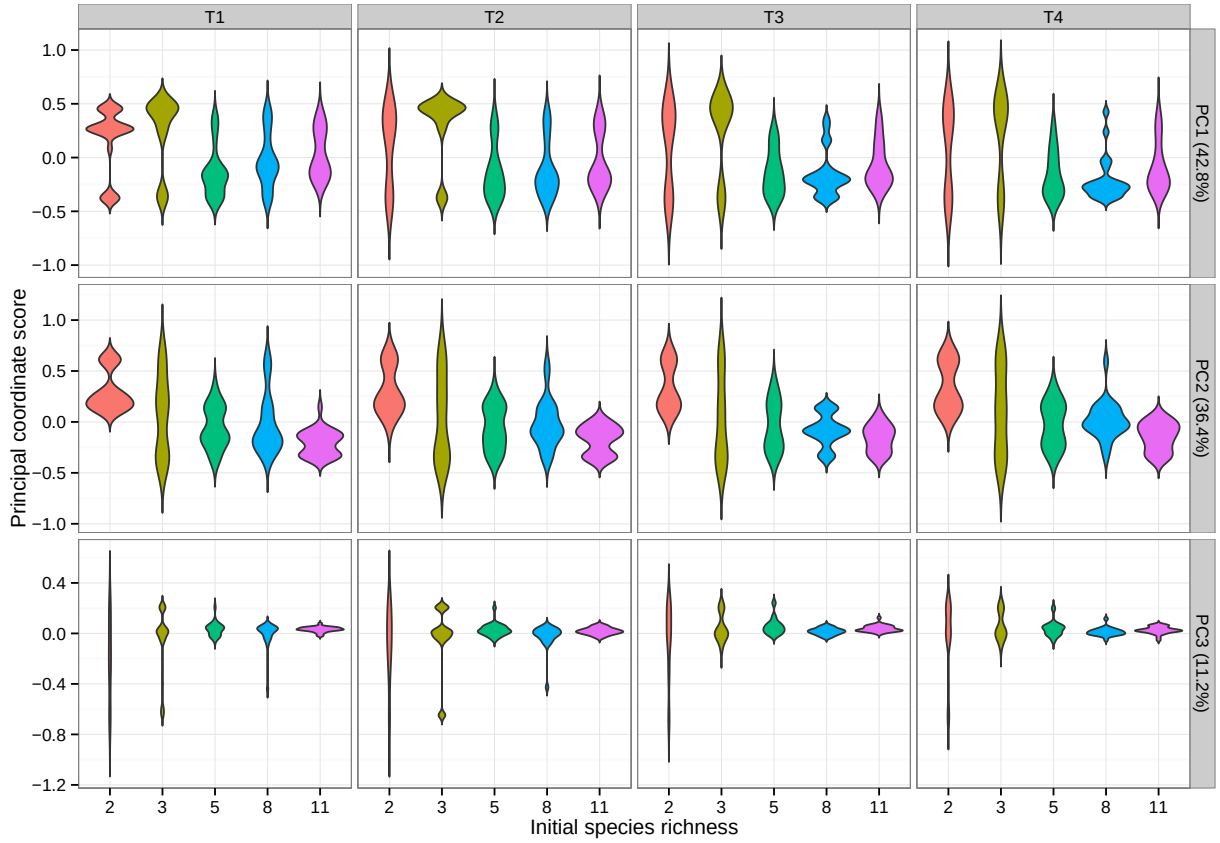


Figure 3.3: Solid media bacterial community structures vary by initial species richness.

Pairwise Bray-Curtis distances were calculated for all solid media samples. The distance matrix was subjected to principal coordinates analysis, and sample scores for top three principal coordinates were violin plotted by initial species richness and faceted by principal coordinate and time point. Percentages indicate the amount of variation explained by each principal coordinate.

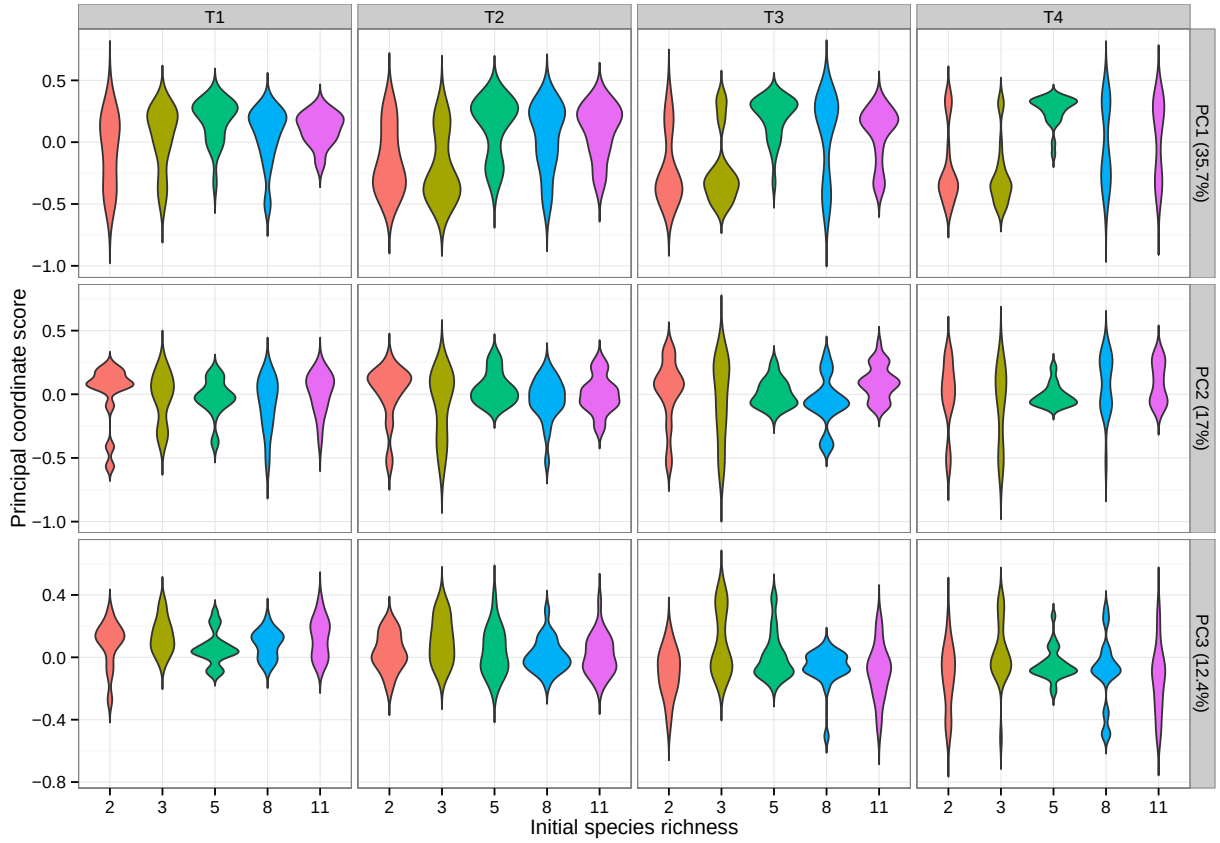
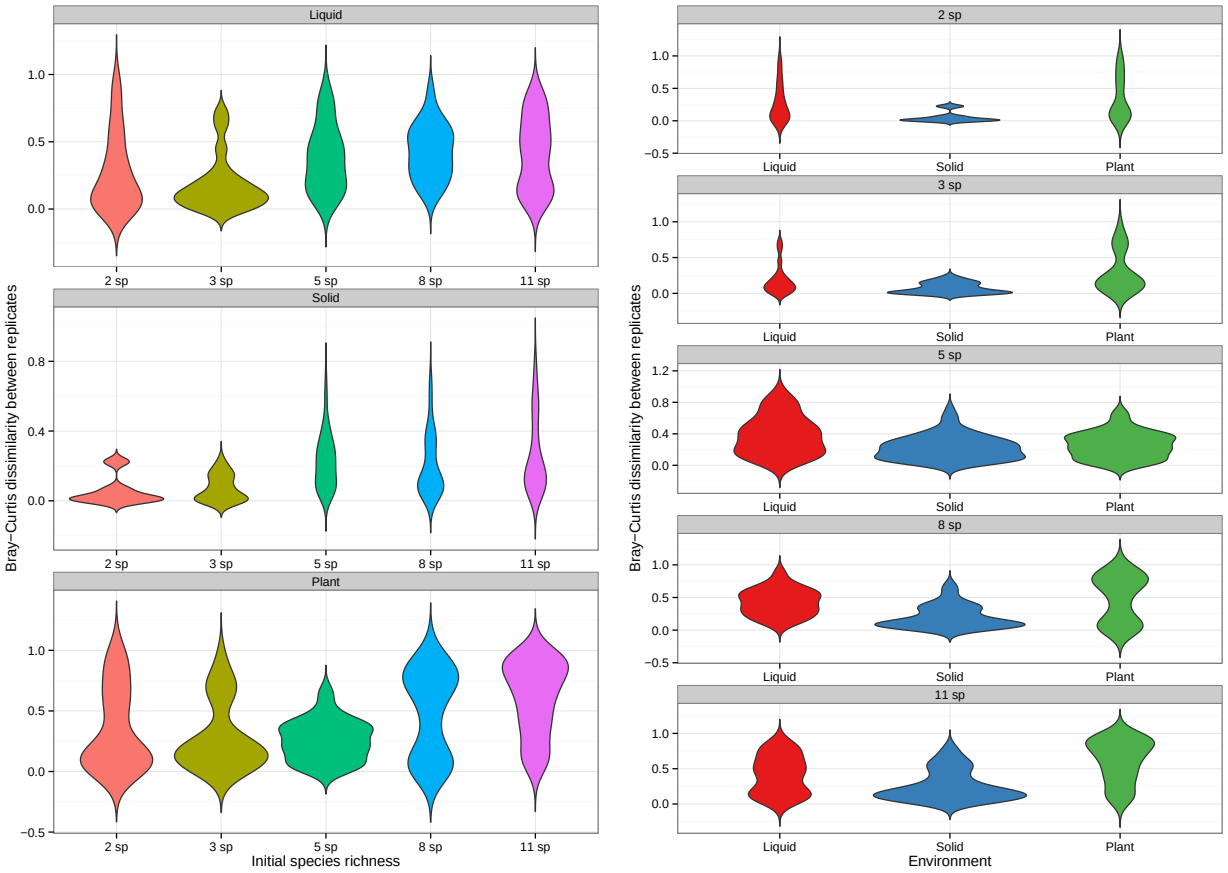


Figure 3.4: Plant bacterial community structures vary by initial species richness.

Pairwise Bray-Curtis distances were calculated for all plant samples. The distance matrix was subjected to principal coordinates analysis, and sample scores for top three principal coordinates were violin plotted by initial species richness and faceted by principal coordinate and time point. Percentages indicate the amount of variation explained by each principal coordinate.



(a) (b)

Figure 3.5: Replicate dissimilarity increases with species richness but not with environmental complexity.

For all time point T4 samples, Bray-Curtis dissimilarities were calculated for pairwise combinations of replicates of each inoculum and environment combination. These dissimilarities were pooled and violin plotted. In (a), dissimilarities are colored by initial species richness and faceted by environment. In (b), dissimilarities are colored by environment and faceted by initial species richness.

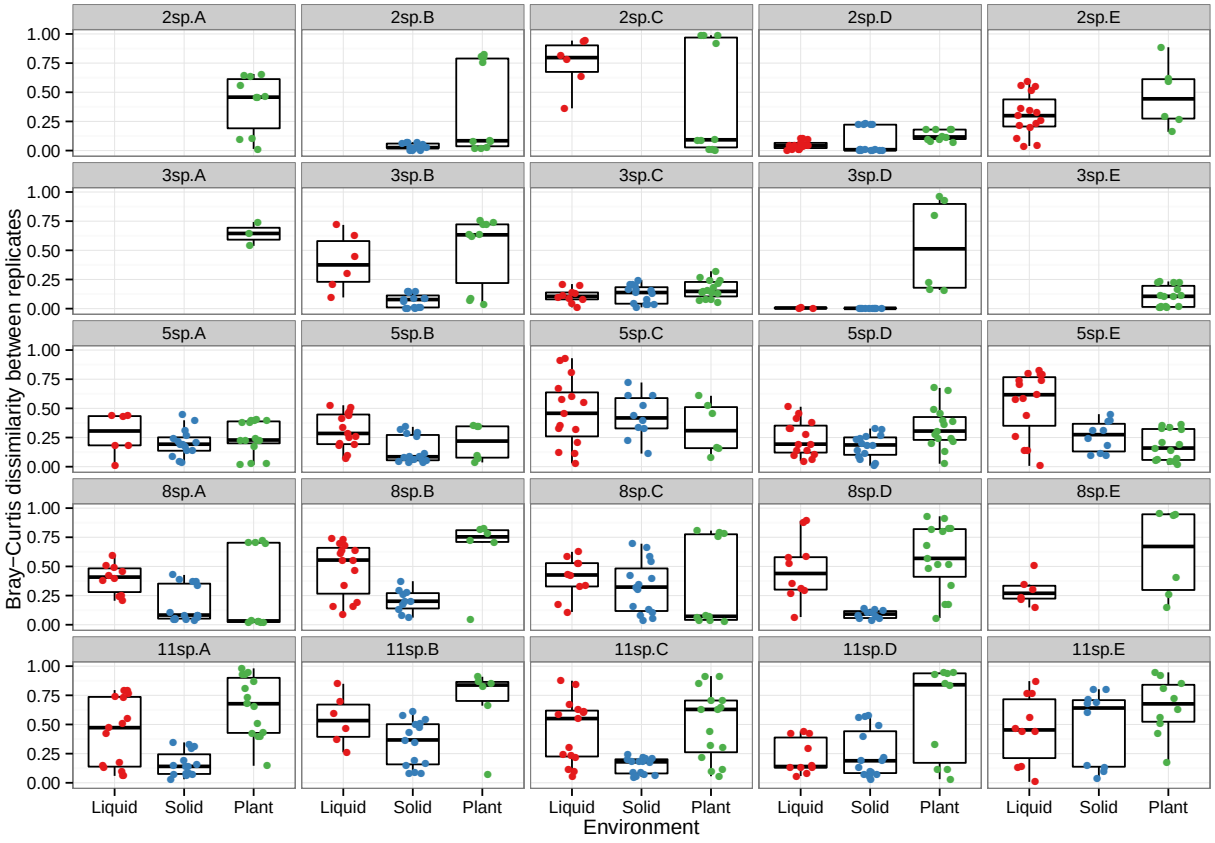


Figure 3.6: Replicate Bray-Curtis dissimilarity for each species combination.

For all time point T4 samples, Bray-Curtis dissimilarities were calculated for pairwise combinations of replicates of each inoculum and environment combination. These dissimilarities were box plotted by environment for each inoculum.

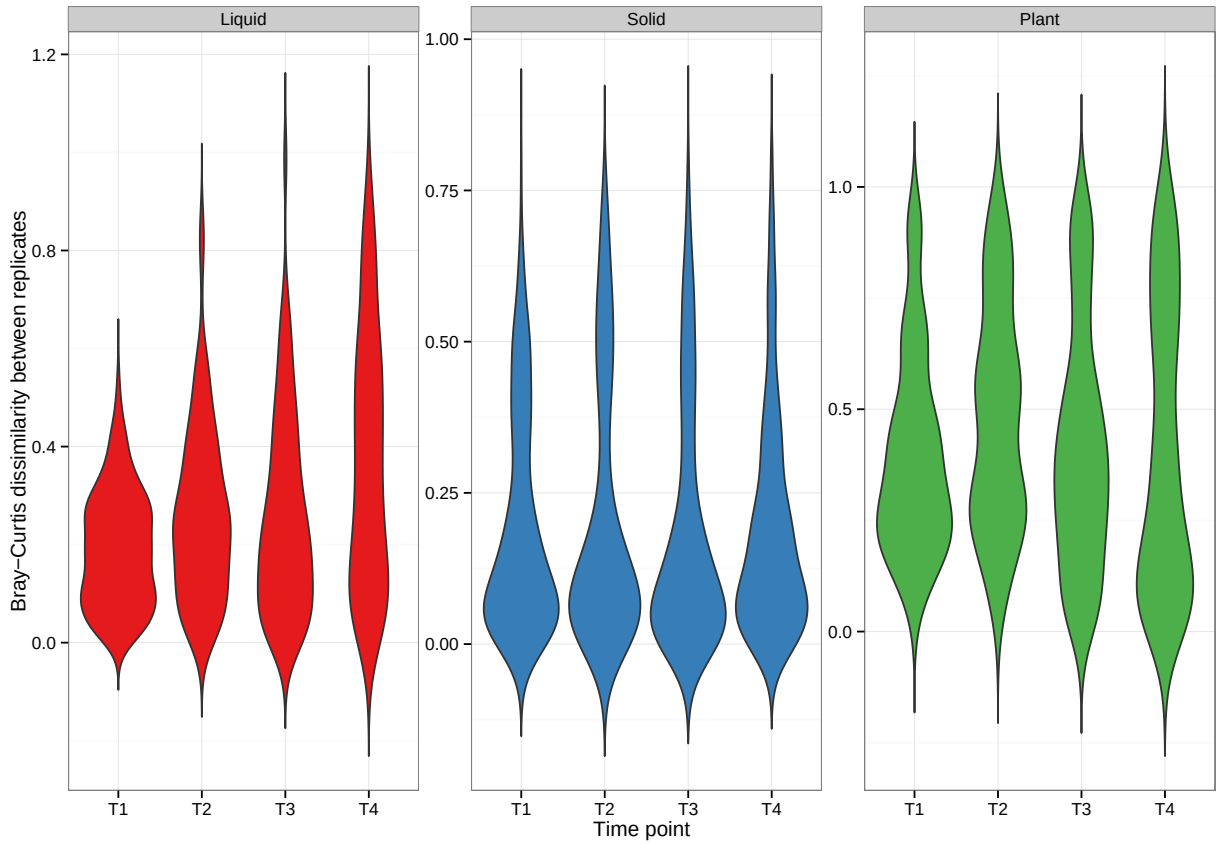


Figure 3.7: Replicate dissimilarity time trends by environment.

Pairwise Bray-Curtis dissimilarities were calculated for each inoculum, environment, and time combination of replicates. These values were pooled and violin plotted by time point and faceted by environment.

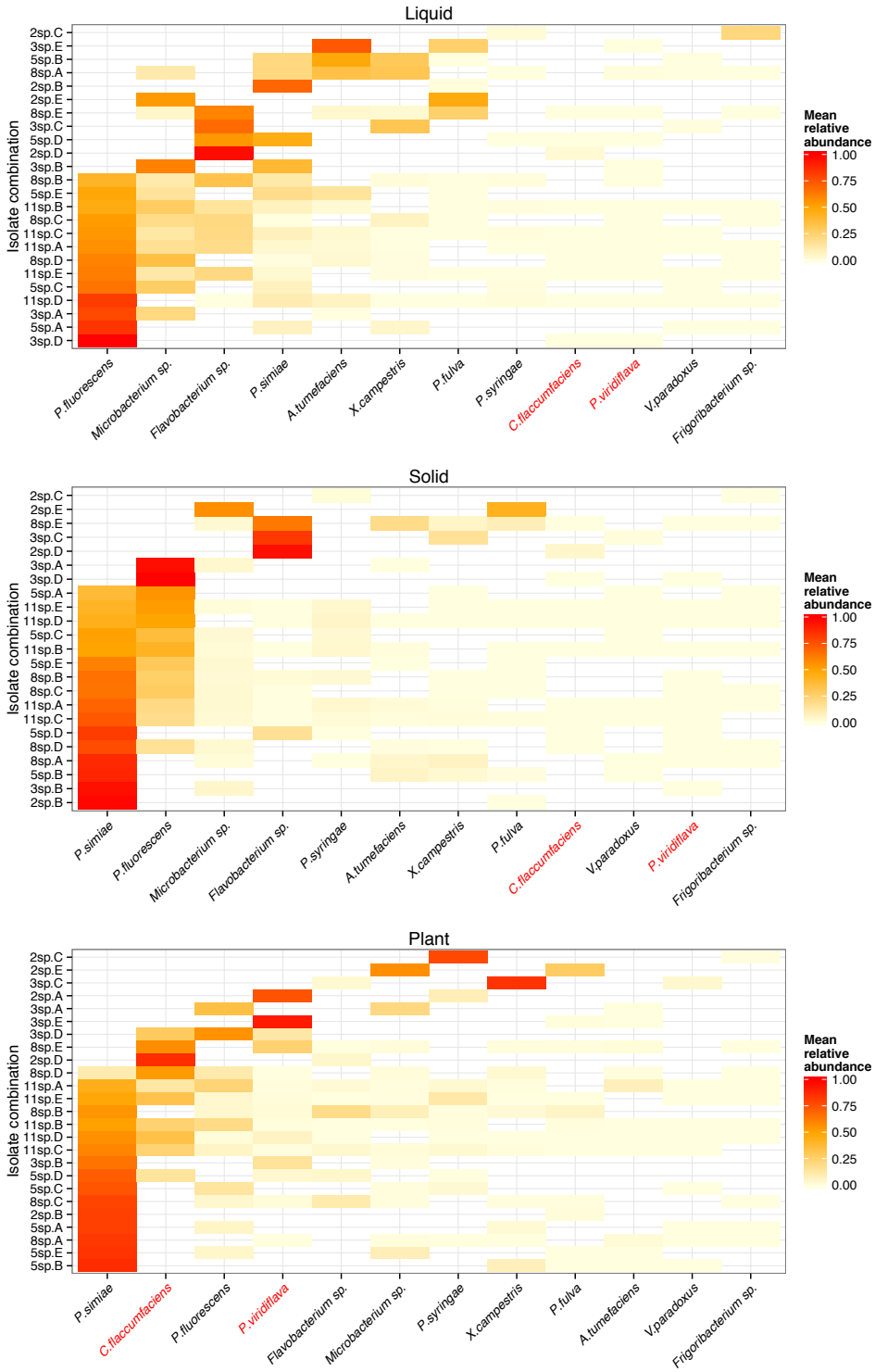


Figure 3.8: Each environment has a different species abundance hierarchy.

For each environment at T4, the inoculum by isolate matrix of mean relative abundances was sorted by the leading eigenvector, revealing hierarchical patterns. The color of each tile is the mean relative abundance of each isolate for each inoculum.

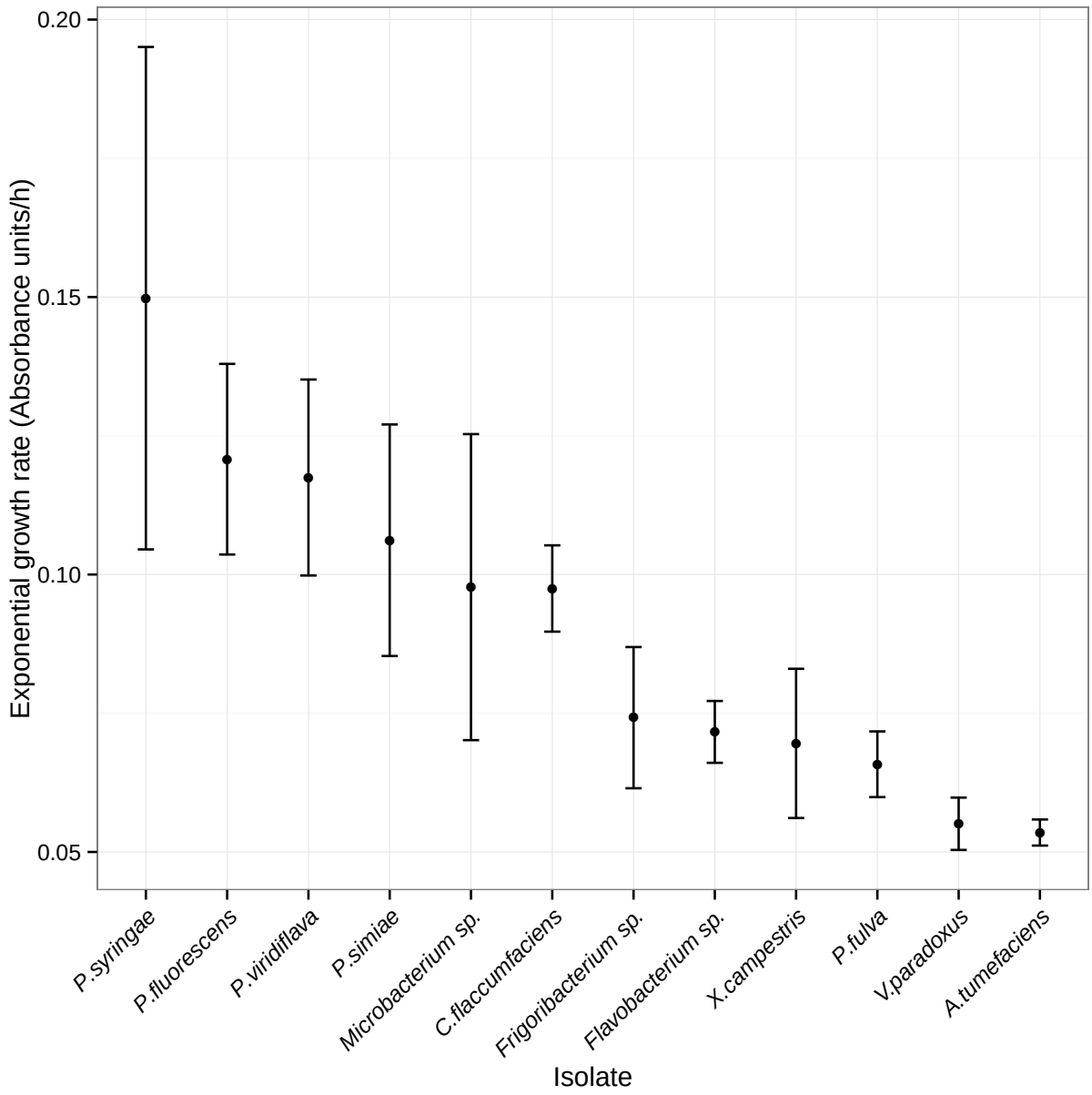


Figure 3.9: Isolate growth rates in liquid SOC media

Each isolate was grown alone (7 replicates) in SOC liquid broth at 28°C with continuous shaking, and absorbance at 600 nm was measured every 10 min. The *grofit* package was used to fit smoothed splines and extract the maximum slope (exponential growth rate) for each replicate's growth curve. The mean and standard deviation of each isolate's growth rate is shown.

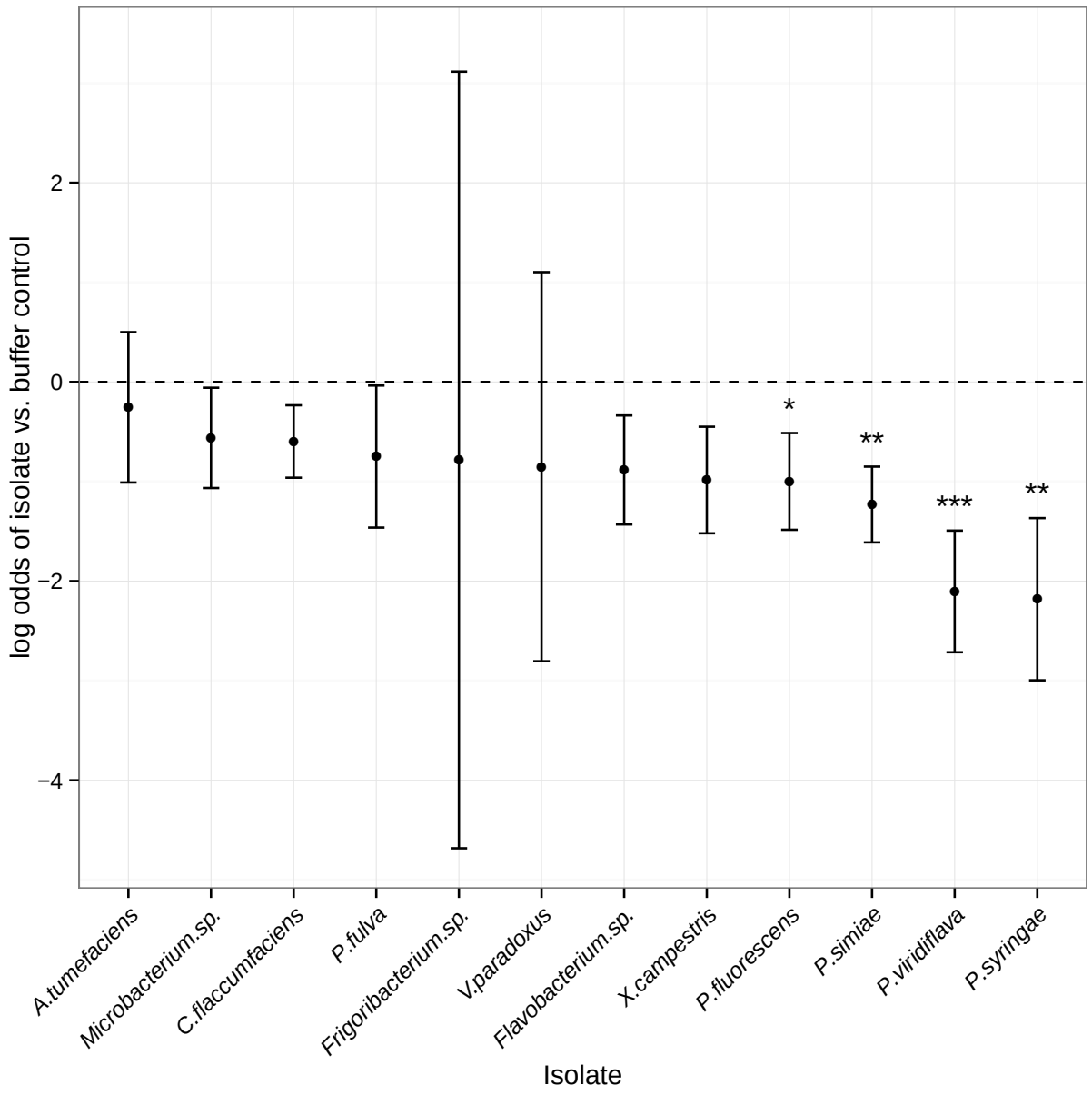


Figure 3.10: *P. viridiflava* and *C. flaccumfaciens* have different effects on plant fresh mass.

For all plant samples at time point T4, the relationship between species' relative abundances and plant fresh mass was tested with a multinomial log-linear model (*multinom* function in nnet package). Based on plant fresh mass, the log odds of each plant containing each isolate versus buffer only was determined. Each isolate's log odds plus and minus standard error is shown. Asterisks indicate significance based on Wald two-tailed z tests: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

CHAPTER 4

SPATIAL AND TEMPORAL DYNAMICS OF *ARABIDOPSIS THALIANA* ASSOCIATED BACTERIAL COMMUNITIES

4.1 Abstract

Plants harbor diverse arrays of microbes that can impact host phenotypes and fitness. Greater insights into the ecological forces that structure these microbial communities are needed to develop effective strategies to promote host health and resist pathogen invasion. In this chapter, I aimed to uncover spatial and temporal factors that structure bacterial communities associated with the tissues of *Arabidopsis thaliana*. Surface sterilized seeds from native accessions were planted at two sites in Southwestern Michigan over two years, and tissues were sampled over the entire plant life cycle. 16S rRNA gene amplicon sequencing was used to quantify bacterial community composition. Plant tissue type had the largest effect on community structure. Temporal observations of root associated communities indicated an initial shift away from soil communities but then a return to soil-like communities during flowering and senescence. Geographic site also had effects on community composition, but these effects varied by tissue. Site effects in soil communities were progressively diminished going aboveground in the roots, leaves, and stems, but were evident in the siliques. The convergence of OTU profiles given different source communities in the soil indicated that plant tissues act as bacterial sieves. From comparison of OTU presence/absence, aboveground tissues had the largest OTU overlap with the roots and not other aboveground tissues. However, based on relative abundances, aboveground tissues enriched for smaller and smaller subsets of these OTUs going from the leaves to stems to siliques. The most abundant OTUs for the *Sphingomonas* genus were enriched in aboveground tissues. Interestingly, the most abundant OTUs in the *Massilia* genus were differentially abundant depending on aboveground versus belowground tissue. These results suggest that the roots initially sieve bacterial taxa from the soil that are then distributed to aboveground tissues. Each tissue enriches for specific

subsets of OTUs.

4.2 Introduction

Plants are colonized inside and out by a diverse set of microbes. Among these colonizers are pathogenic, beneficial, and commensal organisms that impact host phenotype and fitness (Bulgarelli *et al.*, 2013). These microbes rarely exert these effects in isolation, but instead as members of complex multi-species communities. For instance, microbial communities in the soil are capable of altering flowering time (Wagner *et al.*, 2014; Panke-Buisse *et al.*, 2015). Understanding the ecological forces which structure these communities has become a intense area of research in hopes of developing strategies to promote plant health and resist pathogen invasion. These forces include abiotic and biotic factors that have different relative impacts on community assembly depending on the plant tissue of interest. Thus far, most studies have focused on microbes inhabiting the roots and surrounding soil (rhizosphere) or the surface and apoplast of leaves (phyllosphere).

The rhizosphere and interstitial areas of roots (endophytic compartment) are nutrient rich environments, so much so, that they have been compared to the human gut (Hacquard *et al.*, 2015). Roots exude sugars and amino acids which are thought to attract a wide range of microbes (Chaparro *et al.*, 2013). Abiotic factors such as soil pH, water content, and C:N ratio influence the identity and quantity of particular bacterial taxa that will even have the chance to interact with the root system (Fierer *et al.*, 2009). From this pool of bacteria, there has been evidence of community filtering by biotic factors going from the bulk soil to rhizosphere to endophytic compartment (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012). By planting different genotypes of surface sterilized *A. thaliana* seeds in soil collected from various geographic locations, Lundberg *et al.* (2012) noted a subset of bacterial operational taxonomic units (OTUs) that were enriched in the endophytic compartment compared to the soil, and this enrichment was concurrent with a reduction in Shannon diversity. Interestingly, biotic factors such as genotype and growth stage (bolting versus entering senescence) had

significant but marginal effects on community structure compared to site effects (Lundberg *et al.*, 2012). In a similar study, Bulgarelli *et al.* (2012) also observed filtering of soil OTUs in the rhizosphere and endophytic compartment of *A. thaliana*. Again, host genotype had little effect compared to the site origin of soil (Bulgarelli *et al.*, 2012), but comparison of community assembly on metabolically active roots compared to wood sticks, revealed a 40% overlap in OTU membership. This finding indicates that the cell wall has a predominant role in filtering soil OTUs, followed by further filtering by a metabolically active environment (Bulgarelli *et al.*, 2012). Plant defenses and immune signaling likely contribute to this filtering based on the recent finding that salicylic acid biosynthesis and signaling pathways are required to assemble a "normal" endophytic compartment community (Lebeis *et al.*, 2015).

From a temporal aspect, few studies have observed bacterial community succession over multiple plant growth stages. Contrasting with Lundberg *et al.* (2012), Chaparro *et al.* (2014) observed different bacterial communities in the *A. thaliana* seedling rhizosphere compared to vegetative, bolting, and flowering stage associated communities. Small sample sizes may have reduced the power to differentiate the communities from the later time points. Different temporal patterns in root exudation may influence community assembly (Chaparro *et al.*, 2013). By growing *A. thaliana* in gnotobiotic conditions, Chaparro *et al.* (2013) observed that sugars and sugar alcohols were primarily exuded at early life stages compared to later stages in which amino acids and phenolics were predominent in exudates. Direct connection of exudation patterns with shifts in bacterial community composition has yet to be shown, indicating a need for more temporal studies, especially in natural environments.

In contrast to the nutrient rich rhizosphere, is the nutrient poor and heterogeneous phyllosphere environment (Vorholt, 2012). This environment is subjected to fluctuating temperature, humidity, and UV conditions (Lindow & Brandl, 2003), but once inside the apoplast, microbes have the opportunity to leach nutrients from active photosynthesis processes. These differences in abiotic and biotic factors between the phyllosphere and rhizosphere contribute

to differences in bacterial community compositions (Bodenhausen *et al.*, 2013). Endo and epiphytic leaf communities were also shown to have decreased richness and evenness compared to rhizosphere and endopytic root communities (Bodenhausen *et al.*, 2013). Just as in the roots, biotic factors are thought to influence community assembly in the leaves. *A. thaliana* treated with salicylic acid were shown to have reduced bacterial titers and increased fitness in the field (Traw *et al.*, 2007). Further, a genome-wide association study of the microbial communities associated with *A. thaliana* in the field resulted in defense response and cell wall gene associations with community compositions (Horton *et al.*, 2014). Clearly, active plant processes influence microbial community assembly.

Temporal observation of phyllosphere communities in relation to microbes inhabiting surrounding environments has suggested that leaves filter microbes from these sources. In a common garden experiment with soybean, canola, and common bean plants, Copeland *et al.* (2015) observed that leaf bacterial communities initially resembled soil communities, but then diverged over the plant life cycle to become more leaf-specific. With merlot grapevines, Zarraonaindia *et al.* (2015) similarly observed that leaf, flower, and grape bacterial communities shared more OTUs in common with soil than other aboveground tissues. Microbes inoculated onto leaves from the air likely also impact community succession in the phyllosphere. By planting *A. thaliana* on initially sterile soil and observing bacterial community succession on the leaves in addition to in the air, Maignien *et al.* (2014) observed that in terms of OTU identities, phyllosphere communities initially resembled microbes in the air before diverging to become more leaf-specific. In terms of relative abundances, leaf communities diverged from each other depending on spatial location in the growth chamber, thus indicating priority effects as forces shaping community assembly (Maignien *et al.*, 2014).

In this chapter, I expand on these previous findings by characterizing the spatial and temporal dynamics of bacterial communities associated with *A. thaliana* at two sites in Southwest Michigan. To appropriately observe natural community succession with native plants, I planted surface sterilized seeds from North American *A. thaliana* accessions (ecotypes)

all belonging to the near isogenic haplogroup-1 (HPG1) (Hagmann *et al.*, 2015). Unlike Eurasian accessions which display fine-scale isolation by distance patterns, North American accessions are highly similar even at the continental scale (Platt *et al.*, 2010). Unlike most previous studies that focused on either root or leaf associated microbes, I isolated bacterial communities from both of these tissues in addition to stems and siliques. I sampled these tissues over the entire life cycle of the plant. In the Midwestern United States, *A. thaliana* germinates in the fall, overwinters as a small rosette, flowers in the spring, and senesces in the early summer. To quantify bacterial composition, I used 16S rRNA gene amplicon sequencing. Comparison of community structures revealed that plant tissue type had the largest effect on community assembly. The temporal aspect of this study allowed observation of root communities that initially diverged in structure from soil communities but then became more soil-like after flowering and senescence. Further comparison of communities between plant tissues, indicated that tissues act as progressive sieves. Roots initially sieve for particular OTUs that are then enriched in aboveground tissues.

4.3 Methods

Field experiments were replicated over the course of two years and at two locations: Michigan State Southwest Michigan Research and Extension Center (ME) in Benton Harbor, MI, USA (42.0853, -86.3588), and University of Chicago Warren Woods Ecological Field Station (WW) in Three Oaks, MI, USA (41.8362, -86.63). Sites were tilled prior to each year's planting.

4.3.1 Planting

Prior to planting, seeds from seven natural *A. thaliana* accessions, primarily collected from the Midwestern United States (Table 4.1) and all belonging to HPG1, were aliquoted into 1.5 mL Eppendorf tubes (100 seeds/tube for Year 1, 50 seeds/tube for Year 2). Each tube corresponded to a pot to be planted. Seeds were surface sterilized by addition of 70% ethanol,

incubation at room temperature for 1 minute, removal of ethanol, addition of 100% ethanol, incubation at room temperature for 1 minute, and removal of ethanol. Seeds were air dried in a sterilized biological cabinet.

For Year 1 planting, the bottoms of 11.43 cm diameter plastic plots were removed before being placed 2-5 cm into the ground. Pots were placed 30.48 cm apart in a grid (ME: 100 rows by 8 columns with rows going north to south; WW: 50 rows by 15 columns with rows going north to south). Surface sterilized seeds were spread in each pot. Ecotype and empty pots were completely randomized within each grid. The ME site was planted on October 12, 2012 and WW site was planted on October 15, 2012.

For Year 2 planting, black landscaping cloth was placed at each site after tilling to reduce native weed and grass growth (ME: 19 days prior to planting; WW: 7 days prior to planting). On the day of planting, pot size holes were cut in the cloth, the bottoms of 6.35 cm square pots were removed, and pots were placed 2-5 cm into the ground. Pots were placed in blocks (4 blocks/site, 1.22 m between blocks going north to south) with each block containing a 16 row by 10 column grid with 10 cm between pots (rows going north to south). Ecotype and empty pots were completely randomized in each block. Both sites were planted on October 28, 2013. After germination, seedlings were thinned (ME: November 20, 2013; WW: November 21, 2013) to 5-10 plants per pot and any native weed and grass was removed. Tweezers were ethanol sterilized between pots while thinning.

4.3.2 Sample collection

At each sampling time point (Table 4.2), the time and surface soil (2-5 cm deep) temperature were recorded for each sample. The sampling order was randomized over the entire grid for Year 1 (8 replicates per ecotype per site per time point) and over each block for Year 2 (2 replicates per ecotype/empty pot per block per site per time point). For empty pot soil samples, a sterile 15 mL conical tube was pushed into the ground (2-5 cm), soil was collected, and the tube was capped and placed on dry ice. For plant samples, a spatula was used to

remove root and above ground tissues. In a plastic tray, tweezers and a razor were used to remove root tissue from aboveground tissue. Each type of tissue was placed in a separate 15 mL conical tube. In between samples, the spatula, tweezers, and razor were flame sterilized with 100% ethanol, and the tray was sterilized with 100% ethanol. Tubes were placed on dry ice until transport back the University of Chicago. Tubes were stored at -80°C until processing.

4.3.3 Plant processing

Prior to processing, all samples were removed from -80°C, thawed, and the plant processing order was randomized, keeping the root/plant tissue together per plant. Samples were placed back at -80°C. Samples were then removed for 10 plants per processing round. To remove loosely associated microbes, each plant sample was washed. For each root or above ground sample, the tissue was removed and added to a 50 mL conical tube with 25 mL of surfactant buffer (Lundberg *et al.*, 2012) (6.33 g $NaH_2PO_4 \cdot H_2O$, 16.5 g $Na_2HPO_4 \cdot 7H_2O$, per 1 L, autoclaved then 200 μL Silwet L-77 added). Tubes were vortexed for 10 s then material was transferred to a new 50 mL conical tube with fresh 25 mL surfactant buffer. Tubes were vortexed again for 10 s. Tissue was removed and placed in a glass tray. Tweezers and a scalpel were used to separate tissues, where applicable, and to cut tissue into pieces small enough to fit into 1.4 mL Matrix tubes (Thermo Scientific, Waltham, MA, USA). Each tissue type was placed into Matrix tubes in 96 well racks and sealed with SepraSeal caps (Thermo Scientific). For large tissues, only enough material was added to allow for bead homogenization. For each soil sample, soil was sieved through a 2 mm sieve onto a glass tray. A spatula was used to add ∼100 mg of soil to the appropriate Matrix tube. Tweezers, spatula, scalpel, and glass tray were flame sterilized and the sieve was ethanol sterilized between samples. Matrix tubes were placed at -80°C overnight. The next morning, caps were removed and plates were covered in parafilm. Over each tube, a small hole was punched and the Matrix plates were lyophilized overnight (LABCONCO FreeZone 4.5, Kansas City,

MO, USA). Parafilm was then removed and tubes were sealed with new caps. Two 2.3 mm silica beads (BioSpec, Bartlesville, OK, USA) were added to each tube, and samples were ground to powder with a 2010 Genogrinder (SPEX, Metuchen, NJ, USA) (1750 RPM, 2 min). Dry mass was recorded and up to 36 mg of material was retained per tube. All tubes were then randomized over thirty 96 well DNA extraction plates, including empty well controls. Plates were kept at room temperature until DNA extraction.

4.3.4 DNA extraction

All following pipetting and shaking steps were completed using custom liquid handling robotics scripts on the Freedom Evo 200 (Tecan, Morrisville, NC, USA), unless indicated differently. All centrifugation steps were completed using the Beckman Coulter Avanti J-25 centrifuge (Beckman Instruments, Munich, Germany). Two 96 well plates were processed each day. To start, ground material was centrifuged at 6600xg for 2 min. Material was re-suspended in TES (10 mM Tris-Cl, 1 mM EDTA, 100 mM NaCl) to a concentration of 0.04 mg/μL. A minimum of 250 μL TES was added per tube. Tubes were sealed with new caps and material was homogenized with the 2010 Genogrinder (SPEX) (two rounds of 2 min 30 s, 1750 RPM). Homogenates (240 μL) were manually transferred to new Nunc deepwell 96 well plates (Thermo Scientific). Ready-lyse lysozyme solution (Epicentre, Madison, WI, USA) was manually added to each well to a final concentration of 50 U/μL. Plates were vortexed for 10 s and incubated at room temperature for 30 min. After a flash centrifugation, proteinase-K (EMD Millipore, Billerica, MA, USA) and SDS were added to each well to final concentrations of 0.5 mg/mL (20 mg/mL stock in 50 mM Tris-Cl pH 8, 3 mM $CaCl_2$, 50% glycerol, filter sterilized through 0.2 μm pore filter) and 1%, respectively. Plates were vortexed for 10 s and incubated in a 55°C water bath for 4 h. After a flash centrifugation, an equal volume of 24:1 chloroform:isoamyl alcohol was manually added to each well and pipetted to mix. Plates were centrifuged at 6600xg for 15 min at 4°C. The top aqueous layer (350 μL) was removed and added to new deepwell plates with 500 μL 100% isopropanol

per well. Plates were inverted 50 times to mix and incubated at -20°C for 1 h. After centrifugation at 6600xg for 15 min at 4°C, isopropanol was removed, and DNA pellets were washed with 500 μL 70% ethanol. Pellets were air dried in a chemical hood with a 96 well blower and resuspended in TE (100 μL, 10 mM Tris-Cl, 1 mM EDTA) by shaking for 4 min. After incubation on ice for 5 min, plates were centrifuged at 6600xg for 12 min at 4°C. DNA supernatants were removed from impurities in the pellets and added to new Nunc 0.5 mL 96 well both for storage (no dilution, kept at -20°C) and PCR amplification (10X dilution in TE, kept at 4°C).

4.3.5 16S rRNA gene amplification

The V5, V6, and V7 regions of the 16S rRNA gene were amplified from each samples using the 799F (5'-AAC MGG ATT AGA TAC CCK G-3') and 1193R (5'ACG TCA TCC CCA CCT TCC-3') primers. These primers minimize amplification from chloroplast DNA and allow for size selection of bacterial amplicons (~450 bp) from mitochondrial amplicons (~900 bp) (Bodenhausen *et al.*, 2013). Primers were designed as in Kozich *et al.* (2013) and also contained Illumina MiSeq adapters, and custom pads, linkers, and barcode sequences (Appendix B). All following pipetting steps were completed using custom liquid handling robotics scripts on the Freedom Evo 200 (Tecan), unless indicated differently. Each PCR was completed in triplicate. The total PCR volume was 25 μL, which contained 1 μL of 10X diluted DNA template, 0.2 μM of each primer, 1X 5PRIME HotMasterMix(5PRIME, Gaithersburg, MD, USA), and 0.8X SBT-PAR additive (5X stock: 750 mM sucrose, 2 mg/mL BSA, 1% Tween-20, 8.5 mM Tris-Cl pH 7.5) (Samarakoon *et al.*, 2013). PCR amplification consisted of an initial denaturation step of 2 minutes at 94°C, followed by 35 cycles of denaturation at 94°C for 30 s, annealing at 54.3°C for 40 s, elongation at 68°C for 40 s, followed by a final elongation of 7 minutes at 68°C.

Replicate reactions were pooled and amplicons purified with an equal volume of Axygen AxyPrep Mag PCR Clean-Up bead solution (Corning, Tewksbury, MA, USA) using the

manufacturer's standard protocol. Amplicon concentrations were quantified by fluorimetry (QUANT-iT PicoGreen dsDNA Assay Kit, Life Technologies, Carlsbad, CA, USA), and 30 ng or a maximum of 30 μL per sample were pooled for 576 samples per sequencing run. The length distribution and purity of the final pools were visualized with agarose gels. Primer dimers and mitochondrial amplicons were removed by first concentrating each amplicon pool 20X (Savant SPD121P SpeedVac Concentrator, Thermo Scientific), followed by BluePippin (Sage Science, Beverly, MA, USA) purification for 300-700 bp using the manufacturer's standard protocol.

The final DNA concentrations were determined, and the amplicon pools were sequenced using the Illumina MiSeq platform and MiSeq V2 Reagent Kits (Illumina, San Diego, CA, USA). Runs consisted of paired-end 250 bp reads (MiSeq Control Software v2.5.0.5) MiSeq Reporter v2.5.1.3 demultiplexed samples based on dual indices.

4.3.6 Sequence processing

Forward raw sequencing reads that aligned to PhiX using Bowtie 2 v2.1.0 (Langmead & Salzberg, 2012) were removed. Primer and adapter sequences were trimmed off reads with cutadapt v1.8.3 (Martin, 2011), prior to quality filtering with QIIME v1.9.1 split_libraries_fastq.py with -q 19 and default parameters (Caporaso *et al.*, 2010b). OTUs were clustered using QIIME pick_open_reference_otus.py (Rideout *et al.*, 2014) with the SILVA 119 97% reference sequences (http://www.arb-silva.de/download/archive/qiime/) for initial reference OTU clustering at 97% similarity. Reads that failed to align to references were clustered *de novo* at 97% similarity using uclust (Edgar, 2010). Representative OTU sequences were aligned with PyNAST Caporaso *et al.* (2010a) with the Greengenes 13_5 85% OTU alignment as a template (McDonald *et al.*, 2012). A gap-filtered alignment was used to construct a phylogenetic tree with FastTree (Price *et al.*, 2009). OTU taxonomy was assigned with uclust against the SILVA 119 97% reference taxonomy (Edgar, 2010). A final OTU table was created with unaligned and plastid aligned OTUs excluded. This table

was further parsed for OTUs that were detected more than 10 times over all samples. To account for differences in sequencing depth, each sample was rarefied to 1000 reads.

4.3.7 Statistical analysis

The R statistical environment (Team, 2015) was used for all analyses along with the ggplot2 package for plotting (Wickham, 2009). The vegan package was used for multivariate distance calculations and tests (Oksanen *et al.*, 2013). To compare multivariate distance metrics, the *vegdist* function was used to calculate pairwise Bray-Curtis distances. QIIME was used to calculate Weighted UniFrac distances (Caporaso *et al.*, 2010b; Lozupone & Knight, 2005). The *pco* function in the labdsv package (https://cran.r-project.org/web/packages/labdsv/index.html) was used for principal coordinates analysis (PCoA) of Bray-Curtis and Weighted UniFrac distances. For Hellinger distances, rarefied counts were Hellinger transformed (Legendre & Gallagher, 2001), and the *pca* function was used for principal components analysis (PCA).

To test for spatial and temporal factor effects on community structure, permutational multivariate analysis of variance (PERMANOVA, *adonis* function) was used to partition Bray-Curtis distance matrices (*vegdist* function) by treatments. Significance was assessed by 999 permutations.

The following model formula was used for PERMANOVA of all samples:

Bray-Curtis distances $\sim$ Tissue + Site + Tissue:Site + Ecotype + Growth stage nested in Year + MiSeq run

For PERMANOVA of site and year effects for samples from each tissue type, the following model formula was used:

Bray-Curtis distances $\sim$ Site + Year + Site:Year

For alpha-diversity calculations, the *diversity* function in the vegan package was used. Rarefaction curves were created with the *rarecurve* function. OTU overlaps were calculated

and plotted with the *venn.diagram* function in the package VennDiagram (https://cran.r-project.org/web/packages/VennDiagram/index.html).

4.4 Results

In this study, we characterized the spatial and temporal dynamics of bacterial communities associated with *A. thaliana*. Surface-sterilized seeds were planted at two locations in Southwest Michigan, USA, and above and below-ground plant samples were taken over the plant's entire life cycle (germinant, rosette, flowering, senescent). These plants were further separated in the laboratory by tissue type (roots, rosette leaves, stems, cauline leaves, flowers, and siliques, depending on growth stage). This field experiment was replicated for a second year at the same locations. In total, 1013 plant or soil pots were sampled, resulting in 2307 tissue and soil samples. To identify bacteria present in each sample, loosely associated microbes were removed, DNA was extracted, and the V5-V7 regions of the 16S rRNA gene were amplified and sequenced. After quality filtering, we had sufficient sequence data ($\geq$1000 reads) for 1180 samples. Sample sizes for flowers and cauline leaves were low and these samples were excluded from downstream analyses. Each sample was rarefied to 1000 reads prior to analysis.

4.4.1 Bacterial communities primarily differ by plant tissue

Given the different levels of temporal and spatial factors, we first tested the influence of each of these factors on bacterial community structure. To test for differences between mean structures (multivariate centroids), we applied a PERMANOVA model to the rarefied OTU table (see methods). The effect that accounted for the most variance was plant tissue ($F_4 = 66.17$, $R^2 = 0.16$, $P < 0.001$), followed by life stage nested in year ($F_9 = 11.02$, $R^2 = 0.062$, $P < 0.001$), interaction between tissue and site ($F_4 = 14.13$, $R^2 = 0.035$, $P < 0.001$), Site ($F_1 = 52.84$, $R^2 = 0.033$, $P < 0.001$), and finally MiSeq run ($F_4 = 5.87$, $R^2 = 0.015$, $P < 0.001$).

Ecotype did not have a significant effect on structure ($F_6 = 1.13$, $R^2 = 0.0042$, $P = 0.099$); therefore, we considered samples from different ecotypes to be equivalent for all downstream analyses. The effect of plant tissue on community structure is most evident via principal coordinates analysis of pairwise multivariate distances between all samples. We compared the results of PCoA with three different distance metrics and observed clear sample clustering by tissue with each metric (Figure 4.1). Bray-Curtis, Weighted UniFrac, and Hellinger distances are all weighted metrics that lessen the influence of rare OTUs. Bray-Curtis and Hellinger distances do not account for relatedness between OTUs, whereas Weighted UniFrac takes into account phylogenetic relatedness (Lozupone & Knight, 2005). The first and second principal coordinates for Weighted UniFrac accounted for more variance (44.4%) than coordinates for Bray-Curtis (20.8%) or Hellinger (18.2%); however, Bray-Curtis and Hellinger distances are advantageous for resolving differences between samples with highly phylogenetically similar OTUs. Hellinger distances also allow for extraction of OTU loadings on each principal coordinate.

4.4.2 Root communities diverge in structure from soil for vegetative plants and then become more soil-like after flowering and senescence

Based on significant PERMANOVA effects of life stage and year, we further explored temporal trends by taking the Bray-Curtis PCoA from Figure 4.1 and subsetting by site, year, and time point (Figure 4.2). We first observed that plant tissue effects on community structures were repeatable over time and geographic location (Figure 4.2). By comparison across time points, we observed an interesting successional pattern in the roots. There was an initial shift in root community structure away from soil for vegetative plants. Upon flowering and especially senescing, root communities became more similar to soil communities (Figure 4.2). These trends were also consistent between sites and years.

4.4.3 Geographic site differences in bacterial community structures vary by plant tissue

To further examine the significant PERMANOVA effects of plant tissue and site interactions, we isolated samples according to each plant part, calculated pairwise Bray-Curtis distances, and subjected the distance matrix to PCoA (Figure 4.3). In Figure 4.3, soil communities strongly cluster by site (PERMANOVA $R^2 = 0.18$, $P < 0.001$). This site differentiation decreased in root communities ($R^2 = 0.083$, $P < 0.001$) and was further diminished in rosette leaf ($R^2 = 0.063$, $P < 0.001$) and stem communities ($R^2 = 0.039$, $P < 0.001$). Interestingly, site differences then increased in silique communities ($R^2 = 0.13$, $P < 0.001$).

4.4.4 Plant tissues act as progressive sieves going from below to above-ground

Decreased site differences in bacterial community structures going from the soil to plant indicated that the plant may act as a sieve. If so, we would expect OTU richness to decrease going from soil to plant-associated communities, along with enrichment of specific OTUs. Indeed, we observed richness and evenness not only sharply decline from soil to root communities, but also, from below to aboveground plant tissues, and these trends were consistent across sites and years (Figure 4.4).

Interestingly, greater overlap in OTU membership did not necessarily correspond with these decreases in alpha-diversity. We observed that the percentage of unique OTUs per tissue did decrease going from below to aboveground tissues (Roots: 38%; Rosette leaves: 15%; Stems: 12%; Siliques: 8.8%), but the greatest overlap in OTUs for aboveground tissue was with roots and not other aboveground tissues (Figure 4.5). For rosette leaves, 75% of OTUs found in leaves were also found in roots. For stems, 74% of OTUs found in stems were also found in roots compared to just 65% overlap for leaves. For siliques, 74% of OTUs found in siliques were also found in roots, compared to 71% and 72% for leaves and stems,

respectively. We did not include comparisons with soil as the rarefaction level undersampled these samples (Figure 4.6). In terms of OTU identities, the roots appear to be the primary sieves that then seed OTUs to aboveground tissues.

Based on decreased richness and evenness going from below to aboveground tissues, we next examined whether certain genera and OTUs were enriched in each plant tissue. To find influential OTUs, we extracted the top 20 (in terms of absolute magnitude) OTU loadings for PC1 in the Hellinger PCoA in Figure 4.1 (Figure 4.7). Among these OTUs, the genera *Sphingomonas* and *Massilia* were enriched. In the case of the genus *Sphingomonas*, we observed increases in overall abundance of the genus from below to aboveground samples, though primarily at the Warren Woods site (Figure 4.8a). These increases were due to all of the most abundant *Sphingomonas* OTUs increasing in relative abundance, as indicated by little differences between the distribution of *Sphingomonas* OTUs by plant part, site, or year (Figure 4.8b). In contrast, the *Massilia* genus was at relatively equal abundance between plant parts, but specific *Massilia* OTUs are enriched in above-ground tissues (Figure 4.9).

4.5 Discussion

In this chapter, I characterized the spatial and temporal dynamics of bacterial communities associated with *A. thaliana*. By planting surface sterilized seeds at two sites over two growing seasons, I examined the effects of life stage and tissue type on the identity and relative abundances of bacterial colonizers. Unlike previous studies that focused on either leaf or root communities, I examined the communities in both of these tissues in addition to in stems and siliques. Plant tissue had the largest effect on bacterial community composition compared to the other spatial and temporal factors. By sampling over the plant's life cycle, I observed an initial shift in the bacterial communities associated with the roots away from communities in the soil. Upon flowering and senescence, root associated communities became more soil-like. By sampling at two geographic locations, I observed strong site effects on community composition in the soil which progressively decreased in the roots, leaves,

and stems, but reappeared in silique associated communities. The convergence in OTUs from different source communities in the soils indicated that these tissues acted as bacterial sieves. Alpha-diversity also progressively decreased going from the roots to leaves to stems to siliques, further supporting the roles of plate tissues as OTU sieves. From comparison of OTU presence/absence, aboveground tissues had a higher degree of overlap with roots than with other aboveground tissues. This overlap was surprising given the compositional differences and reduced diversity in aboveground tissues. All these results suggest that the roots primarily filter OTUs from the soil and that aboveground tissues differentially enrich for the presence and relative abundance of specific OTUs from this filtered OTU pool.

The finding that *A. thaliana* roots and leaves harbor different bacterial communities has been previously observed (Bodenhausen *et al.*, 2013), but our work has uncovered that the stems and siliques also contain unique communities separate from the leaves. To calculate the multivariate distances between samples, we used three different metrics: Bray-Curtis, Weighted UniFrac, and Hellinger. All three metrics are weighted, meaning that abundant OTUs influence the dissimilarity between two samples more than rare OTUs. With our low thresholds for OTU retention (>10 reads over all samples), we found a large number of OTUs, with the vast majority being rare. Weighted metrics are needed to avoid equal weighting of these rare and possibly artifactual OTUs. PCoA with Weighted UniFrac distances resulted in the largest variance explained by the first two principal coordinates compared to PCoA with Bray-Curtis or Hellinger distances. The increased explanatory power of UniFrac indicates that bacterial OTUs found in each tissue type are related and we can explain more compositional differences by considering OTU phylogeny (Lozupone & Knight, 2005). This finding corresponds with the progressive filtering of OTUs going from belowground to aboveground tissues. To delve into fine scale differences in OTU abundances by site and/or tissue, we used Bray-Curtis instead of Weighted UniFrac. As observed in Figure 4.9, closely related OTUs can have different abundance patterns which were better captured for sample differentiation by Bray-Curtis distances.

In contrast to previous studies that have used thoroughly defined extraction conditions for microbes associated with the rhizosphere, rhizoplane, and endophytic compartment of roots (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012), we chose a more high-throughput extraction procedure that could also be applied to additional plant tissues. With a surfactant buffer wash, we removed loosely associated microbes from all tissues. However, we cannot distinguish between closely associated microbes on the surface versus within each tissue.

Though our processing conditions were different, we confirmed the finding of the root as an OTU sieve, filtering microbes from the surrounding soil (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012). By examining other tissue types, we further observed progressive filtering going from the roots to the leaves, leaves to the stems, and stems to the siliques. Previous work with bacterial communities associated with different tissue of merlot grapevines identified higher degrees of overlap between the OTUs found in the soil and aboveground tissues than between aboveground tissues (Zarraonaindia *et al.*, 2015). In this study, we observed more extensive OTU membership overlap between communities associated with aboveground tissues and the roots than the soil, though we did not fully sample soil communities. OTU overlaps between aboveground tissues and the roots were higher than between aboveground tissues. These findings suggest a model in which the roots initially filter specific bacteria from surrounding soil. These OTUs are then distributed across tissues with each tissue enriching for specific but different OTUs. Whether plant tissues act as active or passive sieves remains to be seen. Our finding that site effects on community composition were more prominent in senescent tissue (siliques versus leaves), indicates that the metabolically active plant environment is selective. Selection may occur through plant defense responses, niche partitioning, or through bacterial interactions with each other.

4.6 Acknowledgments

Thanks to Dave Francis at the Michigan State Southwest Michigan Research and Extension Center for providing and prepping field plots. Thanks to Carlos Sahagun for plant-

ing prep and assistance in the field. Thanks to Timothy Morton, Benjamin Brachi, Talia Karasov, Manfred Ruddat, and Roderick Woolley for help with sampling. DNA extraction and 16S rRNA gene amplicon sequencing was developed in collaboration with Benjamin Brachi. Thanks to Alison Anastasio for experimental design advice.

4.7 Tables and figures

Accession	Ecotype ID	Collection site	Latitude/Longitude
BRR4	470	Watseka, IL, USA	40.8313, -87.735
LI-WP-041	546	Nissequogue, NY, USA	40.9076, -73.2089
L-R-10	1797	Union Pier, MI, USA	41.847, -86.67
MNF-Che-47	1942	Manistee National Forest, MI, USA	43.5251, -86.1843
Pent-7	2191	Pentwater, MI, USA	43.7623, -86.3929
PT1.85	8057	Hanna, IN, USA	41.3423, -86.7368
SLSP-69	2285	Silver Lake, MI, USA	43.665, -86.496

Table 4.1: *A. thaliana* natural accessions (ecotypes) used in this study. All accessions are nearly isogenic and belong to haplogroup-1 (HPG1) (Hagmann *et al.*, 2015).

Year	Date	Time point	Collection site	Plant stage	Sample types
1	October 12, 2012	T1	ME	NA	Soil
1	October 15, 2012	T1	WW	NA	Soil
1	October 26, 2012	T2	ME	Two leaf	Roots, Rosette leaves
1	October 29, 2012	T2	WW	Two leaf	Roots, Rosette leaves
1	November 29, 2012	T3	ME, WW	Four leaf	Roots, Rosette leaves
1	February 13, 2013	T4	ME	Six leaf	Roots, Rosette leaves
1	March 29, 2013	T5	ME, WW	Eight leaf	Roots, Rosette leaves
1	May 1, 2013	T6	ME, WW	Flowering	Roots, Rosette leaves, Stems, Siliques, Flowers
1	June 14, 2013	T7	ME, WW	Senescent	Roots, Stems, Siliques
2	October 28, 2013	T1	ME, WW	NA	Soil
2	December 4, 2013	T2	ME, WW	Two leaf	Soil, Roots, Rosette leaves
2	April 16, 2014	T3	ME, WW	Six leaf	Soil, Roots, Rosette leaves
2	May 15, 2014	T4	ME, WW	Flowering	Soil, Roots, Rosette leaves, Stems, Siliques, Flowers
2	July 3, 2014	T5	ME, WW	Senescent	Soil, Roots, Stems, Siliques

Table 4.2: Field sample summary

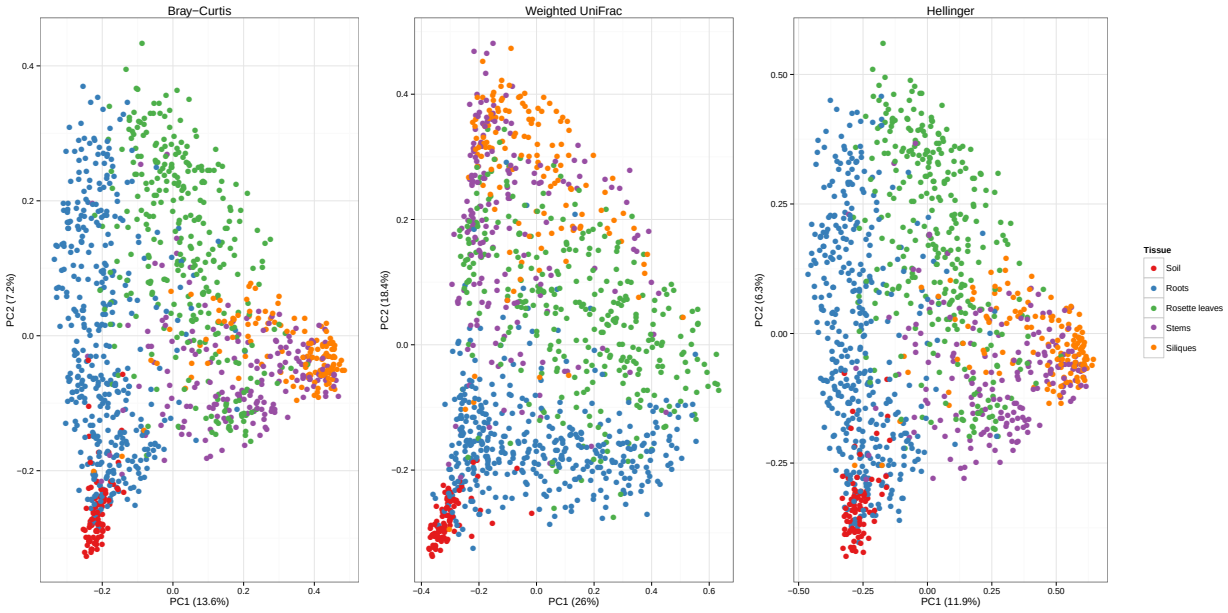


Figure 4.1: Bacterial communities primarily differ by plant tissue.

Pairwise Bray-Curtis, Weighted UniFrac, or Hellinger distances were calculated between all samples and subjected to principal coordinates analysis (PCoA). Samples are colored by plant part.

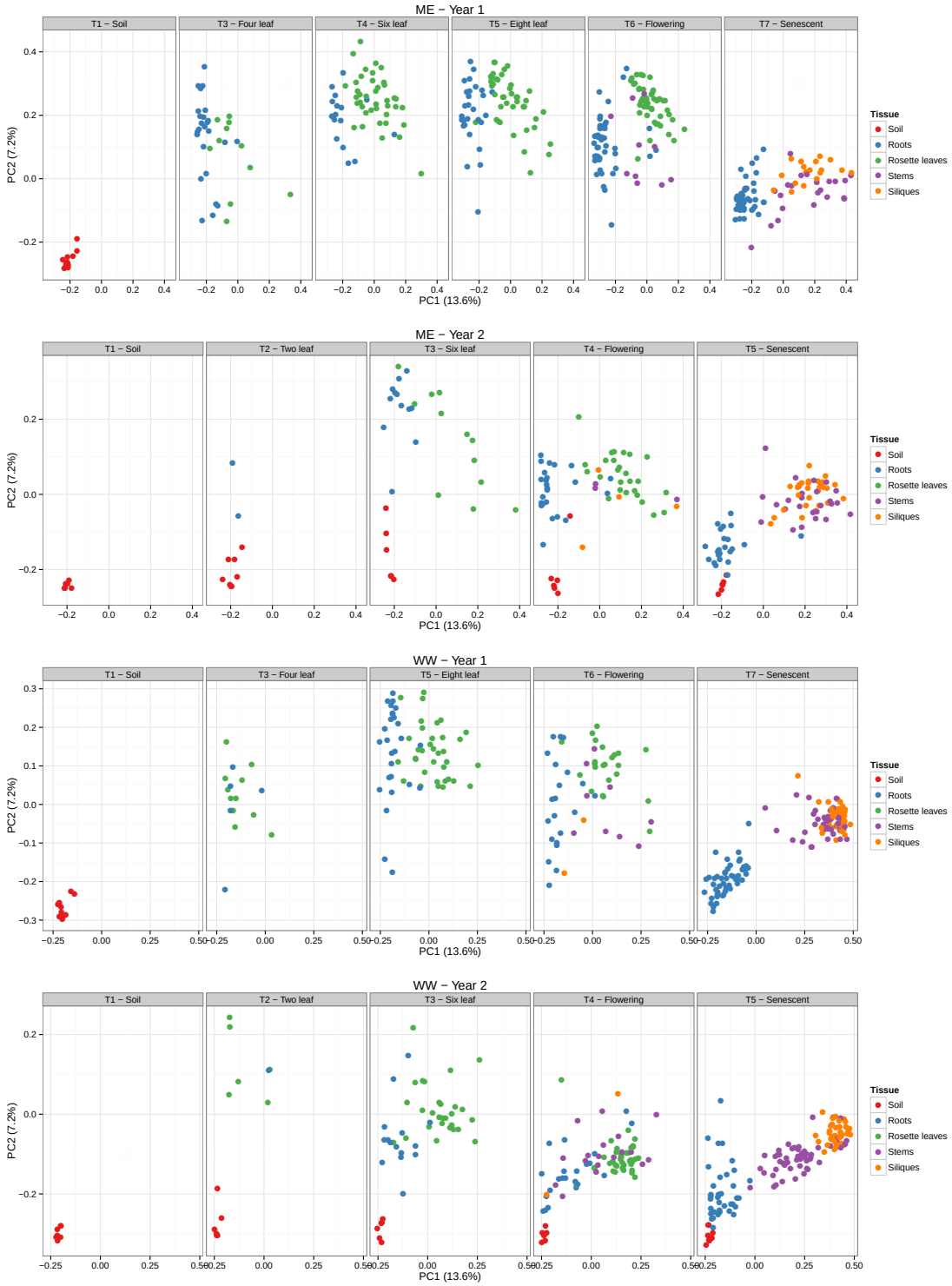


Figure 4.2: Tissue effects are repeatable over sites, years, and time points.

Pairwise Bray-Curtis distances were calculated between all samples and subjected to principal coordinates analysis (PCoA). Samples are colored by plant part and facetted by site, year, and time point.

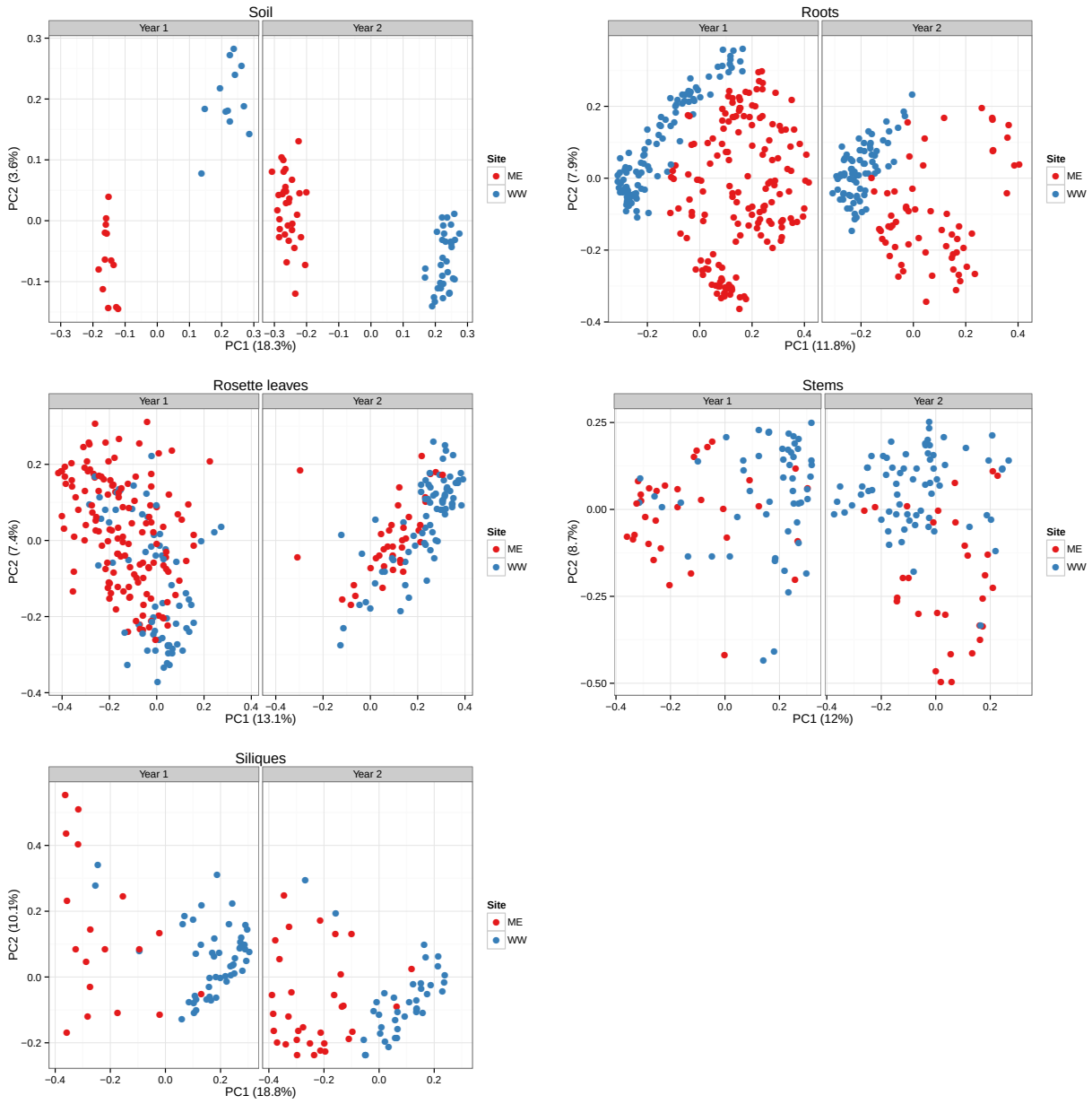
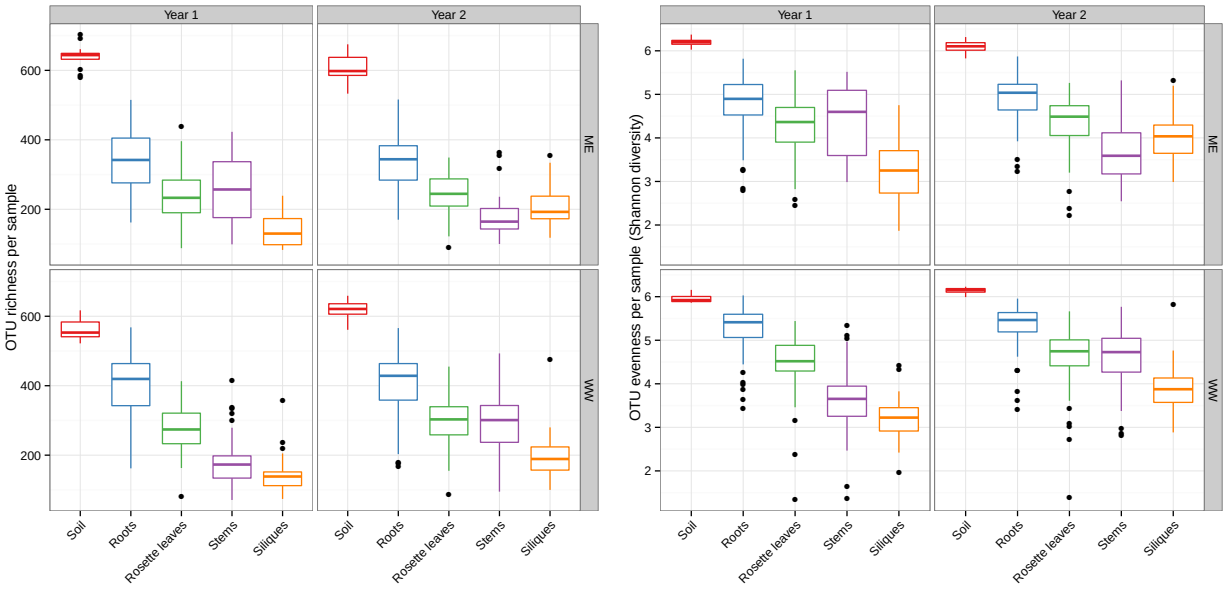


Figure 4.3: Site effects on community structure by plant tissue

For samples from each plant tissue, pairwise Bray-Curtis distances were calculated and subjected to principal coordinates analysis (PCoA). Samples are colored by site and facetted by year for each tissue.



(a) Richness

(b) Evenness

Figure 4.4: OTU richness and evenness decrease from below to aboveground plant tissues.

For each plant part by site and year, the number of OTUs (richness) and Shannon diversity (evenness) was calculated for all corresponding samples with boxplots shown.

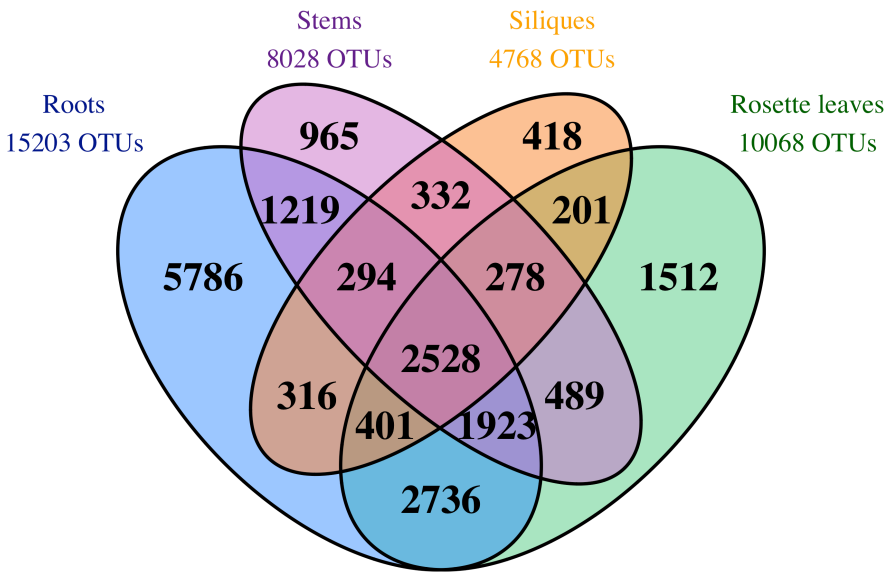


Figure 4.5: OTU overlap venn diagram

OTU overlaps were calculated by plant tissue. Overlap was based on presence/absence and not weighted by abundances.

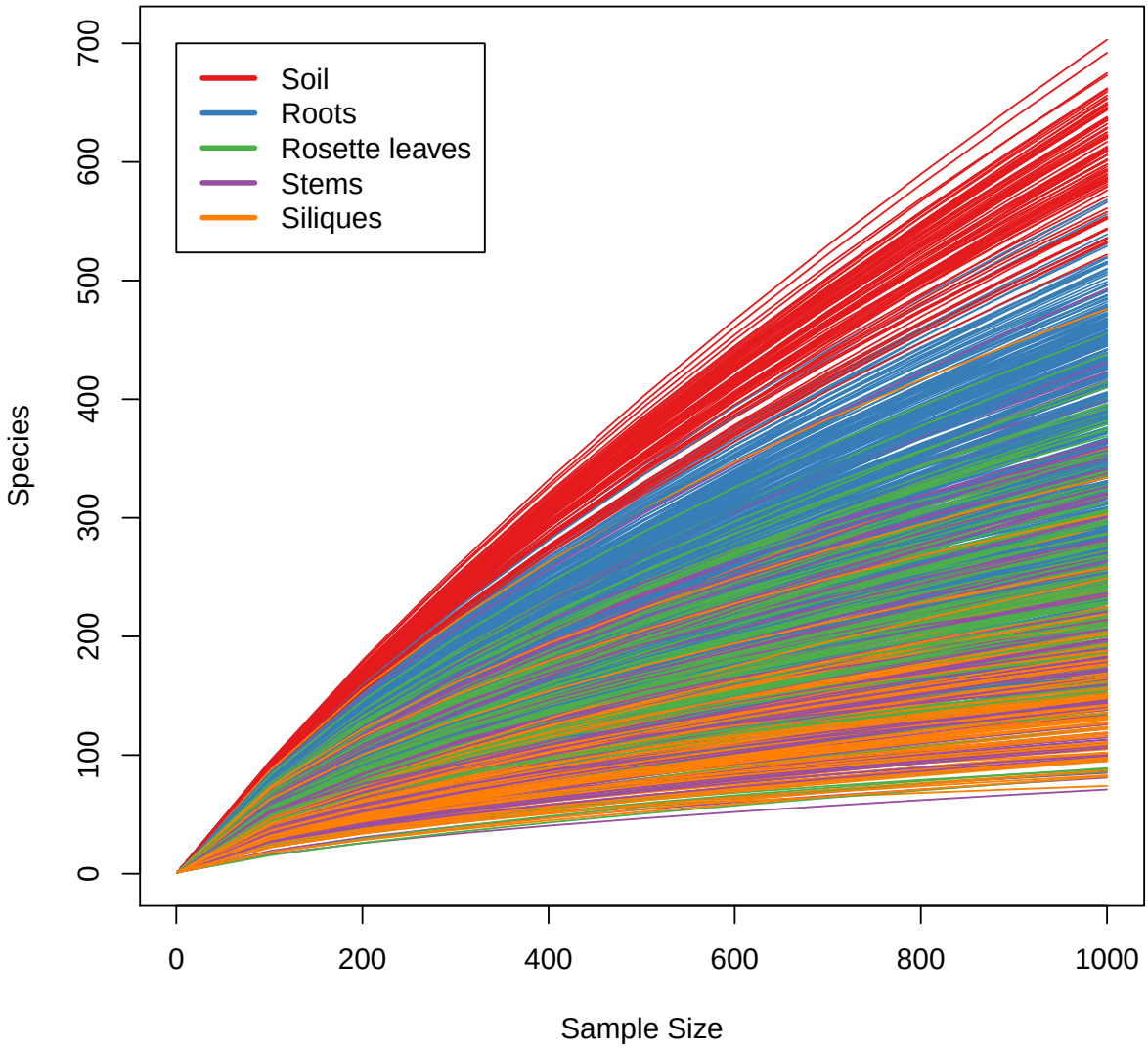


Figure 4.6: Rarefaction curves for OTUs

For a step size of 100 sequencing reads, rarefied samples (1000 reads) were further rarefied. Each line corresponds to a separate sample. The number of 97% OTUs (Species) is shown for each sample size and colored by the tissue type.

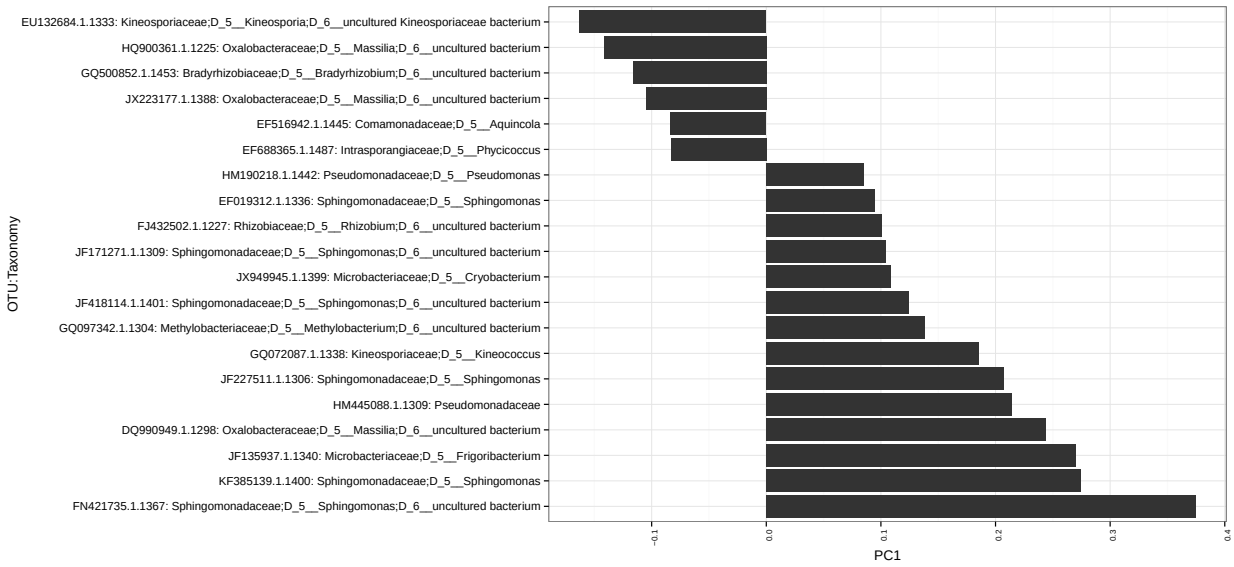
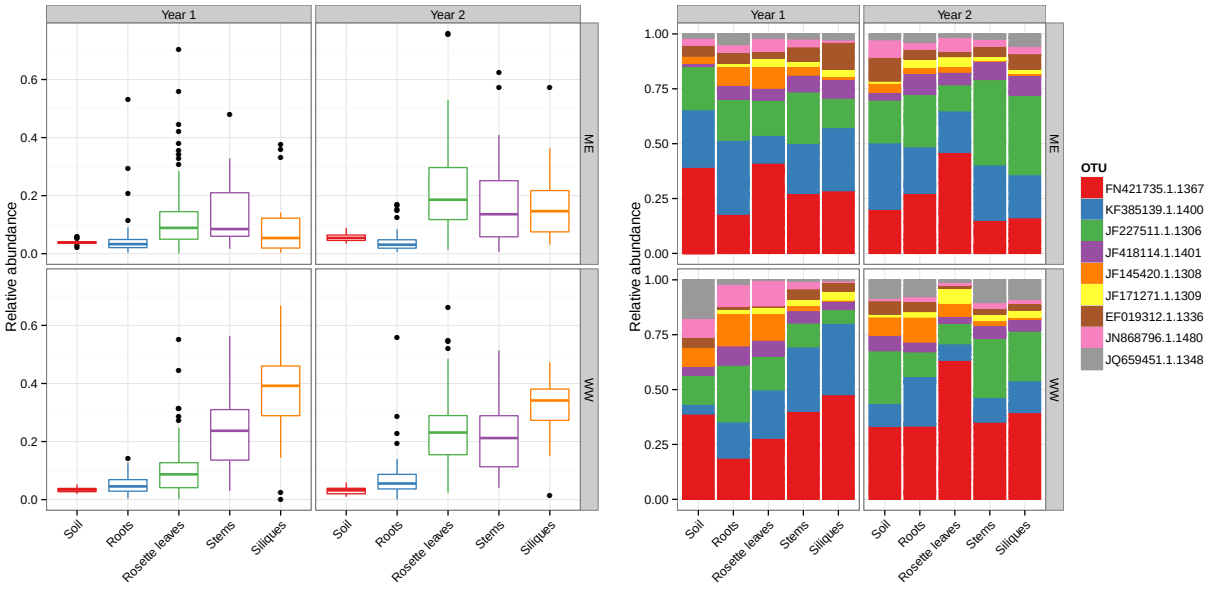


Figure 4.7: Top OTU loadings for Hellinger PC1

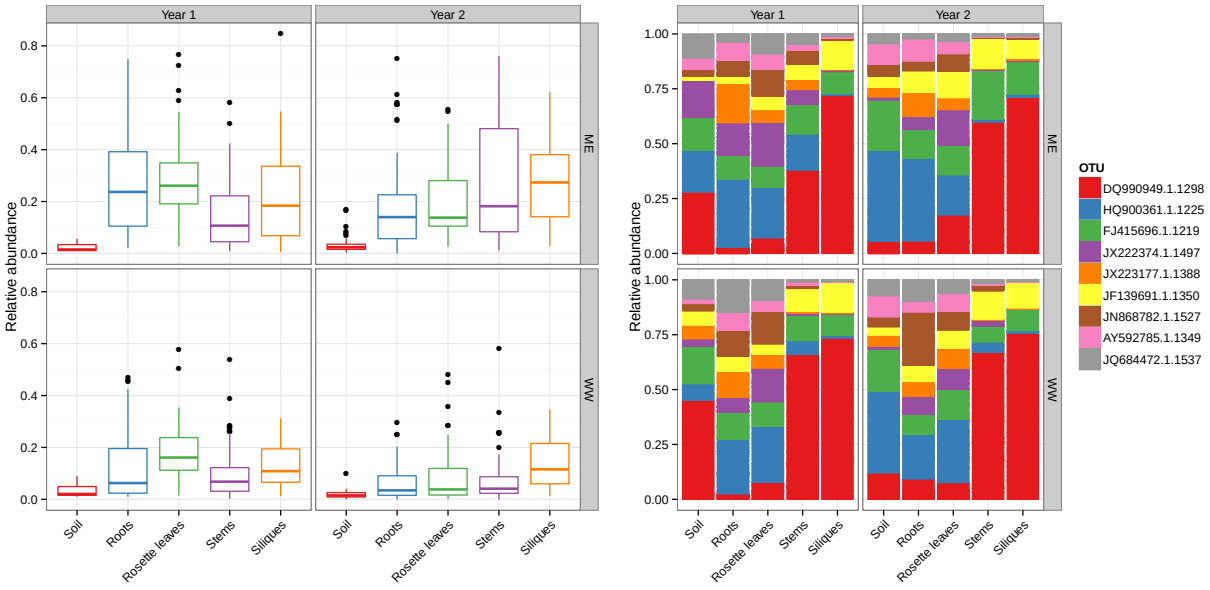
For the Hellinger PCoA in Figure 4.1, the top 20 OTU loadings for PC1 were extracted (as measured by absolute magnitude).



(a) (b)

Figure 4.8: Aboveground plant tissues enrich for most abundant *Sphingomonas* OTUs

(a) Relative abundances for all OTUs classified as the genus *Sphingomonas* were summed for each sample and boxplotted by plant tissue. (b) The relative abundances, within the *Sphingomonas* genus only, for the nine most abundant OTUs.



(a) (b)

Figure 4.9: Aboveground plant tissues enrich for specific *Massilia* OTUs

(a) Relative abundances for all OTUs classified as the genus *Massilia* were summed for each sample and boxplotted by plant tissue. (b) The relative abundances, within the *Massilia* genus only, for the nine most abundant OTUs.

CHAPTER 5

DISCUSSION

Plant-associated bacterial communities are highly diverse. Each plant tissue type harbors hundreds of OTUs (Figure 4.4). I have been fascinated by this coexistence which is seemingly contradictory to the principle of competitive exclusion. To uncover the ecological mechanisms that maintain diversity, I took a multi-faceted approach involving the analysis of bacterial competition first outside the plant in liquid and solid media and then inside gnotobiotic plants. The complexity of environment altered community composition and the predictability of succession was surprisingly low for bacterial communities in liquid media compared to solid media. To reveal environmental and host factors that structure communities, I sampled natural communities and analyzed how these communities change over time and space. Sampling of different plant tissues and comparison of OTU overlaps indicated that tissues acted as bacterial sieves. Roots first filtered OTUs from the soil that then became enriched in aboveground tissues.

Natural communities displayed a remarkable level of richness, with few highly abundant OTUs and a long tail of rare OTUs. This distribution is common to almost every environment sampled thus far (Sogin *et al.*, 2006). The role of these rare bacteria has been debated. These OTUs could act as a seed bank, blooming to high abundance given the right set of conditions (Shade *et al.*, 2014). Interestingly, I rarely encountered a sample with an OTU with >25% abundance in my field samples. In this list of dominant OTUs, were known *A. thaliana* pathogens such as *P. viridiflava* and *P. syringae*, as well as beneficial genera such as *Sphingomonas* that are capable of preventing *P. syringae* mediated disease (Vogel *et al.*, 2012).

To disentangle the bacterial interspecific interactions that promote this coexistence, I took advantage of cultured species originally isolated from natural *A. thaliana*. Though my controlled community experiments involved some of the most abundant leaf endophytic species, I never recapitulated the richness and evenness observed in the field. Even with

eleven species inoculated into media or gnotobiotic plants, only a subset (3-5 species) were detected at later time points. I observed the lowest diversity levels in solid media followed by liquid media then plant. The collapse in diversity *in vitro* is likely due to strong competition for nutrients. This result may not have been surprising given that Foster & Bell (2012) noted that in pairwise species mixtures in aquatic microcosms, competition was the predominant interaction type.

In my *in vitro* experiments, I confirmed each isolate's capacity to grow in the rich SOC media, but growth rates were not sufficient to explain the lack of coexistence. Both liquid and solid media communities displayed different species' abundance hierarchies, that for liquid media, did not follow decreasing order of growth rates (Figure 3.8 and Figure 3.9). Interestingly, I observed a different hierarchy from the pairwise competitions in liquid KB media (Table 2.3) than that observed in SOC liquid media. This discrepancy is likely due to lack of iron in KB media. KB media was primarily developed as a medium to culture *Pseudomonas* species, and these colonies are commonly fluorescent due to production of siderophores, known as pyoverdines, in response to low iron concentrations (King *et al.*, 1954). *Pseudomonas* species produce specifically modified pyoverdines to scavenger iron from the surrounding environment and prevent its uptake into unrelated organisms (Andrews *et al.*, 2003). Interestingly, when the top competitor in liquid KB media and solid SOC media (*P. simiae*) is drip inoculated on top of a lawn of another isolate on a KB plate, an inhibition zone forms around *P. simiae* (data not shown). The mechanism of inhibition is bacteriostatic as streaking bacteria from the inhibition zone onto a fresh plate results in uninhibited growth (data not shown). Adding iron to KB reduces the size of the inhibition zone (data not shown). This hypothesis of pyoverdine mediated competition that results in the observed abundance hierarchies will need to be confirmed with pyoverdine knockout mutants.

Species mixtures inoculated into gnotobiotic plants also resulted in diversity collapse, though not to the extent exhibited *in vitro*. The variety of spatial and nutrient niches *in planta* likely promotes coexistence. In Chapter 2, I developed FISH and spectral imaging

methods to properly interrogate the spatial distributions of species in mixed isolate communities. Future optimization of these techniques for visualization of multiple bacterial species within leaves will fully test for the presence of spatial niches. In FISH and bandpass imaging of epiphytic bacteria, non-random spatial distributions were observed (Remus-Emsermann *et al.*, 2014), indicating species predilections for certain areas of the leaf surface. The lack of diversity in gnotobiotic compared to natural plant-associated communities may be due to phenotypic differences caused by sterile conditions. Gnotobiotic plant leaves do not have a thick cuticle layer and cannot withstand concentrated bacterial inoculations that are tolerated by greenhouse or field grown plants (data not shown). In a similar synthetic community experiment, Bodenhausen *et al.* (2014) observed coexistence of seven isolates inoculated into gnotobiotic plant leaves, though commonly detected pathogens were not included in this mix. Bodenhausen *et al.* (2014) also allowed communities to form over the course of four weeks, in contrast to eight days in my experiments. Perhaps the communities I observed did not reach steady states.

The sudden introduction of thousands of bacteria onto the surface of leaves is also not a natural inoculation route. Natural plants encounter microbes from the soil and air. To observe how air inoculation affects community succession on the surface of leaves, Maignien *et al.* (2014) grew surface sterilized seeds on sterilized soil and monitored epiphytic and air bacterial communities over the life cycle of the plant. Interestingly, epiphytic communities initially resembled air communities in terms of OTU membership, but plant communities became more plant specific over time (Maignien *et al.*, 2014). In terms of relative abundances, epiphytic community succession followed two separate paths depending on spatial location in the growth chamber (Maignien *et al.*, 2014). This successional stochasticity hints at the roles of priority effects in structuring leaf-associated communities. Reduced diversity in my synthetic community experiments may be due to priority effects. My inoculations contained 10^3 CFU per plant, which for an eleven species mixture equals $\sim$90 CFU per species. Slight differences in initial species abundances can become amplified by priority effects, resulting

in overrepresentation of certain strains. Though microbes inoculated onto the aboveground surface of the plant have the ability to enter and colonize the plant, recent studies point to the soil as the main reservoir for plant-associated OTUs (Copeland *et al.*, 2015; Zarraonaindia *et al.*, 2015). My field experiments also indicate the soil as the source of microbes that are first filtered by the roots and then distributed to aboveground tissues.

In regards to ecological theory, my research has provided novel insights into the maintenance of diversity in plant-associated bacterial communities. By approaching bacterial coexistence from multiple angles, including studies of both *in vitro* and natural communities, I observed evidence for both niche and neutral theories. Under controlled conditions, I observed clear competitive hierarchical patterns in species' abundances. These results indicate that species are non-equivalent competitors in media and gnotobiotic plants, thus arguing against neutral theory. However, there were large multivariate dissimilarities between replicate communities which indicated a role for stochastic forces. Demographic stochasticity in initial inoculums may have contributed to replicate discordance. From a niche theory perspective, one would expect that fastest growing isolate in liquid media would be the most robust inhabitant of that niche. However, some of the fastest growing isolates were the worst competitors. This prediction assumes equal distribution of nutrients to each isolate, which is likely not the case. Nutrient acquisition factors such as siderophores and production of other bacteriostatic or bactericidal factors in response to the presence of other species, makes the outcomes of multi-species competitions difficult to predict. These competition factors create evolving conditions and niches. Under natural conditions, there are already niches due to biotic and abiotic conditions. Evolving competitive niches could provide an incredibly large number of different conditions that could support a large number of species. Perhaps, these types of niches account for the observed species distributions in which there are few highly abundant species and a long tail of rare species. These abundant species can compete and create the right conditions for a rare species to bloom and continue the cycle of boom-and-bust.

From my field experiments, I observed species predilections for specific plant tissues, and these species were consistently found in high abundances. The taxa that I observed associated with the roots and leaves, were the same taxa that have been observed in previous studies conducted at different times and at different locations around the world (Delmotte *et al.*, 2009; Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Bodenhausen *et al.*, 2013). These consistent findings indicate strong structural forces for plant-associated bacterial communities. Evolving competitive niches likely play a role, but the added trophic level of the host adds feedback mechanisms. I observed evidence for host environment feedback on community structure by comparison of site effects by plant tissue and life stage. Interestingly, site effects on community composition were strongest in the soil and senescent tissues. The magnitude of these effects was dampened in vegetative tissues in which the same or similar OTUs were filtered from different source communities (Figure 4.3). Metabolically active plant tissue creates a selective environment for specific bacterial species. Future experiments are needed to decipher whether filtering is an active plant process or due to passive effects of that environment. From a neutral theory perspective, this filtering process could be explained by dispersal limitation. However, the highly abundant taxa found *in planta* would need to have comparably high intrinsic dispersal rates compared to other metacommunity members. The consistency of detected *in planta* taxa and the OTU filtering that I observed in Chapter 4 strongly advocate for niche over neutral theory in natural bacterial communities.

Overall, I have strived to answer the perplexing questions of how so many bacterial species can coexist within a plant. Furthermore, what is our ability to predict coexistence and community succession. These complex questions have complex answers. Based on the highly variable replicate community compositions that I observed in Chapter 3, we do not have the ability to predict succession given a known starting mixture of bacterial species. By averaging over replicates, I observed species abundance patterns, but I could not predict the behavior of each replicate to a high degree of certainty. Contrary to Chapter 3, in Chapter 4, I found that natural succession of bacterial communities associated with various plant tissues

was repeatable over two years at two sites. This discrepancy is likely due to the types, timing, and number of selective forces in my controlled versus field environments. The controlled environments of media and gnotobiotic plants are nutrient rich, and co-inoculation of multiple species results in a free-for-all while other environmental conditions such as temperature and humidity are held constant. The selective forces in these experiments were primarily due to bacterial competition during nutrient acquisition. Gnotobiotic plants imposed additional spatial and host immune forces, but this environment was also nutrient rich.

In comparison, I observed natural plant colonization and succession to follow repeatable paths. Similar to the controlled conditions, a sprouting plant seed is immediately introduced to a large and diverse mixture of bacterial species in the soil; however, the initial *in planta* environment is nutrient poor. Based on my observation that vegetative root associated communities are diverged from the surrounding soil communities, the metabolically active plant tissue environment imposes a strong selection on community composition. I hypothesize that this selective mechanism involves plant production of specific metabolites and nutrients that support the competitively dominant species for those factors. The plant likely benefits from this relationship by protection from pathogenic species and growth promotion. Preliminary evidence for this hypothesis comes from the *in vitro* and *in planta* hierarchies in Chapter 3. The strongest competitors were commensal or growth promoting isolates (*P. simiae*, *P. fluorescens*, *C. flaccumfaciens*, *Microbacterium sp.*) not known pathogenic species (*P. syringae*, *P. viridiflava*, *X. campestris*). Add in the selective forces of plant defense factors and fluctuating abiotic conditions such as temperature and humidity, and community succession becomes much more predictable.

Therefore, if we wish to manipulate bacterial communities in bioreactors or *in planta*, we must impose strong selective forces on community composition and function. For plant-associated microbiomes, the initial soil community is crucial as the roots filter OTUs from this pool that become enriched in aboveground tissues. Based on unpredictability of succession *in vitro*, I also believe probiotic treatments will be largely unpredictable and unsuccessful.

Knowing the initial community composition before treatment is insufficient to understand the near infinite multi-trophic interactions that must occur for establishment of the probiotic strain. Furthermore, there is immense strain variation between isolates at the species and even sub 97% OTU identity (Eren *et al.*, 2013). For instance, in Chapter 4, highly similar OTUs from the genus *Massilia* show different tissue predilections. On the other hand, prebiotic treatments have much greater potential for microbiome manipulation. My findings indicate that bacterial competition can be unpredictable under non-selective conditions, but addition of prebiotics such as specific nutrients in the soil has the potential to select and promote the growth of desired taxa. This strategy also takes advantage of native strains that are likely to be adapted to the local environment. If these strains are not present, a combination of pro and prebiotic treatments may be effective.

These recommendations derive from observation of plant-associated bacterial communities both under controlled and natural conditions. By taking this multi-faceted approach, I have uncovered ecological forces that structure these communities and influence successional predictability. I have put forth new hypotheses in this discussion based on this groundwork and I am excited to see the field of plant microbial ecology continue to grow and answer these questions.

REFERENCES

Alonso, D., Etienne, R. & McKane, A. (2006). The merits of neutral theory. *Trends in Ecology & Evolution*, 21, 451–457.

Amann, R.I., Binder, B.J., Olson, R.J., Chisholm, S.W., Devereux, R. & Stahl, D.A. (1990). Combination of 16S rRNA-targeted oligonucleotide probes with flow cytometry for analyzing mixed microbial populations. *Applied and Environmental Microbiology*, 56, 1919–1925.

Andrews, S.C., Robinson, A.K. & Rodríguez-Quiñones, F. (2003). Bacterial iron homeostasis. *FEMS Microbiology Reviews*, 27, 215–237.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. & Zagnitko, O. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, 9, 75.

Barberán, A., Casamayor, E.O. & Fierer, N. (2014). The microbial contribution to macroecology. *Frontiers in microbiology*, 5, 203.

Barrett, L.G., Bell, T., Dwyer, G. & Bergelson, J. (2011). Cheating, trade-offs and the evolution of aggressiveness in a natural pathogen population. *Ecology Letters*, 14, 1149–1157.

Bodenhausen, N., Bortfeld-Miller, M., Ackermann, M. & Vorholt, J.A. (2014). A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. *PLoS Genetics*, 10, e1004283.

Bodenhausen, N., Horton, M.W. & Bergelson, J. (2013). Bacterial Communities Associated with the Leaves and the Roots of Arabidopsis thaliana. *PLoS ONE*, 8, e56329.

Bolger, A.M., Lohse, M. & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120.

Bordenstein, S.R. & Theis, K.R. (2015). Host Biology in Light of the Microbiome: Ten Principles of Holobionts and Hologenomes. *PLoS Biology*, 13, e1002226.

Brynildsrud, O., Snipen, L.G. & Bohlin, J. (2015). CNOGpro: Detection and quantification of CNVs in prokaryotic whole-genome sequencing data. *Bioinformatics*, p. btv070.

Buell, C.R. (2002). Interactions Between Xanthomonas Species and Arabidopsis thaliana. *The Arabidopsis Book*, 1, e0031.

Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., Peplies, J., Gloeckner, F.O., Amann, R., Eickhorst, T. & Schulze-Lefert, P. (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature*, 488, 91–95.

Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E. & Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants. *Annual review of plant biology*, 64, 807–838.

Caporaso, J.G., Bittinger, K., Bushman, F.D., DeSantis, T.Z., Andersen, G.L. & Knight, R. (2010a). PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26, 266–267.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A., McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J., Yatsunenko, T., Zaneveld, J. & Knight, R. (2010b). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7, 335–336.

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

Celiker, H. & Gore, J. (2014). Clustering in community structure across replicate ecosystems following a long-term bacterial evolution experiment. *Nature Communications*, 5, 4643.

Chaparro, J.M., Badri, D.V., Bakker, M.G., Sugiyama, A., Manter, D.K. & Vivanco, J.M. (2013). Root exudation of phytochemicals in Arabidopsis follows specific patterns that are developmentally programmed and correlate with soil microbial functions. *PLoS ONE*, 8, e55731.

Chaparro, J.M., Badri, D.V. & Vivanco, J.M. (2014). Rhizosphere microbiome assemblage is affected by plant development. *The ISME Journal*, 8, 790–803.

Chave, J. (2004). Neutral theory and community ecology. *Ecology Letters*, 7, 241–253.

Chave, J. & Leigh, E.G. (2002). A spatially explicit neutral model of beta-diversity in tropical forests. *Theoretical population biology*, 62, 153–168.

Chesson, P. (2000). Mechanisms of maintenance of species diversity. *Annual review of Ecology and Systematics*.

Choi, K.H. & Schweizer, H.P. (2006). mini-Tn7 insertion in bacteria with single attTn7 sites: example Pseudomonas aeruginosa. *Nature Protocols*.

Cocking, E.C. (2003). Endophytic colonization of plant roots by nitrogen-fixing bacteria. *Plant and soil*.

Copeland, J.K., Yuan, L., Layeghifard, M., Wang, P.W. & Guttman, D.S. (2015). Seasonal Community Succession of the Phyllosphere Microbiome. *Molecular Plant-Microbe Interactions*, 28, 274–285.

Delmotte, N., Knief, C., Chaffron, S., Innerebner, G., Roschitzki, B., Schlapbach, R., von Mering, C. & Vorholt, J.A. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proceedings of the National Academy of Sciences*, 106, 16428–16433.

Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C. & Fitter, A.H. (2009). Relative roles of niche and neutral processes in structuring a soil microbial community. *The ISME Journal*, 4, 337–345.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26, 2460–2461.

Eiler, A., Heinrich, F. & Bertilsson, S. (2012). Coherent dynamics and association networks among lake bacterioplankton taxa. *The ISME Journal*, 6, 330–342.

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

Falk, M.W., Song, K.G., Matiasek, M.G. & Wuertz, S. (2009). Microbial community dynamics in replicate membrane bioreactors–natural reproducible fluctuations. *Water research*, 43, 842–852.

Falkowski, P.G., Fenchel, T. & Delong, E.F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320, 1034–1039.

Fierer, N., Strickland, M.S., Liptzin, D., Bradford, M.A. & Cleveland, C.C. (2009). Global patterns in belowground communities. *Ecology Letters*, 12, 1238–1249.

Finkel, O.M., Burch, A.Y., Lindow, S.E., Post, A.F. & Belkin, S. (2011). Geographical Location Determines the Population Structure in Phyllosphere Microbial Communities of a Salt-Excreting Desert Tree. *Applied and Environmental Microbiology*, 77, 7647–7655.

Flores, G.E., Caporaso, J.G., Henley, J.B., Rideout, J.R., Domogala, D., Chase, J., Leff, J.W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., Dunn, R.R. & Fierer, N. (2014). Temporal variability is a personalized feature of the human microbiome. *Genome Biology*, 15, 531.

Foster, K.R. & Bell, T. (2012). Competition, not cooperation, dominates interactions among culturable microbial species. *Current biology : CB*, 22, 1845–1850.

Fox, J.L. (2014). Microbes unite Novozymes and Monsanto. *Nature Biotechnology*, 32, 211–211.

Fuhrman, J.A., Hewson, I., Schwalbach, M.S., Steele, J.A., Brown, M.V. & Naeem, S. (2006). Annually reoccurring bacterial communities are predictable from ocean conditions. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 13104–13109.

Gause, G.F. (1934). Experimental Analysis of Vito Volterra's Mathematical Theory of the Struggle for Existence. *Science*, 79, 16–17.

Gilbert, J.A., Jansson, J.K. & Knight, R. (2014). The Earth Microbiome project: successes and aspirations. *BMC biology*, 12, 69.

Gilbert, J.A. & Neufeld, J.D. (2014). Life in a World without Microbes. *PLoS Biology*, 12, e1002020.

Gilbert, J.A., Steele, J.A., Caporaso, J.G., Steinbrück, L., Reeder, J., Temperton, B., Huse, S., McHardy, A.C., Knight, R., Joint, I., Somerfield, P., Fuhrman, J.A. & Field, D. (2011). Defining seasonal marine microbial community dynamics. *The ISME Journal*, pp. 1–11.

Godfrey, S.A.C., Mansfield, J.W., Corry, D.S., Lovell, H.C., Jackson, R.W. & Arnold, D.L. (2010). Confocal Imaging of Pseudomonas syringae pv. phaseolicola Colony Development in Bean Reveals Reduced Multiplication of Strains Containing the Genomic Island PPHGI-1. *MPMI*, 23, 1294–1302.

Greenblum, S., Carr, R. & Borenstein, E. (2015). Extensive Strain-Level Copy-Number Variation across Human Gut Microbiome Species. *Cell*, 160, 583–594.

Grice, E.A. & Segre, J.A. (2012). The Human Microbiome: Our Second Genome. *Annual review of genomics and human genetics*, 13, 151–170.

Van der Gucht, K., Cottenie, K., Muylaert, K., Vloemans, N., Cousin, S., Declerck, S., Jeppesen, E., Conde-Porcuna, J.M., Schwenk, K., Zwart, G., Degans, H., Vyverman, W. & De Meester, L. (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*, 104, 20404–20409.

Haas, D. & Défago, G. (2005). Biological control of soil-borne pathogens by fluorescent pseudomonads. *Nature Publishing Group*, 3, 307–319.

Hacquard, S., Garrido-Oter, R., Gonzalez, A., Spaepen, S., Ackermann, G., Lebeis, S., McHardy, A.C., Dangl, J.L., Knight, R., Ley, R. & Schulze-Lefert, P. (2015). Microbiota and Host Nutrition across Plant and Animal Kingdoms. *Cell Host and Microbe*, 17, 603–616.

Hagmann, J., Becker, C., Müller, J., Stegle, O., Meyer, R.C., Wang, G., Schneeberger, K., Fitz, J., Altmann, T., Bergelson, J., Borgwardt, K. & Weigel, D. (2015). Century-scale Methylome Stability in a Recently Diverged Arabidopsis thaliana Lineage. *PLoS Genetics*, 11, e1004920.

Horton, M.W., Bodenhausen, N., Beilsmith, K., Meng, D., Muegge, B.D., Subramanian, S., Vetter, M.M., Vilhjálmsson, B.J., Nordborg, M., Gordon, J.I. & Bergelson, J. (2014). Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nature Communications*, 5, 5320.

Hsiao, E.Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A., Chow, J., Reisman, S.E., Petrosino, J.F., Patterson, P.H. & Mazmanian, S.K. (2013). Microbiota Modulate Behavioral and Physiological Abnormalities Associated with Neurodevelopmental Disorders. *Cell*, 155, 1451–1463.

Hubbell, S.P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography.* Princeton University Press.

Hutchinson, G.E. (1961). The Paradox of the Plankton. *The American Naturalist*, 95, 137–145.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., Giglio, M.G., Hallsworth-Pepin, K., Lobos, E.A., Madupu, R., Magrini, V., Martin, J.C., Mitreva, M., Muzny, D.M., Sodergren, E.J., Versalovic, J., Wollam, A.M., Worley, K.C., Wortman, J.R., Young, S.K., Zeng, Q., Aagaard, K.M., Abolude, O.O., Allen-Vercoe, E., Alm, E.J., Alvarado, L., Andersen, G.L., Anderson, S., Appelbaum, E., Arachchi, H.M., Armitage, G., Arze, C.A., Ayvaz, T., Baker, C.C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M.J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G.A., Buhay, C.J., Busam, D.A., Campbell, J.L., Canon, S.R., Cantarel, B.L., Chain, P.S.G., Chen, I.M.A., Chen, L., Chhibba, S., Chu, K., Ciulla, D.M., Clemente, J.C., Clifton, S.W., Conlan, S., Crabtree, J., Cutting, M.A., Davidovics, N.J., Davis, C.C., DeSantis, T.Z., Deal, C., Delehaunty, K.D., Dewhirst, F.E., Deych, E., Ding, Y., Dooling, D.J., Dugan, S.P., Michael Dunne, W., Scott Durkin, A., Edgar, R.C., Erlich, R.L., Farmer, C.N., Farrell, R.M., Faust, K., Feldgarden, M., Felix, V.M., Fisher, S., Fodor, A.A., Forney, L.J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D.C., Fronick, C.C., Fulton, L.L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M.Y., Goldberg, J.M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B.J., Hamilton, H.A., Harris, E.L., Hepburn, T.A., Herter, B., Hoffmann, D.E., Holder, M.E., Howarth, C., Huang, K.H., Huse, S.M., Izard, J., Jansson, J.K., Jiang, H., Jordan, C., Joshi, V., Katancik, J.A., Keitel, W.A., Kelley, S.T., Kells, C., King, N.B., Knights, D., Kong, H.H., Koren, O., Koren, S., Kota, K.C., Kovar, C.L., Kyrpides, N.C., La Rosa, P.S., Lee, S.L., Lemon, K.P., Lennon, N., Lewis, C.M., Lewis, L., Ley, R.E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.C., Lozupone, C.A., Dwayne Lunsford, R., Madden, T., Mahurkar, A.A., Mannon, P.J., Mardis, E.R., Markowitz, V.M., Mavromatis, K., McCorrison, J.M., McDonald, D., McEwen, J., McGuire, A.L., McInnes, P., Mehta, T., Mihindukulasuriya, K.A., Miller, J.R., Minx, P.J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S.M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K.S., Pop, M., Priest, M.E., Proctor, L.M., Qin, X., Raes, J., Ravel, J., Reid, J.G., Rho, M., Rhodes, R., Riehle, K.P., Rivera, M.C., Rodriguez-Mueller, B., Rogers, Y.H., Ross, M.C., Russ, C., Sanka, R.K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J.A., Schloss, P.D., Schmidt, T.M., Scholz, M., Schriml, L., Schubert, A.M., Segata, N., Segre, J.A., Shannon, W.D., Sharp, R.R., Sharpton, T.J., Shenoy, N., Sheth, N.U., Simone, G.A., Singh, I., Smillie, C.S., Sobel, J.D., Sommer, D.D., Spicer, P., Sutton, G.G., Sykes, S.M., Tabbaa, D.G., Thiagarajan, M., Tomlinson, C.M., Torralba, M., Treangen, T.J., Truty, R.M., Vishnivetskaya, T.A., Walker, J., Wang, L., Wang, Z., Ward, D.V., Warren, W., Watson, M.A., Wellington, C., Wetterstrand, K.A.,

White, J.R., Wilczek-Boney, K., Wu, Y., Wylie, K.M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B.P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J.D., Birren, B.W., Gibbs, R.A., Highlander, S.K., Methé, B.A., Nelson, K.E., Petrosino, J.F., Weinstock, G.M., Wilson, R.K. & White, O. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486, 207–214.

Jakob, K., Goss, E.M., Araki, H., Van, T., Kreitman, M. & Bergelson, J. (2002). Pseudomonas viridiflava and P. syringae–natural pathogens of Arabidopsis thaliana. *Molecular Plant-Microbe Interactions*, 15, 1195–1203.

Kahm, M. & Hasenbrink, G. (2010). grofit: fitting biological growth curves with R. *Journal of Statistical ....*

Karasov, T.L., Kniskern, J.M., Gao, L., DeYoung, B.J., Ding, J., Dubiella, U., Lastra, R.O., Nallu, S., Roux, F., Innes, R.W., Barrett, L.G., Hudson, R.R. & Bergelson, J. (2014). The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature*, 512, 436–440.

Kembel, S.W., Wu, M., Eisen, J.A. & Green, J.L. (2012). Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance. *PLoS Computational Biology*, 8, e1002743.

King, E.O., Ward, M.K. & Raney, D.E. (1954). Two simple media for the demonstration of pyocyanin and fluorescin. *The Journal of laboratory and clinical medicine*, 44, 301–307.

Kinross, J.M., Darzi, A.W. & Nicholson, J.K. (2011). Gut microbiome-host interactions in health and disease. *Genome Med*.

Kniskern, J.M., Traw, M.B. & Bergelson, J. (2007). Salicylic acid and jasmonic acid signaling defense pathways reduce natural bacterial diversity on *Arabidopsis thaliana*. *Molecular Plant-Microbe Interactions*, 20, 1512–1522.

Koenig, J.E., Spor, A., Scalfone, N., Fricker, A.D., Stombaugh, J., Knight, R., Angenent, L.T. & Ley, R.E. (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, 108, 4578–4585.

Kozich, J.J., Westcott, S.L., Baxter, N.T., Highlander, S.K. & Schloss, P.D. (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.

Langenheder, S. & Székely, A.J. (2011). Species sorting and neutral processes are both important during the initial assembly of bacterial communities. *The ISME Journal*, pp. 1–9.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., Clemente, J.C., Burkepile, D.E., Vega Thurber, R.L., Knight, R., Beiko, R.G. & Huttenhower, C. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, 31, 814–821.

Langmead, B. & Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9, 357–359.

Lebeis, S.L., Paredes, S.H., Lundberg, D.S., Breakfield, N., Gehring, J., McDonald, M., Malfatti, S., Glavina del Rio, T., Jones, C.D., Tringe, S.G. & Dangl, J.L. (2015). Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science*, 349, 860–864.

Lee, C., Lee, S., Shin, S.G. & Hwang, S. (2008). Real-time PCR determination of rRNA gene copy number: absolute and relative quantification assays with *Escherichia coli*. *Applied Microbiology and Biotechnology*, 78, 371–376.

Lee, Z.M.P., Bussema, C. & Schmidt, T.M. (2009). *rrn*DB: documenting the number of rRNA and tRNA genes in bacteria and archaea. *Nucleic Acids Research*, 37, D489–93.

Legendre, P. & Gallagher, E.D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271–280.

Levine, J.M. & HilleRisLambers, J. (2009). The importance of niches for the maintenance of species diversity. *Nature*, 461, 254–257.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25, 2078–2079.

Lindow, S.E. & Brandl, M.T. (2003). Microbiology of the Phyllosphere. *Applied and Environmental Microbiology*, 69, 1875–1883.

Logue, J.B. & Lindström, E.S. (2010). Species sorting affects bacterioplankton community composition as determined by 16S rDNA and 16S rRNA fingerprints. *The ISME Journal*, 4, 729–738.

Lotka, A.J. (1920). Analytical Note on Certain Rhythmic Relations in Organic Systems. *Proceedings of the National Academy of Sciences of the United States of America*, 6, 410–415.

Lozupone, C. & Knight, R. (2005). UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71, 8228–8235.

Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., del Rio, T.G., Edgar, R.C., Eickhorst, T., Ley, R.E., Hugenholtz, P., Tringe, S.G. & Dangl, J.L. (2012). Defining the core Arabidopsis thaliana root microbiome. *Nature*, 488, 86–90.

Maignien, L., DeForce, E.A., Chafee, M.E., Eren, A.M. & Simmons, S.L. (2014). Ecological succession and stochastic variation in the assembly of Arabidopsis thaliana phyllosphere communities. *mBio*, 5, e00682–13.

Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M., Verdier, V., Beer, S.V., Machado, M.A., Toth, I., Salmond, G. & Foster, G.D. (2012). Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology*, 13, 614–629.

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

McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R. & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, 6, 610–618.

Monier, J.M. & Lindow, S.E. (2005). Spatial Organization of Dual-Species Bacterial Aggregates on Leaf Surfaces. *Applied and Environmental Microbiology*, 71, 5484–5493.

Morgan, J.L., Darling, A.E. & Eisen, J.A. (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS ONE*, 5, e10209.

van Nood, E., Vrieze, A., Nieuwdorp, M., Fuentes, S., Zoetendal, E.G., de Vos, W.M., Visser, C.E., Kuijper, E.J., Bartelsman, J.F.W.M., Tijssen, J.G.P., Speelman, P., Dijkgraaf, M.G.W. & Keller, J.J. (2013). Duodenal infusion of donor feces for recurrent Clostridium difficile. *The New England journal of medicine*, 368, 407–415.

O'Dwyer, J.P., Kembel, S.W. & Sharpton, T.J. (2015). Backbones of evolutionary history test biodiversity theory for microbes. *Proceedings of the National Academy of Sciences*, 112, 8356–8361.

Ofiteru, I.D., Lunn, M., Curtis, T.P., Wells, G.F., Criddle, C.S., Francis, C.A. & Sloan, W.T. (2010). Combined niche and neutral effects in a microbial wastewater treatment community. *Proceedings of the National Academy of Sciences*, 107, 15345–15350.

O'Kane, S.L. & Al-Shehbaz, I.A. (1997). A Synopsis of Arabidopsis (Brassicaceae). *Novon*, 7, 323.

Oksanen, J., Blanchet, F.G. & Kindt, R. (2013). Vegan: community ecology package, R package version 2.2-1 . *Community ecology ...*.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A.R., Xia, F. & Stevens, R. (2013). The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research*, 42, D206–D214.

Pagaling, E., Strathdee, F., Spears, B.M., Cates, M.E., Allen, R.J. & Free, A. (2014). Community history affects the predictability of microbial ecosystem development. *The ISME Journal*, 8, 19–30.

Panke-Buisse, K., Poole, A.C., Goodrich, J.K., Ley, R.E. & Kao-Kniffin, J. (2015). Selection on soil microbiomes reveals reproducible impacts on plant function. *The ISME Journal*, 9, 980–989.

Perisin, M., Vetter, M., Gilbert, J.A. & Bergelson, J. (2015). 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies. *The ISME Journal*.

Periwal, V. & Scaria, V. (2014). Insights into structural variations and genome rearrangements in prokaryotic genomes. *Bioinformatics*, 31, 1–9.

Platt, A., Horton, M., Huang, Y.S., Li, Y., Anastasio, A.E., Mulyati, N.W., Ågren, J., Bossdorf, O., Byers, D., Donohue, K., Dunning, M., Holub, E.B., Hudson, A., Le Corre, V., Loudet, O., Roux, F., Warthmann, N., Weigel, D., Rivero, L., Scholl, R., Nordborg, M., Bergelson, J. & Borevitz, J.O. (2010). The Scale of Population Structure in Arabidopsis thaliana. *PLoS Genetics*, 6, e1000843.

Price, M.N., Dehal, P.S. & Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular biology and evolution*, 26, 1641–1650.

Price, R.M. & Bonett, D.G. (2002). Distribution-Free Confidence Intervals for Difference and Ratio of Medians. *Journal of Statistical Computation and Simulation*, 72, 119–124.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. & Glöckner, F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, 35, 7188–7196.

Quinlan, A.R. & Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26, 841–842.

Redford, A.J. & Fierer, N. (2009). Bacterial succession on the leaf surface: a novel system for studying successional dynamics. *Microbial Ecology*, 58, 189–198.

Remus-Emsermann, M.N.P., Lücker, S., Müller, D.B., Potthoff, E., Daims, H. & Vorholt, J.A. (2014). Spatial distribution analyses of natural phyllosphere-colonizing bacteria on Arabidopsis thaliana revealed by fluorescence in situ hybridization. *Environmental Microbiology*, pp. n/a–n/a.

Rideout, J.R., He, Y., Navas-Molina, J.A., Walters, W.A., Ursell, L.K., Gibbons, S.M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., Clemente, J.C., Gilbert, J.A., Huse, S.M., Zhou, H.W., Knight, R. & Caporaso, J.G. (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*, 2, e545.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P. & Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499, 431–437.

Rodríguez, H., Fraga, R., Gonzalez, T. & Bashan, Y. (2006). Genetics of phosphate solubilization and its potential applications for improving plant growth-promoting bacteria. *Plant and soil*, 287, 15–21.

Rohde, H., Qin, J., Cui, Y., Li, D., Loman, N.J., Hentschke, M., Chen, W., Pu, F., Peng, Y., Li, J., Xi, F., Li, S., Li, Y., Zhang, Z., Yang, X., Zhao, M., Wang, P., Guan, Y., Cen, Z., Zhao, X., Christner, M., Kobbe, R., Loos, S., Oh, J., Yang, L., Danchin, A., Gao, G.F., Song, Y., Li, Y., Yang, H., Wang, J., Xu, J., Pallen, M.J., Wang, J., Aepfelbacher, M., Yang, R. & E. coli O104:H4 Genome Analysis Crowd-Sourcing Consortium (2011). Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *The New England journal of medicine*, 365, 718–724.

Round, J.L. & Mazmanian, S.K. (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews Immunology*, 9, 313–323.

Samarakoon, T., Wang, S.Y. & Alford, M.H. (2013). Enhancing PCR amplification of DNA from recalcitrant plant specimens using a trehalose-based additive. *Applications in plant sciences*, 1, 1200236.

Schloss, P.D. & Handelsman, J. (2006). Toward a Census of Bacteria in Soil. *PLoS Computational Biology*, 2, e92.

Shade, A., Caporaso, J.G., Handelsman, J., Knight, R. & Fierer, N. (2013). A meta-analysis of changes in bacterial and archaeal communities with time. *The ISME Journal*, 7, 1493–1506.

Shade, A., Jones, S.E., Caporaso, J.G., Handelsman, J., Knight, R., Fierer, N. & Gilbert, J.A. (2014). Conditionally rare taxa disproportionately contribute to temporal changes in microbial diversity. *mBio*, 5, e01371–14–e01371–14.

Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.M., Huse, S.M., Neal, P.R., Arrieta, J.M. & Herndl, G.J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America*, 103, 12115–12120.

Staniczenko, P.P.A., Kopp, J.C. & Allesina, S. (2013). The ghost of nestedness in ecological networks. *Nature Communications*, 4, 1391.

Stoddard, S.F., Smith, B.J., Hein, R., Roller, B.R.K. & Schmidt, T.M. (2014). *rrn*DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Research*, 43, D593–D598.

Team, R.C. (2015). *R: A language and environment for statistical computing 2014*. URL http://www. R-project. org.

Traw, M.B., Kniskern, J.M. & Bergelson, J. (2007). SAR increases fitness of *Arabidopsis thaliana* in the presence of natural bacterial pathogens. *Evolution*, 61, 2444–2449.

Valm, A.M., Mark Welch, J.L. & Borisy, G.G. (2012). CLASI-FISH: principles of combinatorial labeling and spectral imaging. *Systematic and applied microbiology*, 35, 496–502.

Valm, A.M., Mark Welch, J.L., Rieken, C.W., Hasegawa, Y., Sogin, M.L., Oldenbourg, R., Dewhirst, F.E. & Borisy, G.G. (2011). Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. *Proceedings of the National Academy of Sciences*, 108, 4152–4157.

Vandenkoornhuyse, P., Quaiser, A., Duhamel, M., Le Van, A. & Dufresne, A. (2015). The importance of the microbiome of the plant holobiont. *New Phytologist*, pp. n/a–n/a.

Vanwonterghem, I., Jensen, P.D., Dennis, P.G., Hugenholtz, P., Rabaey, K. & Tyson, G.W. (2014). Deterministic processes guide long-term synchronised population dynamics in replicate anaerobic digesters. *The ISME Journal*, 8, 2015–2028.

Venables, W.N. & Ripley, B.D. (2013). *Modern Applied Statistics with S-PLUS*. Statistics and Computing. Springer Science & Business Media, New York, NY.

Vogel, C., Innerebner, G., Zingg, J., Guder, J. & Vorholt, J.A. (2012). Forward genetic in planta screen for identification of plant-protective traits of Sphingomonas sp. strain Fr1 against Pseudomonas syringae DC3000. *Applied and Environmental Microbiology*, 78, 5529–5535.

Volterra, V. (1926). Fluctuations in the Abundance of a Species considered Mathematically. *Nature*, 118, 558–560.

Vorholt, J.A. (2012). Microbial life in the phyllosphere. *Nature Publishing Group*, 10, 828–840.

Wagner, M.R., Lundberg, D.S., Coleman-Derr, D., Tringe, S.G., Dangl, J.L. & Mitchell-Olds, T. (2014). Natural soil microbes alter flowering phenology and the intensity of selection on flowering time in a wild Arabidopsis relative. *Ecology Letters*, 17, 717–726.

Westram, R., Bader, K., Prüsse, E., Kumar, Y., Meier, H., Glöckner, F.O. & Ludwig, W. (2011). *ARB: A Software Environment for Sequence Data*. Metagenomics and Complementary Approaches. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Wickham, H. (2009). ggplot2: Elegant graphics for data analysis.

Wilson, M. & Lindow, S.E. (1994). Coexistence among Epiphytic Bacterial Populations Mediated through Nutritional Resource Partitioning. *Applied and Environmental Microbiology*, 60, 4468–4477.

Wood, D. & Salzberg, S. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15, R46.

Woodcock, S., Van Der Gast, C.J., Bell, T., Lunn, M., Curtis, T.P., Head, I.M. & Sloan, W.T. (2007). Neutral assembly of bacterial communities. *FEMS Microbiology Ecology*, 62, 171–180.

Yilmaz, L.S., Parnerkar, S. & Noguera, D.R. (2011). mathFISH, a web tool that uses thermodynamics-based mathematical models for in silico evaluation of oligonucleotide probes for fluorescence in situ hybridization. *Applied and Environmental Microbiology*, 77, 1118–1122.

Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19, 1586–1592.

Zarraonaindia, I., Owens, S.M., Weisenhorn, P., West, K., Hampton-Marcell, J., Lax, S., Bokulich, N.A., Mills, D.A., Martin, G., Taghavi, S., van der Lelie, D. & Gilbert, J.A. (2015). The soil microbiome influences grapevine-associated microbiota. *mBio*, 6, e02527–14.

Zerbino, D.R. & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18, 821–829.

Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, 14 Suppl 11, S1.

Zhou, J., Liu, W., Deng, Y., Jiang, Y.H., Xue, K., He, Z., Van Nostrand, J.D., Wu, L., Yang, Y. & Wang, A. (2013). Stochastic assembly leads to alternative communities with distinct functions in a bioreactor microbial community. *mBio*, 4, e00584–12–e00584–12.

APPENDIX A

CHAPTER 3 PRIMERS FOR 16S RIBOSOMAL RNA GENE AMPLIFICATION AND SEQUENCING

Primers for paired-end 16S community sequencing (single index) on the Illumina MiSeq platform:

MV574F (forward primer) PCR primer sequence:

Field number (space-delimited), description:
1. 5' Illumina adapter
2. Forward primer pad
3. Forward primer linker
4. Forward primer

AATGATACGGCGACCACCGAGATCTACAC TATGGTAGTT AT ACAATGGGCGMAAGCCTGAT

MP984Rv2 (reverse primer) PCR primer sequence (each sequence contains different barcode):

Field number (space-delimited), description:
1. Reverse complement of 3' Illumina adapter
2. Golay barcode (Listed for 8 x 96 well plates of primers in Table A.1)
3. Reverse primer pad
4. Reverse primer linker
5. Reverse primer

CAAGCAGAAGACGGCATACGAGAT NNNNNNNNNNNN ATACACTCAG AG ACCAGGGTATCTAAKCCTG

Read 1 sequencing primer: 574.984.seq1

Field number (space-delimited), description:
1. Forward primer pad
2. Forward primer linker
3. Forward primer

TATGGTAGTT AT ACAATGGGCGMAAGCCTGAT

Read 2 sequencing primer: 574.984.seq2

Field number (space-delimited), description:
1. Reverse primer pad
2. Reverse primer linker
3. Reverse primer

ATACACTCAG AG ACCAGGGTATCTAAKCCTG

Index sequence primer: 574.984.bc

Field number (space-delimited), description:

1. RC of reverse primer
2. RC of reverse primer linker
3. RC of reverse primer pad

CAGGMTTAGATACCCTGGT CT CTGAGTGTAT

Barcode number	Golay barcode	Barcode number	Golay barcode	Barcode number	Golay barcode	Barcode number	Golay barcode	Barcode number	Golay barcode	Barcode number	Golay barcode	Barcode number	Golay barcode	Barcode number	Golay barcode
rcbc0	TCCCTTGTCTCC	rcbc96	TACCGCTTCTTC	rcbc192	GTCGAATTTGCG	rcbc288	TCTGAGGTTGCC	rcbc384	GGAATTATCGGT	rcbc480	CACCGAAATCTG	rcbc576	TGGAGCCTTGTC	rcbc672	CTACCGATTGCG
rcbc1	ACGAGACTGATT	rcbc97	TGTGCGATAACA	rcbc193	GCATCAGAGTTA	rcbc289	GATCATTCTCTC	rcbc385	CATCAAGCATAG	rcbc481	TGACGTAGAACT	rcbc577	CTCGATGTAAGC	rcbc673	TCACCCAAGGTA
rcbc2	GCTGTACGGATT	rcbc98	GATTATCGACGA	rcbc194	GTGGTCATCGTA	rcbc290	AGACATACCGTA	rcbc386	CATCGCGTTGAC	rcbc482	CTATGCCGGCTA	rcbc578	AGCTTCGACAGT	rcbc674	AGCCAGTCATAC
rcbc3	ATCACCAGGTGT	rcbc99	GCCTAGCCCAAT	rcbc195	CTGAAGGGCGAA	rcbc291	GATCCTCATGCG	rcbc387	GCACATAGTCGT	rcbc483	GTGGTATGGGAG	rcbc579	ATACGCATCAAG	rcbc675	TAACGGCGCTCT
rcbc4	TGGTCAACGATA	rcbc100	GATGTATGTGGT	rcbc196	CGCTCACAGAAT	rcbc292	ATTATCGTCCCT	rcbc388	GGCAAATACACT	rcbc484	TGTACCAACCGA	rcbc580	AGATGTCCGTCA	rcbc676	GTTTGCTCGAGA
rcbc5	ATCGCACAGTAA	rcbc101	ACTCCTTGTGTT	rcbc197	ATTCGGTAGTGC	rcbc293	CCAGACCGCTAT	rcbc389	GTCATGCTCCAG	rcbc485	AGGGTACAGGGT	rcbc581	GCACCTGTTGAA	rcbc677	CAAACGCACTAA
rcbc6	GTCGTGTAGCCT	rcbc102	GTCACGGACATT	rcbc198	CGAGCTGTTACC	rcbc294	AGCTCTAGAAAC	rcbc390	CCTAGTAAGCTG	rcbc486	AGAGTGCTAATC	rcbc582	CCTAGAGAAACT	rcbc678	GAACAAAGAGCG
rcbc7	AGCGGAGGTTAG	rcbc103	GCGAGCGAAGTA	rcbc199	CAACACATGCTG	rcbc295	TCCATCGACGTG	rcbc391	TTACCGACGAGT	rcbc487	TTGGCGGGTTAT	rcbc583	GAGGTTCTTGAC	rcbc679	GCTAAGTGATGT
rcbc8	ATCCTTTGGTTC	rcbc104	ATCTACCGAAGC	rcbc200	ATTCTCTCACGT	rcbc296	CGATGTGTGGTT	rcbc392	GCTTAGATGTAG	rcbc488	CACGATGGTCAT	rcbc584	CTGTAAAGGTTG	rcbc680	AAGGGACAAGTG
rcbc9	TACAGCGCATAC	rcbc105	ACTTGGTGTAAG	rcbc201	CGACTCTAAACG	rcbc297	GCGAAGTTGGGA	rcbc393	AAGACGTAGCGG	rcbc489	GTCACCAATCCG	rcbc585	TGAGTCATTGAG	rcbc681	AGTGTCGATTCG
rcbc10	ACCGGTATGTAC	rcbc106	TCTTGGAGGTCA	rcbc202	GTCTTCAGCAAG	rcbc298	GCATTCGGCGTT	rcbc394	TTACCTTACACC	rcbc490	CACTAACAAACG	rcbc586	TACGGCAGTTCA	rcbc682	CTATTAAGCGGC
rcbc11	AATTGTGTCGGA	rcbc107	TCACCTCCTTGT	rcbc203	CGGATAACCTCC	rcbc299	CGCCATTGTGCA	rcbc395	TGACTAATGGCC	rcbc491	TTCCAGGCAGAT	rcbc587	CTCTAGAAGAGT	rcbc683	CCTACCATTGTT
rcbc12	TGCATACACTGG	rcbc108	GCACACCTGATA	rcbc204	AGGGTGACTTTA	rcbc300	TCCAACTGCAGA	rcbc396	CTCTCTCACTTG	rcbc492	TATGGTACCCAG	rcbc588	TGCACAGTCGCT	rcbc684	GAGTCCGTTGCT
rcbc13	AGTCGAACGAGG	rcbc109	GCGACAATTACA	rcbc205	GACTTCATGCGA	rcbc301	TAAAGACCCGTA	rcbc397	ATTGCAAGCAAC	rcbc493	CACGACTTGACA	rcbc589	CATGCGGATCCT	rcbc685	GATAACTGTACG
rcbc14	ACCAGTGACTCA	rcbc110	TCATGCTCCATT	rcbc206	GCCTGTCTGCAA	rcbc302	TGTATCTTCACC	rcbc398	CACGTGACATGT	rcbc494	CTTGGAGGCTTA	rcbc590	TGCTCCGTAGAA	rcbc686	TAAACCTGGACA
rcbc15	GAATACCAAGTC	rcbc111	AGCTGTCAAGCT	rcbc207	ACTGATGGCCTC	rcbc303	GACTGACTCGTC	rcbc399	CACAGTTGAAGT	rcbc495	ACGTGGTTCCAC	rcbc591	TGATAGGTACAC	rcbc687	CCGAATTGACAA
rcbc16	GTAGATCGTGTA	rcbc112	GAGAGCAACAGA	rcbc208	TTCGATGCCGCA	rcbc304	TCGTGGATAGCT	rcbc400	CTAGGATCACTG	rcbc496	GACGCTTTGCTG	rcbc592	CGAGTTCATCGA	rcbc688	CTGGCATCTAGC
rcbc17	TAACGTGTGTGC	rcbc113	TACTCGGGAACT	rcbc209	TGTGGCTCGTGT	rcbc305	GACGCACTAACT	rcbc401	GATGACCCAAAT	rcbc497	ACAGGGTTTGTA	rcbc593	AAGCAGATTGTC	rcbc689	GGTGGTCGTTCT
rcbc18	CATTATGGCGTG	rcbc114	CGTGCTTAGGCT	rcbc210	AACTTTCAGGAG	rcbc306	GGCGATTTACGT	rcbc402	ACCGGAGTAGGA	rcbc498	GCCTATGAGATC	rcbc594	TAGAGGCGTAGG	rcbc690	ACTATGGGCTAA
rcbc19	CCAATACGCCTG	rcbc115	TACCGAAGGTAT	rcbc211	TGCACGTGATAA	rcbc307	TAAGGCATCGCT	rcbc403	TGAGGACTACCT	rcbc499	CAAACCTATGGC	rcbc595	TCAGCGCCGTTA	rcbc691	GCATTGAGTTCG
rcbc20	GATCTGCGATCC	rcbc116	CACTCATCATTC	rcbc212	GTTCGGTGTCCA	rcbc308	ACCCATACAGCC	rcbc404	CAATCGGCTTGC	rcbc500	ATCGCTTAAGGC	rcbc596	TAGACCGACTCC	rcbc692	GTTGCTGAGTCC
rcbc21	CAGCTCATCAGC	rcbc117	GTATTTCGGACG	rcbc213	AAGACAGCTATC	rcbc309	CGCACTACGCAT	rcbc405	AACACTCGATCG	rcbc501	ACCATCCAACGA	rcbc597	GTCAACGCTGTC	rcbc693	CTATGGTGAACC
rcbc22	CAAACAACAGCT	rcbc118	TATCTATCCTGC	rcbc214	ATTGACCGGTCA	rcbc310	CAGTCGTTAAGA	rcbc406	TGACCGGCTGTT	rcbc502	GCAATAGGAGGA	rcbc598	ACAGGAGGGTGT	rcbc694	GGACCAAGGGAT
rcbc23	GCAACACCATCC	rcbc119	TTGCCAAGAGTC	rcbc215	TTCTCCATCACA	rcbc311	CTACGAAAGCCT	rcbc407	GGAGGAGCAATA	rcbc503	CCGAACGTCACT	rcbc599	GCTGTCGTCAAC	rcbc695	GTATTGGTCAGA
rcbc24	GCGATATATCGC	rcbc120	AGTAGCGGAAGA	rcbc216	CGTAGGTAGAGG	rcbc312	ATAATTGCCGAG	rcbc408	AGCGACGAAGAC	rcbc504	ACACCAACACCA	rcbc600	ATAGAGGCCATT	rcbc696	AGAACCGTCATA
rcbc25	CGAGCAATCCTA	rcbc121	GCAATTAGGTAC	rcbc217	ATTTAGGACGAC	rcbc313	GGCATGTTATCG	rcbc409	CTTCCCTAACTC	rcbc505	CCATCACATAGG	rcbc601	AAGCTTGAAACC	rcbc697	AACTGGAACCCT
rcbc26	AGTCGTGCACAT	rcbc122	CATACCGTGAGT	rcbc218	GGATAGCCAAGG	rcbc314	AGGCACAGTAGG	rcbc410	TGGAAGAACGGC	rcbc506	CGACACGGAGAA	rcbc602	TAAGCGTCTCGA	rcbc698	ATACTCGGCTGC
rcbc27	GTATCTGCGCGT	rcbc123	ATGTGTGTAGAC	rcbc219	TGGTTGGTTACG	rcbc315	CTACTTACATCC	rcbc411	GCTAGACACTAC	rcbc507	GAACCTATGACA	rcbc603	ATAGCTTCGTGG	rcbc699	ACGCTTAACGAC
rcbc28	CGAGGGAAAGTC	rcbc124	CCTGCGAAGTAT	rcbc220	GTCGTCCAAATG	rcbc316	CTCTTCTGATCA	rcbc412	TTGGATTGAACG	rcbc508	ATGCCGGTAATA	rcbc604	CGGGATCAAATT	rcbc700	AGCTTACCGACC
rcbc29	CAAATTCGGGAT	rcbc125	TTCTCTCGACAT	rcbc221	CAACGTGCTCCA	rcbc317	ATGCTAACCACG	rcbc413	GATATACCAGTG	rcbc509	GAACAGCTCTAC	rcbc605	AGTCATCGAATG	rcbc701	AGGGCTATAGTT
rcbc30	AGATTGACCAAC	rcbc126	GCTCTCCGTAGA	rcbc222	TACACAAGTCGC	rcbc318	ACCAATCTCGGC	rcbc414	AACAAACTGCCA	rcbc510	GTGAGTCATACC	rcbc606	ATCTTGGAGTCG	rcbc702	TGTCTCGCAAGC
rcbc31	AGTTACGAGCTA	rcbc127	GTTAAGCTGACC	rcbc223	GCGTCCATGAAT	rcbc319	TATCCAAGCGCA	rcbc415	GTAGACATGTGT	rcbc511	TGGCCGTTACTG	rcbc607	AGCACCGGTCTT	rcbc703	CAGCCGCATATC
rcbc32	GCATATGCACTG	rcbc128	ATGCCATGCCGT	rcbc224	GTAATGCGTAAC	rcbc320	GTACTGAAGATC	rcbc416	TACAGTTACGCG	rcbc512	TAGAGCTGCCAT	rcbc608	GCAAATCAGCCT	rcbc704	GATACGTTCGCA
rcbc33	CAACTCCCGTGA	rcbc129	GACATTGTCACG	rcbc225	GTCGCCGTACAT	rcbc321	TCGCCGTGTACA	rcbc417	CAAGCCCTAGTA	rcbc513	ATCTAGTGGCAA	rcbc609	GCAAGCTGTCTC	rcbc705	CCAAGATTCGCC
rcbc34	TTGCGTTAGCAG	rcbc130	GCCAACAACCAT	rcbc226	GGAATCCGATTA	rcbc322	AACTGCGATATG	rcbc418	TAGTGTCGGATC	rcbc514	CCTTCAATGGGA	rcbc610	AGCGGCCTATTA	rcbc706	GAGGCTGATTTA
rcbc35	TACGAGCCCTAA	rcbc131	ATCAGTACTAGG	rcbc227	CACCCGATGGTT	rcbc323	CTTCCAACTCAT	rcbc419	CTGAGCTCTGCA	rcbc515	TTGACGACATCG	rcbc611	TCTTCAACTACC	rcbc707	GAGTTAGCATCA
rcbc36	CACTACGCTAGA	rcbc132	TCCTCGAGCGAT	rcbc228	TTCTGAGAGGTA	rcbc324	GAGATCGCCTAT	rcbc420	CTTCGACTTTCC	rcbc516	ACATACTGAGCA	rcbc612	TGGAATTCGGCT	rcbc708	TGTAGTATAGGC
rcbc37	TGCAGTCCTCGA	rcbc133	ACCCAAGCGTTA	rcbc229	ATCCCTACGGAA	rcbc325	TGTACATCGCCG	rcbc421	GTCATAAGAACC	rcbc517	GGCTAAACTATG	rcbc613	TAAGATGCAGTC	rcbc709	CTCACGCAATGC
rcbc38	ACCATAGCTCCG	rcbc134	TGCAGCAAGATT	rcbc230	GGTTCCATTAGG	rcbc326	TGTTAAGCAGCA	rcbc422	GTCCGCAAGTTA	rcbc518	AAGAGCAGAGCC	rcbc614	TGCCGAGTAATC	rcbc710	GTCCCGTGAAAT
rcbc39	TCGACATCTCTT	rcbc135	AGCAACATTGCA	rcbc231	GTGTTCCCAGAA	rcbc327	ACGGCGTTATGT	rcbc423	CGTAGAGCTCTC	rcbc519	GGAGAGATCACG	rcbc615	ACCTTGACAAGA	rcbc711	GGACAGTGTATT
rcbc40	GAACACTTTGGA	rcbc136	GATGTGGTGTTA	rcbc232	CCGAGGTATAAT	rcbc328	ACTTTGCTTTGC	rcbc424	CCTCTGAGAGCT	rcbc520	TCAACCCGTGAA	rcbc616	GTAACCACCACC	rcbc712	ACACGACTATAG
rcbc41	GAGCCATCTGTA	rcbc137	CAGAAATGTGTC	rcbc233	AGCGTAATTAGC	rcbc329	CAAAGCGGTATT	rcbc425	CCTCGATGCAGT	rcbc521	GTTTGAAACACG	rcbc617	CATAGCTCGGTC	rcbc713	GTGTAGGTGCTT
rcbc42	TTGGGTACACGT	rcbc138	GTAGAGGTAGAG	rcbc234	CTCGTGAATGAC	rcbc330	CGAAACTACGTA	rcbc426	GCGGACTATTCA	rcbc522	AGAGAGACAGGT	rcbc618	AACCATGCCAAC	rcbc714	TGAACTAGCGTC
rcbc43	AAGGCGCTCCTT	rcbc139	CGTGATCCGCTA	rcbc235	AGGTGAGTTCTA	rcbc331	GAGGACCAGCAA	rcbc427	CGTGCACAATTG	rcbc523	TCGCCAGTGCAT	rcbc619	TATGGAGCTAGT	rcbc715	TCCGAGTCACCA
rcbc44	TAATACGGATCG	rcbc140	GGTTATTTGGCG	rcbc236	CCTGTCCTATCT	rcbc332	AATAGCATGTCG	rcbc428	CGGCCTAAGTTC	rcbc524	GCTCAGGACTCT	rcbc620	ACTACCTCTTCA	rcbc716	TCCTCTTTGGTC
rcbc45	TCGGAATTAGAC	rcbc141	GGATCGTAATAC	rcbc237	GGTTTAACACGC	rcbc333	CGGAGTAATCCT	rcbc429	AGCGCTCACATC	rcbc525	CACTTTGGGTGC	rcbc621	GATGATAACCCA	rcbc717	TCCACCCTCTAT
rcbc46	TGTGAATTCGGA	rcbc142	GCATAGCATCAA	rcbc238	AGACAGTAGGAG	rcbc334	CTGTGTCCATGG	rcbc430	TGGTTATGGCAC	rcbc526	TCTAGCCTGGCA	rcbc622	GGCCCAATATAA	rcbc718	TCGTGACGCTAA
rcbc47	CATTCGTGGCGT	rcbc143	GTGTTAGATGTG	rcbc239	GCCACGACTTAC	rcbc335	CTTCGCGGATGT	rcbc431	CGAGGTTCTGAT	rcbc527	AATGCAATGCGT	rcbc623	TTGTATGACAGG	rcbc719	ACGGCTAGTTCC
rcbc48	TACTACGTGGCC	rcbc144	TTAGAGCCATGC	rcbc240	ATTGTTCCTACC	rcbc336	ATAGGCTGTAGT	rcbc432	AACTCCTGTGGA	rcbc528	CGAATGAGTCAT	rcbc624	GGTAAGTTTGAC	rcbc720	GCACTGGCATAT
rcbc49	GGCCAGTTCCTA	rcbc145	TGAACCCTATGG	rcbc241	GCCGTAAACTTG	rcbc337	TGTGTAGCCATG	rcbc433	TAATGGTCGTAG	rcbc529	CAACGCTAGAAT	rcbc625	CTACCACGGTAC	rcbc721	GGCATTAGTTGA
rcbc50	GATGTTCGCTAG	rcbc146	AGAGTCTTGCCA	rcbc242	GCAGATTTCCAG	rcbc338	AAGGGCGCTGAA	rcbc434	TTGCACCGTCGA	rcbc530	ATCAGAGCCCAT	rcbc626	CGGTCTGTCTGA	rcbc722	CGGTAGTTGATC
rcbc51	CTATCTCCTGTC	rcbc147	ACAACACTCCGA	rcbc243	AGATGATCAGTC	rcbc339	GTTTCCGTGGTG	rcbc435	TGCTACAGACGT	rcbc531	TCTGTAGAGCCA	rcbc627	GTACATGTCGCC	rcbc723	TGAAAGCGGCGA
rcbc52	ACTCACAGGAAT	rcbc148	CGATGCTGTTGA	rcbc244	GAGACGTGTTCT	rcbc340	AGGAACCAGACG	rcbc436	ATGGCCTGACTA	rcbc532	CCGACTCTAGGT	rcbc628	TTCTAGAGTGCG	rcbc724	GGTTACGGTTAC
rcbc53	ATGATGAGCCTC	rcbc149	ACGACTGCATAA	rcbc245	TATCACCGGCAC	rcbc341	TAATGCCCAGGT	rcbc437	ACGCACATACAA	rcbc533	ATCCTACGAGCA	rcbc629	ACGGATGTTATG	rcbc725	ACATCAGGTCAC
rcbc54	GTCGACAGAGGA	rcbc150	ACGCGAACTAAT	rcbc246	TATGCCAGAGAT	rcbc342	TATGAACGTCCG	rcbc438	TGAGTGGTCTGT	rcbc534	GACAACGAATCT	rcbc630	TTGAGGCTACAA	rcbc726	GTTGATACGATG
rcbc55	TGTCGCAAATAG	rcbc151	AGCTATGTATGG	rcbc247	AGGTCCAAATCA	rcbc343	CCACATTGGGTC	rcbc439	GATAGCACTCGT	rcbc535	TGCGGTTGACTC	rcbc631	GTAGGAACCGGA	rcbc727	CAGACACTTCCG
rcbc56	CATCCCTCTACT	rcbc152	ACGGGTCATCAT	rcbc248	ACCGTGCTCACA	rcbc344	TCAGTCAGATGA	rcbc440	TAGCGCGAACTT	rcbc536	TGAGAAGAAAGG	rcbc632	ACATCTAGCAGA	rcbc728	TCACCATCCGAG
rcbc57	TATACCGCTGCG	rcbc153	GAAACATCCCAC	rcbc249	CTCCCTTTGTGT	rcbc345	AAGTCACACACA	rcbc441	CATACACGCACC	rcbc537	TCGGATCTGTGA	rcbc633	CCGACATTGTAG	rcbc729	ACCCACCACTAG
rcbc58	AGTTGAGGCATT	rcbc154	CGTACTCTCGAG	rcbc250	AGCTGCACCTAA	rcbc346	GCTGTGATTCGA	rcbc442	ACCTCAGTCAAG	rcbc538	GCCGGTACTCTA	rcbc634	CATGTAAGGCTC	rcbc730	CAGAAGGTGTGG
rcbc59	ACAATAGACACC	rcbc155	TCAGTTCTCGTT	rcbc251	CCTTGACCGATG	rcbc347	CTAGCTATGGAC	rcbc443	TCGACCAAACAC	rcbc539	CACAGGATTACC	rcbc635	TGCAAGCTAAGT	rcbc731	GAAGCTTGAATC
rcbc60	CGGTCAATTGAC	rcbc156	TCGTGCGTGTTG	rcbc252	CTATCATCCTCA	rcbc348	CTTGACGAGGTT	rcbc444	CCACCCAGTAAC	rcbc540	CGATATCAGTAG	rcbc636	GTGTGTGCCATA	rcbc732	ACTAGGATCAGT
rcbc61	GTGGAGTCTCAT	rcbc157	GTTATCGCATGG	rcbc253	ACTCTAGCCGGT	rcbc349	ACCTGGGAATAT	rcbc445	ATATCGCGATGA	rcbc541	CATAAGGGAGGC	rcbc637	TGACAACCGAAT	rcbc733	GCTCCTTAGAAG
rcbc62	GCTCGAAGATTC	rcbc158	GATCACGAGAGG	rcbc254	CGATAGGCCTTA	rcbc350	CTCTGCCTAATT	rcbc446	CGCCGGTAATCT	rcbc542	TGTGTTACTCCT	rcbc638	TAGGCTCGTGCT	rcbc734	TCCCATTCCCAT
rcbc63	AGGCTTACGTGT	rcbc159	GTAAATTCAGGC	rcbc255	AATGACCTCGTG	rcbc351	ATATGACCCAGC	rcbc447	CCGATGCCTTGA	rcbc543	GGTACCTGCAAT	rcbc639	CTCCTTAAGGCG	rcbc735	TGGCGTCATTCG
rcbc64	TCTCTACCACTC	rcbc160	AGTGTTTCGGAC	rcbc256	CTTAGGCATGTG	rcbc352	CTCTATTCCACC	rcbc448	AGCAGGCACGAA	rcbc544	TCGCCTATAAGG	rcbc640	TTGCCTGGGTCA	rcbc736	AATCCTCGGAGT
rcbc65	ACTTCCAACTTC	rcbc161	ACACGCGGTTTA	rcbc257	CCAGATATAGCA	rcbc353	ATTGAGTGAGTC	rcbc449	TACGCAGCACTA	rcbc545	AGTGGCACTATC	rcbc641	CAATTCTGCTTC	rcbc737	CTGGACGCATTA
rcbc66	CTCACCTAGGAA	rcbc162	TGGCAAATCTAG	rcbc258	GAGAGTCCACTT	rcbc354	TTATGGTACGGA	rcbc450	CGCTTAGTGCTG	rcbc546	TAACCCGATAGA	rcbc642	ACTGGCAAACCT	rcbc738	ACCGATTAGGTA
rcbc67	GTGTTGTCGTGC	rcbc163	CACCTTACCTTA	rcbc259	GAACGGGACGTA	rcbc355	GCTAGTTATGGA	rcbc451	CAAAGTTTGCGA	rcbc547	GTGTGCTAACGT	rcbc643	AATCAGAGCTTG	rcbc739	ATGTGCTGCTCG
rcbc68	CCACAGATCGAT	rcbc164	TTAACCTTCCTG	rcbc260	ACGTGTAGGCTT	rcbc356	CAGATTAACCAG	rcbc452	TCGAGCCGATCT	rcbc548	CTTGCGGCAATC	rcbc644	CAATGTAGACAC	rcbc740	TACGTACGAAAC
rcbc69	TATCGACACAAG	rcbc165	TGCCGTATGCCA	rcbc261	GGTCTCCTACAG	rcbc357	GGCTGCATACTC	rcbc453	CTCATCATGTTC	rcbc549	TGAGGTTTGATG	rcbc645	TGGCGATACGTT	rcbc741	ATCACATTCTCC
rcbc70	GATTCCGGCTCA	rcbc166	CGTGACAATAGT	rcbc262	ACTGACTTAAGG	rcbc358	TTGGTAAAGTGC	rcbc454	CCAGGGACTTCT	rcbc550	ATTGCTGGTCGA	rcbc646	GCCTTACGATAG	rcbc742	AGCCTGGTACCT
rcbc71	CGTAATTGCCGC	rcbc167	CGCTACAACTCG	rcbc263	GATGCTGCCGTT	rcbc359	AAGTGGCTATCC	rcbc455	GCAATCCTTGCG	rcbc551	AAGAAGCCGGAC	rcbc647	TACCTGTGTCTT	rcbc743	GCTAAAGTCGTA
rcbc72	GGTGACTAGTTC	rcbc168	TTAAGACAGTCG	rcbc264	TTCCTAGGCCAG	rcbc360	AACCGATGTACC	rcbc456	CCTGCTTCCTTC	rcbc552	ACGGGATACAGG	rcbc648	AACGAGGCAACG	rcbc744	TCTCAGCGCGTA
rcbc73	ATGGGTTCCGTC	rcbc169	TCTGCACTGAGC	rcbc265	ATTAAGCCTGGA	rcbc361	TCGATTGGCCGT	rcbc457	CAAGGCACAAGG	rcbc553	AAGAGTCTCTAG	rcbc649	GAAGACAGCGAC	rcbc745	GACCCTAGACCT
rcbc74	TAGGCATGCTTG	rcbc170	CGCAGATTAGTA	rcbc266	TGGCTTTCTATC	rcbc362	GCATTACTGGAC	rcbc458	GGCCTATAAGTC	rcbc554	TCCGTCATGGGT	rcbc650	ACACCTGCGATC	rcbc746	TATTCAGCGGAC
rcbc75	AACTAGTTCAGG	rcbc171	TGGGTCCCACAT	rcbc267	ACAGCTCAAACA	rcbc363	TTGGGCCACATA	rcbc459	TCCATTTCATGC	rcbc555	AGATCTATGCAG	rcbc651	GGCGTTGCATTC	rcbc747	GTTCCGGATTAG
rcbc76	ATTCTGCCGAAG	rcbc172	CACTGGTGCATA	rcbc268	GAGCGTATCCAT	rcbc364	CACACAAAGTCA	rcbc460	TCGGCGATCATC	rcbc556	GCACAAGGCAAG	rcbc652	ACTAGCGTTCAG	rcbc748	GCGTGTAATTAG
rcbc77	AGCATGTCCCGT	rcbc173	AACGTAGGCTCT	rcbc269	ATGGGCGAATGG	rcbc365	GCCAAGGATAGG	rcbc461	GTTTCACGCGAA	rcbc557	CGGCAAACACTT	rcbc653	TTGCGACAAAGT	rcbc749	CTGTAGCTTGGC
rcbc78	GTACGATATGAC	rcbc174	AGTTGTAGTCCG	rcbc270	GATCTCTGGGTA	rcbc366	CGCCACGTGTAT	rcbc462	ACAAGAACCTTG	rcbc558	GCGAGTTCCTGT	rcbc654	TGCGAGTATATG	rcbc750	ATGCCTCGTAAG
rcbc79	GTGGTGGTTTCC	rcbc175	TCGTCAAACCCG	rcbc271	CATCATACGGGT	rcbc367	GCAACCGATTGT	rcbc463	TACTCTCTTAGC	rcbc559	TTCCGAATCGGC	rcbc655	TACCACAACGAA	rcbc751	ACCTATGGTGAA
rcbc80	TAGTATGCGCAA	rcbc176	TAATCGGTGCCA	rcbc272	TACGGATTATGG	rcbc368	CATGTGCTTAGG	rcbc464	AACTGTTCGCGC	rcbc560	TACCTAGTGAGA	rcbc656	TCTGGAACGGTT	rcbc752	CTGTTACAGCGA
rcbc81	TGCGCTGAATGT	rcbc177	TTGATCCGGTAG	rcbc273	ATAGCGAACTCA	rcbc369	GTTCCTCCATTA	rcbc465	CGAAGCATCTAC	rcbc561	CGTTCTGGTGGT	rcbc657	GTACTACCTCGG	rcbc753	CAGTCAGGCCTT
rcbc82	ATGGCTGTCAGT	rcbc178	CGGGTGTTTGCT	rcbc274	TAACGCTGTGTG	rcbc370	ACCTGTCCTTTC	rcbc466	GTTTGGCCACAC	rcbc562	TTGGTCTCCTCT	rcbc658	TTCCTGTTAACC	rcbc754	ACTGAGCTGCAT
rcbc83	GTTCTCTTCTCG	rcbc179	TTGACCGCGGTT	rcbc275	AACCAAACTCGA	rcbc371	GTTCACGCCCAA	rcbc467	TCAGGTTGCCCA	rcbc563	CTGCATACTGAG	rcbc659	CTATCCAAGTGG	rcbc755	ACGAAGTCTACC
rcbc84	CGTAAGATGCCT	rcbc180	GTGCAACCAATC	rcbc276	GCCGTCTCGTAA	rcbc372	CGATCGAACACT	rcbc468	TCATTCCACTCA	rcbc564	CAGGGCCTTTGT	rcbc660	CAGTCTAGTACG	rcbc756	ACCGTCTTTCTC
rcbc85	GCGTTCTAGCTG	rcbc181	GCTTGAGCTTGA	rcbc277	CTGGGTATCTCG	rcbc373	CATGCCAACATG	rcbc469	GTCACATCACGA	rcbc565	CGATGAATATCG	rcbc661	GTGTCCGGATTC	rcbc757	AGTCTGTCTGCG
rcbc86	GTTGTTCTGGGA	rcbc182	CGCTGTGGATTA	rcbc278	GACTACCCGTTG	rcbc374	GAGTACAGTCTA	rcbc470	CGACATTTCTCT	rcbc566	GTCAATTAGTGG	rcbc662	TGTGGTGATGTA	rcbc758	CCGCACTCAAGT
rcbc87	GGACTTCCAGCT	rcbc183	CTGTCAGTGACC	rcbc279	GCGTTGCAAACT	rcbc375	CCTACATGAGAC	rcbc471	GGACGTTAACTA	rcbc567	AGTACGCAGTCT	rcbc663	CTTTCGTTCAAC	rcbc759	TGTGGAAACTCC
rcbc88	CTCACAACCGTG	rcbc184	ACGATTCGAGTC	rcbc280	AACCGCATAAGT	rcbc376	TCCGTGGTATAG	rcbc472	TAGCAGTTGCGT	rcbc568	AGCAGCTATTGC	rcbc664	CCGAAGATTCTG	rcbc760	TTAGGCAGGTTC
rcbc89	CTGCTATTCCTC	rcbc185	GGTTCGGTCCAT	rcbc281	ACCTTACACCTT	rcbc377	TCTACGGCACGT	rcbc473	CACGCTATTGGA	rcbc569	CTCGGATAGATC	rcbc665	GTTGGCGTTACA	rcbc761	TAAGACTACTGG
rcbc90	ATGTCACCGCTG	rcbc186	CTGATCCATCTT	rcbc282	GTAGGTGCTTAC	rcbc378	ATGCTGCAACAC	rcbc474	AACTTCACTTCC	rcbc570	TTCCCGAAACGA	rcbc666	GAAGTAGCGAGC	rcbc762	CGCGAAGTTTCA
rcbc91	TGTAACGCCGAT	rcbc187	TATGTGCCGGCT	rcbc283	CGCATTTGGATG	rcbc379	TTCTCATGGAGG	rcbc475	CCAGTGGATATA	rcbc571	GAACTTTAGCGC	rcbc667	TTGCGGACCCTA	rcbc763	CGATACACTGCC
rcbc92	AGCAGAACATCT	rcbc188	TGGTCGCATCGT	rcbc284	ATAACATGTGCG	rcbc380	CATAGTGATTGG	rcbc476	TGTGTGTAACGC	rcbc572	TCCTTAGAAGGC	rcbc668	GCGGAAACATGG	rcbc764	TTGAAATCCCGG
rcbc93	TGGAGTAGGTGG	rcbc189	TGTAAGACTTGG	rcbc285	CTTGAGAAATCG	rcbc381	GCTATCAAGACA	rcbc477	CCAATCGTGCAA	rcbc573	GATGGACTTCAA	rcbc669	AACGTTAGTGTG	rcbc765	GTTAGGGAGCGA
rcbc94	TTGGCTCTATTC	rcbc190	CGGATCTAGTGT	rcbc286	CTACACAGCACA	rcbc382	CCGTGACAACTC	rcbc478	AGGCTAGCAGAG	rcbc574	TACTGAGCCTCG	rcbc670	TGCATGACAGTC	rcbc766	TTACTGTGGCCG
rcbc95	GATCCCACGTAC	rcbc191	CGATCTTCGAGC	rcbc287	GAAATGCTACGT	rcbc383	CGTTCCTTGTTA	rcbc479	GTCACTCCGAAC	rcbc575	AGAAGGCCTTAT	rcbc671	TCAATCGCTTTC	rcbc767	ATATAAGGCCCA

Table A.1: Golay barcodes used with MP984Rv2 primers. Each column lists barcodes for a 96 well plate of primers.

APPENDIX B

CHAPTER 4 PRIMERS FOR 16S RIBOSOMAL RNA GENE AMPLIFICATION AND SEQUENCING

Primers for paired-end 16S community sequencing (dual indices) on the Illumina MiSeq platform:

799F (forward primer) PCR primer sequence:

Field number (space-delimited), description:
1. 5' Illumina adapter
2. i5 index (Listed in Table B.1)
3. Forward primer pad
3. Forward primer linker
4. Forward primer

AATGATACGGCGACCACCGAGATCTACAC NNNNNNNN TACCCCCCTC GT AACMGGATTAGATACCCKG

1193R (reverse primer) PCR primer sequence (each sequence contains different barcode):

Field number (space-delimited), description:
1. Reverse complement of 3' Illumina adapter
2. i7 index (Listed in Table B.1)
3. Reverse primer pad
4. Reverse primer linker
5. Reverse primer

CAAGCAGAAGACGGCATACGAGAT NNNNNNNN TCATTCCTGG GC ACGTCATCCCCACCTTCC

Read 1 sequencing primer: Read1_799f

Field number (space-delimited), description:
1. Forward primer pad
2. Forward primer linker
3. Forward primer

TACCCCCCTC GT AACMGGATTAGATACCCKG

Read 2 sequencing primer: Read2_1193r

Field number (space-delimited), description:
1. Reverse primer pad
2. Reverse primer linker
3. Reverse primer

TCATTCCTGG GC ACGTCATCCCCACCTTCC

Index 1 sequence primer: IndexRead_1193r

Field number (space-delimited), description:
1. RC of reverse primer
2. RC of reverse primer linker
3. RC of reverse primer pad

GGAAGGTGGGGATGACGT GC CCAGGAATGA

i5 name	i5 index	i7 name	i7 index
799f.SA501	ATCGTACG	1193r.SA701	AACTCTCG
799f.SA502	ACTATCTG	1193r.SA702	ACTATGTC
799f.SA503	TAGCGAGT	1193r.SA703	AGTAGCGT
799f.SA504	CTGCGTGT	1193r.SA704	CAGTGAGT
799f.SA505	TCATCGAG	1193r.SA705	CGTACTCA
799f.SA506	CGTGAGTG	1193r.SA706	CTACGCAG
799f.SA507	GGATATCT	1193r.SA707	GGAGACTA
799f.SA508	GACACCGT	1193r.SA708	GTCGCTCG
799f.SB501	CTACTATA	1193r.SA709	GTCGTAGT
799f.SB502	CGTTACTA	1193r.SA710	TAGCAGAC
799f.SB503	AGAGTCAC	1193r.SA711	TCATAGAC
799f.SB504	TACGAGAC	1193r.SA712	TCGCTATA
799f.SB505	ACGTCTCG	1193r.SB701	AAGTCGAG
799f.SB506	TCGACGAG	1193r.SB702	ATACTTCG
799f.SB507	GATCGTGT	1193r.SB703	AGCTGCTA
799f.SB508	GTCAGATA	1193r.SB704	CATAGAGA
799f.SC501	ACGACGTG	1193r.SB705	CGTAGATC
799f.SC502	ATATACAC	1193r.SB706	CTCGTTAC
799f.SC503	CGTCGCTA	1193r.SB707	GCGCACGT
799f.SC504	CTAGAGCT	1193r.SB708	GGTACTAT
799f.SC505	GCTCTAGT	1193r.SB709	GTATACGC
799f.SC506	GACACTGA	1193r.SB710	TACGAGCA
799f.SC507	TGCGTACG	1193r.SB711	TCAGCGTT
799f.SC508	TAGTGTAG	1193r.SB712	TCGCTACG
799f.SD501	AAGCAGCA	1193r.SC701	ACCTACTG
799f.SD502	ACGCGTGA	1193r.SC702	AGCGCTAT
799f.SD503	CGATCTAC	1193r.SC703	AGTCTAGA
799f.SD504	TGCGTCAC	1193r.SC704	CATGAGGA
799f.SD505	GTCTAGTG	1193r.SC705	CTAGCTCG
799f.SD506	CTAGTATG	1193r.SC706	CTCTAGAG
799f.SD507	GATAGCGT	1193r.SC707	GAGCTCAT
799f.SD508	TCTACACT	1193r.SC708	GGTATGCT
		1193r.SC709	GTATGACG
		1193r.SC710	TAGACTGA
		1193r.SC711	TCACGATG
		1193r.SC712	TCGAGCTC
		1193r.SD701	ACCTAGTA
		1193r.SD702	ACGTACGT
		1193r.SD703	ATATCGCG
		1193r.SD704	CACGATAG
		1193r.SD705	CGTATCGC
		1193r.SD706	CTGCGACT
		1193r.SD707	GCTGTAAC
		1193r.SD708	GGACGTTA
		1193r.SD709	GGTCGTAG
		1193r.SD710	TAAGTCTC
		1193r.SD711	TACACAGT
		1193r.SD712	TTGACGCA

Table B.1: Illumina indices used with 799F and 1193R primers.