8-2013

VALIDATING APPROACHES FOR STUDYING MICROBIAL DIVERSITY TO CHARACTERIZE COMMUNITIES FROM ROOTS OF Populus deltoides

Migun Shakya
mshakya@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Migun Shakya entitled "VALIDATING APPROACHES FOR STUDYING MICROBIAL DIVERSITY TO CHARACTERIZE COMMUNITIES FROM ROOTS OF Populus deltoides." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Mircea Podar, Major Professor

We have read this dissertation and recommend its acceptance:

Christopher W. Schadt, Mitch Doktycz, Alison Buchan, Igor Jouline

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

8-2013

VALIDATING APPROACHES FOR STUDYING MICROBIAL DIVERSITY TO CHARACTERIZE COMMUNITIES FROM ROOTS OF Populus deltoides

Migun Shakya
mshakya@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Migun Shakya entitled "VALIDATING APPROACHES FOR STUDYING MICROBIAL DIVERSITY TO CHARACTERIZE COMMUNITIES FROM ROOTS OF Populus deltoides." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Mircea Podar, Major Professor

We have read this dissertation and recommend its acceptance:

Christopher W. Schadt, Mitch Doktycz, Alison Buchan, Igor Jouline

Accepted for the Council:
Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

VALIDATING APPROACHES FOR STUDYING MICROBIAL DIVERSITY TO CHARACTERIZE COMMUNITIES FROM ROOTS OF *Populus deltoides*

A Dissertation
Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Migun Shakya

August 2013

For my parents,

Buddha Ratna Shakya

Mina Shakya

loving wife,

Urshula Shakya

and my beloved aunt,

Late Anjali Shakya

Acknowledgements

This dissertation research would not have been possible without the help and support of my mentors, friends, and family. First and foremost, I would like to express my sincere gratitude to my advisors, Drs. Mircea Podar and Christopher W. Schadt. They have worked tirelessly to improve my scientific execution, writing, publication, and overall scientific way of thinking. Thank you for your support, mentorship, and dedication.

I would not have been able to be a scientist that I am today without my committee members in Drs. Mitch Doktycz, Alison Buchan, and Igor B. Jouline. Thank you for your suggestions, constructive criticisms, and insightful comments.

During my graduate studies, I met some of the most amazing colleagues and friends in Alisha Campbell, James Campbell, Scott Hamilton-Brehm, Mike Robeson, Anil Somenahally, Richard Hurt, Neil Gottel, Marilyn Kerley, Santosh Bhatt, Zamin K. Yang, Hector Castro, Jerreme Jackson, and Ritin Sharma. I would like to acknowledge all of them for their friendship, support, and comments on matters both inside and outside of the lab. Throughout the graduate school, I have also had the pleasure of collaborating with a number of fellow scientist including Anna Louise Reysenbach and Gilberto Flores at Portland State University, Dwayne Elias at Oak Ridge National Lab, and Christopher Quince at University of Glasglow. I am grateful for the opportunity.

None of this would have been possible without constant support and sacrifice from my family. My parents, grand parents, uncles, and aunts have all contributed in one way or another towards my success. I will be forever indebted to them. Lastly, a special thanks to my wife Urshula for keeping me happy during arduous parts of my dissertation.

"Science is a way of thinking

much more than it is a body of knowledge."

-Carl Sagan, *The fine art of baloney detection*

Abstract

Microbial (archaeal, bacterial, and fungal) communities associated with plant roots are central to its health, survival, and growth. However, a robust understanding of root microbiota and the factors that govern their community structure and dynamics have remained elusive, especially in mature perennial plants from natural settings. Although the advent of Next Generation Sequencing (NGS) technologies have changed the scale of microbial ecological studies by enabling exhaustive characterization of microbial communities, the accuracy of taxonomic and quantitative inferences are affected by multiple experimental and computational steps and lack of knowledge of the true ecological diversity. To test for inaccuracies and biases, I assembled diverse bacterial and archaeal 'synthetic communities' from genomic DNAs of sequenced organisms. I tested and compared different approaches that included metagenomic and small subunit rRNA (SSU rRNA) amplicon sequencing. The outcome was dependent on primer pairs, analysis parameters, and sequencing platforms. Nevertheless, new approaches in processing and classifying amplicons were able to recapitulate microbial diversity with high reproducibility within primer sets, even though all tested primers sets showed taxon-specific biases. Consequently, inferences from 'synthetic communities' study were implemented in experimental design and analysis of microbial communities from roots of naturally occurring mature riparian plants of *Populus deltoides*. *Thaumarchaeota*, *Proteobacteria* and *Ascomycota* dominated the overall archaeal, bacterial, and fungal communities respectively. Further, I investigated relationships of bacterial and fungal communities in rhizosphere and endosphere with soil and environmental properties, host genotype, season, and geographic setting. The variation of bacterial and fungal communities between each sampled roots were explained on the basis of seasonal, soil properties, and geographical settings (4% to 23%), however, most variations remain

unexplained. I also tested if rhizosphere of *P. deltoides* and mature trees in general select for higher diversity of archaea than surrounding soil. I discovered a slightly higher diversity of archaea in the trees compared to corresponding bulk soil, but the results were not specific to *P. deltoides*. In summary, this dissertation validates current microbial diversity approaches, characterizes microbial communities of an important plant, and decipher drivers that are controlling root associated community structure.

Table of Contents

List of Tables

List of Figures

Nomenclature

$amoA$	ammonia monooxygenase subunit A
AOA	Ammonia Oxidizing Archaea
AOB	Ammonia Oxidizing Bacteria
CCA	Canonical Correspondence Analysis
$db-RDA$	distance based Redundancy Analysis
DNA	Deoxyribonucleic Acid
$DSMZ$	Deutsche Sammlung von Mikroorganismen
$gDNA$	genomic DNA
NGS	Next Generation Sequencing
OTU	Operational Taxonomic Unit
$PCNM$	Principal Co-ordinate of Neighborhood Matrices
$PCoA$	Principal Co-ordinate Analysis
PCR	Polymerase Chain Reaction
SSR	Simple Sequence Repeats
$SSUrRNA$	Small Subunit ribosomal RNA

Chapter 1

Introduction

Given the profound role of microbes in global nutrient and energy cycling (Schimel and Gulledge, 1998), understanding their diversity, dynamics, and interactions in natural habitats is one of the most important and challenging task facing microbial ecologists. In nature, microbes are found in wide range of environments such as the ocean, soil, hydrothermal vents. Additionally, these microbes are also found associated with living things like plants and animals. The interface created by plants and microbes affects the growth and development of the host and also has direct implications on global ecosystem, climate change, and energy availability (Singh et al., 2004). It has become increasingly evident that important phenotypes of plants are directly or indirectly influenced by the structure of their associated microbial community(Mendes et al., 2011). Understanding the mechanisms and factors that govern the community structure and dynamics is thus of great interest.

With the striking decline in cost, Next Generation Sequencing (NGS) technologies have changed the scale of microbial ecological studies and have enabled exhaustive characterization of microbial communities (MacLean et al., 2009). However, determining the real diversity and distinguishing novel or rare organisms from experimental and computational artifacts remains a challenge. It is pivotal that these methods and technologies be verified for accuracy and effectiveness. As a part of my dissertation research, I investigated the nature of potential errors and biases using an in house set of 'synthetic communities', genomic DNA mixes with known concentration from archaea and bacteria with completely sequence genomes (Chapter 2). The results were then implemented in the experimental design and analysis of sequence data in chapter 3, where I characterized bacterial and fungal communities associated with the roots of *Populus deltoides* and then deciphered important factors that are driving the structure of these assemblages. Likewise, in chapter 4, I characterized archaeal communities in roots of *P. deltoides* and then tested for rhizosphere effect on archaea by comparing its community structure to surrounding soils and non *Populus* rhizosphere.

1.1 Background

Microbes in different parts of plants.

Terrestrial plants provide a range of habitats for microorganisms. These habitats can be categorized into three different sections called the phyllosphere, the endosphere, and the rhizosphere. Phyllosphere encompasses surface of above ground plant (Lindow and Leveau, 2002). Endosphere designates area inside the plant (Wilson, 1995). And, rhizosphere is defined as the area surrounding plant roots where plants exert its influence (Hiltner, 1904). Although microorganisms residing in above ground parts of the plant are important, they will not be discussed in detail here, as the focus of this dissertation is on the below ground communities. Regardless, these habitats provide microorganisms with nutrients, energy, protection, and place for attachment. In exchange, most resident microbes provide plant with nutrients, protection against pathogens, and storage (Buée et al., 2009).

Interactions between plants and microbes.

Bacteria residing in rhizosphere contribute towards the health of the host plant. For instance, species of *Bacillus, Erwinia, Pseudomonas, Rhizobia Serratia*, and *Xanthomonas* protect the host plant against soil borne plant diseases by inducing the systemic resistance and competition for substrates (Doornbos et al., 2011). A group of beneficial bacteria known as plant growth promoting rhizobacteria (PGPR) are known to promote growth of the plants and increase its tolerance to stress caused by drought, salinity, and low nutrients (Yang et al., 2009). Similarly, endophytic bacteria posit mutualistic benefit to the host plants by producing antagonistic metabolites against pathogens, secreting phytohormones, and influencing host metabolism (Holland, 1997; Jallow et al., 2004; Schulz and Boyle, 2005; Benhamou et al., 2000). Additionally, bacterial endophytes in *Populus* are known to make it more effective at tolerating and degrading xenobiotic compounds in the soil (van der Lelie et al., 2009). Some endophytes also have direct effect on growth rate of plants (Taghavi et al., 2009).

Fungi associated with roots - both mycorrhizas and endophytes - are capable of performing many functions that are beneficial to their host. For example, fungi provide plants with mineral nutrients that are otherwise inaccessible, mobilize organic forms

of nitrogen and phosphorus, transfer carbons and nutrients between plants through interconnected roots, suppress pathogens and deter herbivores, and degrade simple and complex substrates (Buée et al., 2009; Cardon, 2007).

Plants in turn provide microbes with a place for attachment, energy source in the form of root exudates and control community structure by promoting beneficial microbes and adjusting environmental parameters like pH (Bais et al., 2006).

Microbes in the rhizosphere.

Lorenz Hiltner crowned the word rhizosphere to describe 'soil compartment influenced by the root' in 1904 (Hiltner, 1904), since then a century of research has described the wide diversity of microbes present here. Molecular studies have discovered bacteria, fungi, and more recently archaea from rhizosphere of most terrestrial plants. For instance, molecular studies have reported archaea from rhizosphere of *Lycopersicon esculentum* (tomato) (Simon et al., 2000), *Zea mays* (maize), pine seedlings (Bomberg et al., 2003), *Oryza sativa* (rice) (Lu and Conrad, 2005), and mosses (Sliwinski and Goodman, 2004). Archaea that reside in rhizosphere belong to either *Thaumarchaeota* - phylum encompassing ammonia oxidizers - or *Euryarchaeota*, phylum that comprise of methanogens, with the former dominating in almost all the studied plants. For example, a survey of archaea from the boreal forest trees revealed that 75% of identifiable sequences belonged to *Thaumarchaeota* (Bomberg et al., 2011). Presently, the only archaea from a plant root with complete genome sequence is Candidatus *Nitrosoarchaeum koreensis*, a *Thaumarchaeota* from the rhizosphere of *Caragana sinica* (Chinese Pea Shrub) (Kim et al., 2011).

Diverse groups of bacteria are found associated with rhizosphere of many plants. For instance, 35 taxonomic orders of bacteria were recorded from just 14 plant species (Cardon, 2007). However, at higher taxonomic rank, 10 bacterial phyla (*Acidobacteria, Actinobacteria, Bacteroidetes, Firmicutes, Gemmatimonadetes, Proteobacteria, Verrucomicrobia, Planctomycetes, and Chloroflexi*), can recapitulate most of the diversity in rhizosphere (Peiffer et al., 2013; Lundberg et al., 2012; Uroz et al., 2010). Table 1.1 is a collection of published cultivation-independent surveys of some of highly characterized plants. It is clear that *Proteobacteria*, *Acidobacteria*, and *Actinobacteria* dominate rhizosphere bacterial

communities, a trend observed in mature tree species like oak ($\sim$ 70%) and poplar ($\sim$ 75%) too (Gottel et al., 2011; Uroz et al., 2010).

Fungi are also found in the rhizosphere of most terrestrial plants. Based on the host, these fungi have been generally grouped into six groups: arbuscular, ecto, ericoid, arbutoid, monotropoid, and orchid (Smith and Read, 2008). Arbuscular mycorrhizae, a widespread and abundant fungi, belongs to division *Glomeromycota*. Ectos, a group that comprise of *Ascomycota*, *Basidiomycota*, and *Zygomycota* are only found associated with certain families of woody gymnosperms and angiosperms. Orchids are specific to *Orchidaceae*, and rest of the group is specific to the plant order *Ericales*, all of which are either *Basidiomycota* or *Ascomycota* (Cardon, 2007). A culture-independent surveys of field grown Pine and naturally occurring Poplar tree revealed *Ascomycota* and *Basidiomycota* to be major fungal division in corresponding rhizospheres (Table 1.1).

Microbes in the endosphere.

Anton de Bary, a German botanist of 19th century coined the term endophyte in 1886 to describe 'microorganisms that colonize internal tissues of stems and leaves' (Wilson, 1995). However, now endosphere is a broad term that implies area within a plant, usually between cells in any part of a plant. Endophytes are microbes that reside within plant tissues in both above and below ground parts of plants without causing any harm to the host (Carroll, 1986).In few cases endophytes have been referred conservatively as organisms that are present in a healthy plant at the the time of sampling (Sieber et al., 2002). The dissertation research, however, only focuses on root endosphere.

The root endosphere houses comparatively lower diversity of bacteria and fungi than rhizosphere as shown by both culture based and culture independent methods(Izumi et al., 2008; Gottel et al., 2011). *Proteobacteria* and *Actinobacteria* are the two most dominant bacterial groups in endosphere while *Ascomycota* and *Basidiomycota* are two most dominant fungal groups (Table 1.1). Likewise, in one study, an archaeon (*Euryarchaeota*) was shown to be present in the endosphere of *O. sativa* (rice) (Sun et al., 2008).

Table 1.1: Microbial phyla dominating rhizosphere and endosphere of some common plants

Host Plants	Dominating phyla						Reference
	Bacteria		Fungi		Archaea		
	Rhizosphere	Endosphere	Rhizosphere	Endosphere	Rhizosphere	Endosphere	
Oak (*Quercus sp.*)	*Proteobacteria* *Actinobacteria* *Acidobacteria*	NR	*Ascomycota* *Basidiomycota*	NR	NR	NR	(Uroz et al., 2010) (Jumpponen et al., 2010)
P. deltoides.	*Proteobacteria* *Acidobacteria* *Verrucomicrobia*	*Proteobacteria* *Actinobacteria*	*Ascomycota* *Chytridiomycota* *Zoopagomycotina*	*Ascomycota* *Basidiomycota* *Zoopagomycotina*	NR	NR	(Gottel et al., 2011)
A. thaliana	*Proteobacteria* *Acidobacteria* *Planctomycetes*	*Proteobacteria* *Actinobacteria* *Bacteroidetes*	NR	NR	NR	NR	(Bulgarelli et al., 2012)
Pinus spp.	*Proteobacteria*	NR	*Basidiomycota* *Ascomycota*	NR	*Thaumarchaeota*	NR	(Lottmann et al., 2010) (Bomberg et al., 2011)

NR:Not reported yet.

Factors influencing root microbiota.

Microbial community structure - the diversity and abundance of microbes - is relevant to its ecological functions. Changes in microbial community structure is directly proportional to its functions as observed in human and plant associated microbial communities. For instance, changes in gut microbial community structure have shown to be linked with metabolic disorders Spencer et al. (2011), obesity (Turnbaugh and Gordon, 2009), and Crohn's disease (Eckburg and Relman, 2007) in humans and disease suppresiveness in plants (Mendes et al., 2011). Therefore, to understand consequences of these shifts we need to characterize the structure and identify the factors that are driving it, and by identifying these factors we can better understand the mechanisms that are structuring these complex communities.

Microbial communities in the rhizosphere are influenced by biotic and abiotic factors. Table 1.2 compiles published studies that revealed such factors that affects fungal and bacterial community structure. These factors can be universal, meaning it affects both bacterial and fungal communities, or it can be specific to either one of them. Additionally, their effect can depend on the host plant type and may not equally affect all plants. I did not list factors that influence below ground archaeal community because very little is known about it. For instance, only pH gradient has been shown by multiple studies as a factor that affect archaeal community (Gubry-Rangin et al., 2011).

Biotic factors comprise of plants or plant-related properties like genotype, developmental stage, and species. All of these factors and few others have been shown to shift both bacterial and fugal communities. Bacterial communities can be plant specific (Kuske et al., 2002; Smalla et al., 2001; Costa et al., 2006; Marschner et al., 2004) and within species, it can differ based on genotype (Aira et al., 2010; LeBlanc et al., 2007) , age (Berg et al., 2005; Aira et al., 2010), plant nutrient status (Yang and Crowley, 2000), and pathogen infections (Cardon, 2007; Yang et al., 2001). Fungal communities can be plant and genotype specific, affected by age and infections, but the number of studies and evidences are limited (Table 1.2). Genotype of plants can be directly correlated to its phenotype. For example, different maize cultivars produce different root exudates (Aira et al., 2010), and as a result different maize genotypes house different bacterial communities. However, these type of effects have

rarely been observed for fungal communities (Hannula et al., 2012) and no direct evidence for any relationship with archaeal community and plant genotype has been reported.

Rhizosphere communities are in intimate contact with the soil, thus soil properties are also major players that affect root communities. At broader scale, soil type is an important factor determining the structure of both fungal and bacterial rhizosphere communities (Berg et al., 2005; Buée et al., 2009; Hannula et al., 2012). At finer scale, only bacterial communities in the rhizosphere correlate with individual soil properties like pH (Marschner et al., 2004; Lauber et al., 2009). Direct evidence of linkage between fungal communities and soil chemical properties are still lacking. In addition to physical factors, root communities are also sensitive to short term environmental changes like season (Lottmann et al., 2010; Smalla et al., 2001) and sampling sites (Peiffer et al., 2013; Costa et al., 2006).

Bacterial and fungal endophytes are also influenced by factors similar to the ones that influences rhizosphere communities. Some of the factors that have been shown to influence bacterial endophytes are plant's developmental stage (Mano et al., 2007), geography or location (A, 2001), plant genotype (Sturz et al., 1999), and soil type (Lundberg et al., 2012; Long et al., 2010). However, the information is lacking in the case of fungal endophytes where the only factors that are known to affect the fungal community structure are genotypes related to plant-defense compounds (Saunders and Kohn, 2009) and agricultural amendments (Seghers et al., 2004). To enhance our ability to utilize beneficial potential of endophytes to improve plant growth and health, it is critical to understand the factors that are structuring endosphere communities.

The microbiome of trees

Most of our knowledge about plant microbiota has been derived from studies of agriculturally important plants like maize (Peiffer et al., 2013), rice, and model plants like *Arabidopsis spp.* (Lundberg et al., 2012; Bulgarelli et al., 2012). Besides the clear importance of mature perennial plants in global economy, energy, and environmental health, the knowledge of tree microbiota is lacking. Only a handful of studies have attempted to characterize microbial community structure in trees (Hernesmaa et al., 2005; Uroz et al., 2010; Gottel et al., 2011) and only few have used NGS technologies (Uroz et al., 2010; Gottel et al., 2011).

Table 1.2: Factors that affect rhizosphere microbiota.

Biotic factor	Abiotic factor
Plant developmental stage	**Soil type**
Bacteria	*Bacteria*
Van Overbeek and Van Elsas (2008)	Lundberg et al. (2012)
Fungi	Marschner et al. (2004)
Hannula et al. (2012)	*Fungi*
Mougel et al. (2006)	Mougel et al. (2006)
Plant species	Viebahn et al. (2005)
Bacteria	**Soil heterogeneity**
Kuske et al. (2002)	*Fungi*
Costa et al. (2006)	Viebahn et al. (2005)
Smalla et al. (2001)	**Season**
Marschner et al. (2004)	*Bacteria*
Fungi	Smalla et al. (2001)
Viebahn et al. (2005)	Van Overbeek and Van Elsas (2008)
Plant genotype	*Fungi*
Bacteria	Hannula et al. (2012)
Van Overbeek and Van Elsas (2008)	**Soil pH**
Aira et al. (2010)	*Bacteria*
LeBlanc et al. (2007)	Marschner et al. (2004)
Fungi	**Sampling site**
Hannula et al. (2012)	*Bacteria*
	Costa et al. (2006)
	Peiffer et al. (2013)
	Fungi
	Costa et al. (2006)
	Hannula et al. (2012)

Populus as a model plant system

Populus is an ideal model system to study plant microbe interface due to its extensive root system that houses a diverse set of microbes that includes archaea, bacteria, and fungi. It is also considered a model plant due to its properties like rapid growth rate, prolific sexual reproduction, small genome (approximately 480Mb), and availability of extensive genomic and molecular tools (Bradshaw et al., 2000; Tuskan et al., 2006; Jansson and Douglas, 2007; Podila et al., 2009). Due to its rapid growth rate and abundance in North America, it is used to produce lumber, pulp, and paper. Moreover, it is an effective phytoremediation agent (Hur et al., 2011) and is currently being studied for its use as a feedstock to produce biofuel. These important functional properties of *Populus* is influenced by the microbes associated with the plant. Thus, studying the plant-microbe system in naturally occurring *Populus* provide a premier opportunity to discover interface functions relevant to Department of Energy missions that includes increase plant biomass yield, ecosystem sustainability, disease control, tolerance, and efficient carbon cycle. This dissertation research is a part of larger initiative aimed to understand and characterize plant microbe interface `http://pmi.ornl.gov`.

NGS approach to study microbial ecology.

Since the identification of SSU rRNA gene (especially 16S rRNA) as a universal marker gene to identify and classify microbes, its amplification and sequencing has been of paramount importance in microbial ecology. Based on its universal presence and relatively uniform rate of evolution, SSU rRNA enabled the discovery and classification of a vast diversity of uncultivated microorganisms spanning all phylogenetic levels (DeLong and Pace, 2001; Tringe and Hugenholtz, 2008; Pace, 2009). The discovery coupled with the advent in NGS technologies have allowed exhaustive characterization and derive statistically robust results from the community studies. NGS technologies allow for characterization of microbes at unprecedented level with fraction of the cost, which keeps decreasing every year. The process, however is rather complex and requires many experimental steps, each of which are prone to errors (Engelbrektson et al., 2010; Haas et al., 2011; Hong et al., 2009; Huse et al., 2010; Polz and Cavanaugh, 1998). Biases due to the hypervariable region in SSU

rRNA, primer pair, PCR and sequencing, and data analysis methods can significantly inflate and misrepresent the actual diversity. Therefore, NGS based microbial ecology studies require specific consideration of methodologies and data analysis for correct interpretation of diversity (Werner et al., 2011; Caporaso et al., 2011; Quince et al., 2011; Schloss et al., 2011; Gihring et al., 2011). An ideal way to test for these errors and biases is through direct, quantitative comparisons between known diversity of a controlled 'synthetic community' to that inferred by rRNA gene and metagenomics sequencing (Caporaso et al., 2011; Haas et al., 2011; Morales and Holben, 2009; Schloss et al., 2011). By quantitatively comparing the accuracy of SSU rRNA gene based microbial diversity analyses and metagenomic sequencing using a controlled 'synthetic community' approach one can infer effects of varying experimental procedures and data analysis strategies on diversity characterization.

1.2 Statement of Hypotheses

This dissertation research addresses two areas of microbial ecology. First, I investigated errors and biases in NGS based microbial microbial diversity characterization methods. Second, I applied such microbial diversity characterization approaches to better understand the microbial (archaea, bacteria, and fungi) community structure and its drivers in roots of an economically and ecologically important mature plant in *P. deltoides*. The research will address following specific hypotheses:

Hypothesis 1: *NGS based microbial ecology studies are associated with errors and biases that can be reduced with robust QA/QC methods.*

Previous studies have revealed the presence of artifacts in NGS based methods (Kunin et al., 2009; Polz and Cavanaugh, 1998), so it is important to understand the nature of these errors and biases before starting a NGS based study. By understanding the nature of biases, we can improve experimental designs and data analysis pipelines of these studies to obtain robust and reliable results.

Hypothesis 2: *Bacterial and fungal communities in the roots of native P. deltoides are controlled by soil properties (pH), plant genotype, and seasonal changes.*

Deciphering the factors that are structuring microbial communities in plants are of utmost interest due to their effect on health and development of plants. Based on previous studies conducted on other plants, soil pH, plant genotype, and seasonal change have emerged as major factors that influence microbial communities in roots. So, here I hypothesize pH, genotype, and season to have significant effect on microbial communities of native *P. deltoides*.

Hypothesis 3: *Roots of P. deltoides and mature trees host higher diversity of archaea than surrounding bulk soils.*

Given the potential role of archaea in nitrogen cycle and evidences of direct effect of microbes in health and growth of *P. deltoides*, it is important to characterize their associated archaea. However, archaeal communities in the roots of *P. deltoides* have not been studied before. Thus, fundamental features of archaeal communities in the roots remain unknown. For instance, a general pattern of selection of specific bacteria from surrounding soil in the rhizosphere is observed in many plants, but same is not known for archaea. Moreover, studies that focused on archaea from mature trees are rare. Here, I test if mature trees in general and *P. deltoides* house higher diversity of archaea than corresponding bulk soils. I also checked if below ground archaeal communities are distinct based on their niches.

1.3 Approach

Experiments to test the above listed hypotheses can be categorized into three parts. First, by assembling 'synthetic communities' with known quantity of genomic DNA from sequenced archaea and bacteria, I tested for errors and biases in NGS based microbial community characterization (SSU rRNA and metagenomics) methods. Second, based partly on findings from previous chapter, a meticulous experimental design and data analysis pipeline that reduced such errors was implemented to test for soil properties (pH), plant genotype, and seasons influence on microbial community structure of *P. deltoides*. Additionally, I also characterized the taxonomy and structure of the resident bacterial and fungal communities. Third, a SSU rDNA and *amoA* based community characterization of archaeal communities

from a subset of our *P. deltoides* samples and surrounding bulk soils was conducted to test for rhizosphere effect.

Chapter 2: Test the efficacy of NGS based microbial (archaeal and bacterial) diversity characterization methods.

There are clear evidences of errors and biases in NGS based microbial community characterization approaches (Kunin et al., 2009; Polz and Cavanaugh, 1998). Therefore, My first step was to understand these errors and biases and use the findings to reduce it during community analysis. Most of the microbial community that are characterized have little to no prior information about diversity and abundance of its community which makes the community characterization method hard to test. Therefore, I assembled 'synthetic communities' with known amount of gDNA from archaeal and bacterial species that have complete genome sequences. This chapter details the construction of three 'synthetic communities' and testing of different experimental and computational techniques using it.

Chapter 3: Characterize *P. deltoides'* root microbiota and elucidate the factors that are structuring it.

Bacterial and fungal communities associated with plant roots - in rhizosphere and endosphere - are central to its health, survival and growth. However, a robust understanding of the factors that shape root microbiota composition and structure has remained elusive. Here, we set out to investigate relationships of bacterial and fungal communities in rhizosphere soils and endosphere of the riparian tree species *Populus deltoides*, with soil parameters, environmental properties (host phenotype and aboveground environmental settings), host plant genotype (Simple Sequence Repeat (SSR) markers), season (Spring vs. Fall), and geographic setting (at scales from regional watersheds to local riparian zones).

Chapter 4: Compare archaeal communities in *P. deltoides* and mature trees' rhizosphere with surrounding bulk soils.

Archaea constitute one of the major organisms in below ground - soil and rhizosphere - microbial communities. In this chapter, I characterized archaeal communities in roots of *P.*

deltoides, other mature trees, and bulk soils from the riparian zone along Caney Fork river in Tennessee using V1-V3 region of 16S rRNA gene and partial region of ammonia oxidizing subunit A gene (*amoA*), a key functional gene for ammonia oxidation. The chapter details two parts of the study. Comparison of archaeal diversity in *P. deltoides* and mature trees with surrounding bulk soil and characterization of archaeal communities in *P. deltoides*, surrounding bulk soil, and non *Populus* trees.

1.4 Significance

Recent advances in the microbial ecology is directly proportional to the advent of sequencing technologies and experimental and computational methodologies. Especially, amplicon based sequencing of phylogenetic marker genes like 16S rRNA and metagenomics have contributed to discovery of novel microbes and functions. However, these approaches are not always accurate and are known to inflate the actual diversity (Sogin et al., 2006; Kunin et al., 2009; Huse et al., 2010). In environmental data sets, distinguishing rare but real OTUs or metagenomic signatures of uncultured taxa from experimental and computational artifacts remains a challenge. Diverse 'synthetic communities' and validation data sets such as the ones presented in chapter 2 enables direct comparison of sequencing, data processing accuracy and effectiveness in sequence binning and assembly for representing the true environmental microbial composition. Although the study mainly focused on 454 sequencing sequencing platform for amplicon sequencing, Illumina is now widely used for amplicon sequencing (Caporaso et al., 2011) as well. Additionally, new sequencing platforms like Ion-Torrent are constantly being introduced in the market. By having access to a mix of microorganisms with all attributes known, we can test new approaches, platforms, and computational tools like AmpliconNoise (Quince et al., 2011), Denoiser (Reeder and Knight, 2010), and Accacia (Bragg et al., 2012) before using it in the actual experiment.

The significance of studying plant microbe systems in naturally occurring perennial plants are two folds. First, by characterizing communities of resident microbes and their major driving factors, we can better understand the plant microbe interface to manipulate and optimize the interface for relevant ecological and environmental functions. Microbes residing in plant roots play an important roles in growth, development, health, and ecological

fitness of the host plants through specific functions like antibiotic production, geochemical cycling of minerals, processing otherwise inaccessible nutrients, and many others (Buée et al., 2009). For better understanding of beneficial functions due to resident microbes or their community, it is important that we characterize the microbes and their community structure. Second, studying microbial systems in naturally occurring perennial plants can lead to discoveries that are relevant to DOE missions. In this regard *Populus* is an ideal plant. It is a model woody plant, a leading candidate for bioenergy application, and widely distributed in North America. Since microbial resident of roots have significant impact on health and proliferation of *Populus*, characterizing the diversity and abundance of resident microbes and deciphering factors that affect them are pivotal for ultimate use of indigenous or engineered *Populus* to increase biomass yield, efficient environmental remediation, and carbon cycling.

Chapter 2

Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities

Disclosure: This chapter was published as:

Shakya, M., Quince, C., Campbell, J.H., Yang, Z.K., Schadt, C.W., and Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology.*

Migun Shakya's contributions include DNA extraction, DNA quantification, assemblage of synthetic communities, sample preparation for 454 sequencing, sequencing, analyzing data, and writing the manuscript as primary author.

2.1 Abstract

Next generation sequencing has dramatically changed the landscape of microbial ecology, large-scale and in-depth diversity studies being now widely accessible. However, determining the accuracy of taxonomic and quantitative inferences and comparing results obtained with different approaches are complicated by incongruence of experimental and computational data types and also by lack of knowledge of the true ecological diversity. Here we used highly diverse bacterial and archaeal synthetic communities assembled from pure genomic DNAs to compare inferences from metagenomic and SSU rRNA amplicon sequencing. Both Illumina and 454 metagenomic data outperformed amplicon sequencing in quantifying the community composition, but the outcome was dependent on analysis parameters and platform. New approaches in processing and classifying amplicons can reconstruct the taxonomic composition of the community with high reproducibility within primer sets, but all tested primers sets lead to significant taxon-specific biases. Controlled synthetic communities assembled to broadly mimic the phylogenetic richness in target environments can provide important validation for fine-tuning experimental and computational parameters used to characterize natural communities.

2.2 Introduction

For over two decades, amplification and sequencing of the small subunit ribosomal RNA (SSU rRNA or 16S rRNA) gene has been the primary approach to assess the abundance and taxonomic identity of microbes in environment. Based on its universal presence and relatively uniform rate of evolution, SSU rRNA enabled the discovery and classification of a vast diversity of uncultivated microorganisms spanning all phylogenetic levels (DeLong and Pace, 2001; Tringe and Hugenholtz, 2008; Pace, 2009). With increasing sequencing depth and throughput, statistically robust quantitative comparisons between communities have become feasible. Direct, metagenomic sequencing of the community DNA pool complements rRNA gene-based characterization by providing insights into physiological potentials and expanding phylogenetic diversity characterization into protein sequence space (Tringe and Rubin, 2005) while metatranscriptomics and metaproteomics offers a direct

access to community physiology (McCarren et al., 2010; VerBerkmoes et al., 2009). A variety of experimental and data analysis alternatives have been developed to allow in-depth characterization and large-scale comparative studies of complex microbial communities (Schloss et al., 2011; Sun et al., 2012). Each experimental and computational step in diversity characterization is prone, however, to errors (Engelbrektson et al., 2010; Haas et al., 2011; Hong et al., 2009; Huse et al., 2010; Polz and Cavanaugh, 1998). Amplification of different hypervariable rRNA gene regions can lead to inconsistent taxonomic coverage and incongruence between datasets. In addition, short read amplicon sequencing requires specific considerations for the methodology, data analysis and interpretation of microbial diversity (Werner et al., 2011; Caporaso et al., 2011; Quince et al., 2011; Schloss et al., 2011; Gihring et al., 2011).

Metagenomic sequencing avoids some of the limitations of rRNA amplicon sequencing by directly accessing the community genomic information. Diversity interpretation is however, complicated by uncertainties in assigning genes to specific organisms (especially for taxa with no cultured representatives) and by bias introduced during sequencing Gomez-Alvarez et al. (2009). Direct, quantitative comparisons between known diversity and that inferred by rRNA gene and metagenomic data are scarce (Caporaso et al., 2011; Haas et al., 2011; Morales and Holben, 2009; Schloss et al., 2011) and have been limited in taxonomic coverage.

Here we quantitatively compared the accuracy of SSU rRNA gene based microbial diversity analyses with metagenomic sequencing using a controlled synthetic community approach. The communities consisted of laboratory-mixed microbial genomic DNAs (gDNA) of known sequence, representing a broad diversity of bacteria and archaea. Species from nearly all phyla with cultured representatives were included, covering a wide range of genetic variation at different taxonomic levels and spanning the full spectrum of genome sizes, (G+C)% (GC) content, genomic divergence, and rRNA operon copy numbers. The effects of varying experimental procedures and data analysis strategies on rRNA based diversity and composition were compared with those determined using two metagenomic sequencing platforms, 454 and Illumina.

2.3 Methods

Collection of gDNA for the synthetic communities

Three distinct synthetic communities with gDNAs from representatives of 17 bacterial and 5 archaeal phyla were assembled. Except for four bacteria that have genomes in high quality draft stage (*Sulfurihydrogenibium yellowstonense SS-5, Sulfitobacter sp. EE-36, Sulfitobacter sp.NAS-14.1 and Desulfovibrio piger*), all other species and strains included in the study have their genome sequences closed. Pure cultures of 27 archaea and bacteria were grown as part of this study in liquid using stocks from ATCC (American Type Culture Collection), DSMZ (Deutsche Sammlung von Mikroorganismen) or from collaborators, using the published media and conditions for each organism. High molecular weight DNA was extracted using a mechanical and organic cell lysis method as described in Ley et al. (2008), dissolved in TE buffer (pH 8) and measured spectrophotometrically for quality and concentration. For 37 archaea and bacteria we received either purified gDNA or cell cultures from collaborators (Table 2.1 and Table 2.2), from which we extracted the gDNA. All gDNA solutions were stored in nuclease-free sylanized tubes (Ambion, Austin, TX), to minimize loss by adsorption to tube walls.

DNA quantification and assembly of synthetic communities

Three different methods were used to determine the quality and concentration of each gDNA. The initial concentration of each gDNA preparation was measured by fluorescence assay against a set of standards using a Qubit 2.0 fluorometer (Invitrogen, Carlsbad CA). For an estimation of the molecular weight, approximately 50 ng DNA was separated and visualized on 1.2% agarose E-gels (Invitrogen) with a set of lambda phage DNA mass standards (10-100 ng). All DNAs used in the assembly of the synthetic communities had average molecular weight exceeding that of the lambda phage and no RNA or small, degraded nucleic acids were detected. Because the DNAs were isolated from very diverse organisms, grown in different media and potentially still contained molecules that could interfere with accurate fluorescence and gel quantification, we used generalized quantitative PCR (qPCR) assays for Bacteria (Fierer et al., 2005) and Archaea (Reysenbach et al., 2006) to guide assembly of the synthetic communities. For each organism to be represented in

the community, qPCR was performed on its purified DNA with either the archaeal or the bacterial primer pair. Sequences of SSU rRNA genes from each organism were screened against the published primer sequences (Eub338-Eub518 and Arc915f-Arc1059r) in silico prior to performing qPCRs. To broaden the specificity of the primers so that the SSU rRNA genes of all the species targeted for inclusion in the synthetic community could be amplified, we modified both forward primers Eub338 and Arc915f (see below). DNA SYBR Green qPCR assays (20μl) were performed in a Bio-Rad CFX96TM (Hercules, CA) thermal cycler using primers synthesized by IDT (Coralville, IA) and Eurofins MWG Operon (Huntsville, AL) and Bio-Rad iQ Supermix. Archaeal assays used primers arc915fmc (5'-AGGAATTGGCGGGRGRGCAC -3') and arc1059r (5'-GCCATGCACCWCCTCT-3') at a final concentration of 350 nM each. Cycling parameters included an initial denaturation at 95°C for 5 min followed by 45 amplification cycles of 95°C for 30 sec, 61°C for 30 sec, 72°C for 1 min and a fluorescence reading. Following amplification cycles, products were denatured at 95°C for 10 sec, and a melt curve was determined over a range of 60-95°C . Standard curves were constructed using *Methanococcus maripaludis S900* genomic DNA diluted from 1×10^7 - 1×10^2 SSU rRNA gene copies per reaction. Bacterial assays used primers Eub338mc (5'-ACTCCTACGGGDGGCWGCAG-3') and Eub518 (5'-ATTACCGCGGCTGCTGG-3') at a final concentration of 500 nM each. Cycling parameters included an initial denaturation of 95°C for 5 min followed by 45 amplification cycles of 95°C for 30 sec, 53°C for 30 sec, 72°C for 1 min and a fluorescence reading. Following amplification cycles, products were denatured at 95°C for 10 sec, and a melt curve was determined over a range of 50-95°C . Standard curves were constructed using *Escherichia coli K12* genomic DNA diluted from 1×10^8 - 1×10^2 SSU rRNA gene copies per reaction After individual organism DNA quantification, in order to achieve a diverse community composition in both taxonomic distribution and abundance we mixed individual gDNAs, obtaining two primary synthetic communities (a bacterial and an archaeal one). The organisms for which we had low amounts of gDNA were represented at lower abundances in the final mix. The genomic abundance for each organism in the two communities was calculated based on the qPCR-determined concentration and the known number of rRNA operons present in each genome (1-10 copies; (Table 2.1 and Table 2.2)). To obtain the Archaea-Bacteria community, aliquots of the two

were mixed and the individual genomic abundances were calculated based on those in the primary communities.

Metagenomic sequencing and analysis

Two metagenomic libraries were constructed for sequencing using the 454 and Illumina platforms. For 454 sequencing, 50 ng of the Archaea-Bacteria synthetic community gDNA was used to prepare an FLX Titanium compatible library using a Nextera^{TM} DNA sample prep kit (Epicentre Biotechnologies, Madison WI) and following manufacturers instructions. Briefly, the DNA was fragmented ("tagmented") using the transposase enzyme mix and purified. 454 sequencing primers, a bar-coded Titanium Adaptor 1 (MID3: AGACGCACTC), were incorporated using 15 cycles of PCR followed by purification and size distribution analysis on a an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany). Insert sizes varied between 500 and 1500 nt. The library was unidirectionally sequenced in-house on one fourth of an FLX Titanium sequencing plate using standard 454/Roche reagents and protocols. The 454.sff sequence file was loaded into the CLC Genomics Workbench 4.8 (CLCBio, Cambridge MA). Low quality reads (limit =0.05), ambiguous nucleotides, 454 and Nextera adaptors were removed and any further remaining reads shorter than 20 nt were discarded. The resulting dataset contained 291,146 reads with an average length of 320 nt, totaling 85.5Mbb. The sequences were mapped to a database containing the 64 reference genomes (combined total length of 205.6 Mbp) using the CLC local aligner algorithm, with a similarity threshold of 0.9, length fraction of 0.5 and default mismatch/indel cost values. The average coverage of the metagenome was 0.39 fold, with 261,385 reads mapped the genomes. A breakdown of the number of reads mapped to each genome and their coverage is shown in Table 2.1 and Table 2.2. The number of reads mapped to each genome was used to calculate the coverage distribution relative to expected values, taking also in account the variable genome size among the represented organisms. Unmapped reads were further analyzed for mapping either to known plasmids of the included organisms (32 plasmids totaling 4.2 Mbp, ranging in size from 3.6 kbp for a *Caldicellulosiruptor bescii* plasmid to >635 kbp for a *Haloferax volcanii* megaplasmid) or back to the reference genomes by decreasing

the similarity threshold in order to accommodate unfiltered sequence artifacts. A total of 5,455 reads mapped to plasmids, reaching the same average coverage obtained for the genomic component (0.39 fold). The identity of reads that did not map to either genomes or plasmids was not further explored but likely include both reads with a higher mutations or sequencing errors frequency and reads that belong to a *Clostridium sp.* contaminant identified in the *Desulfovibrio vulgaris* culture, for which a genome sequence is not available. For Illumina sequencing, 1μg of the Archaea-Bacteria synthetic community gDNA was physically sheared by Covaris Inc. (Woburn, MA), to an average fragment size of 250bp. The fragmented DNA was sequenced bi-directionally (100 bp each direction) on a lane of Illumina HighSeq 2000 using V3 sequencing reagents at the Genome Sciences Resource Center of Vanderbilt University (Nashville, TN). Read quality was analyzed using FastQC (Brabahan Bioinformatics). Filtering out sequences shorter than 50 nt, removal of low-quality reads and of those with ambiguous nucleotides in CLC Genomics Workbench 4.8 resulted in two datasets (forward-reverse reads) of >53.5 and >53.7 million reads, respectively, with an average length of 100 nt and totaling over 10.7 Gbp. Mapping reads to reference genomes with CLC Genomics Workbench 4.8 followed the same approach except that a higher sequence fraction match (0.8) was used as threshold. Over 96 million reads were mapped, achieving an average 46-fold coverage across the metagenome (1,500-fold maximum region coverage), with many genomes being covered over >95% of their length (Table 2.1 and Table 2.2). Two million reads mapped to the 32 known plasmids, with some regions reaching >1,000-fold coverage (average 50-fold). An accuracy of detection ratio for each species within each sample was calculated by dividing the fraction of its sequences in the metagenomic dataset by its known abundance (Q-PCR-based), normalized to genome size, within the corresponding synthetic community. A matrix containing species accuracy detection within each sample sequenced was constructed relative to the standard Q-PCR-based estimates (always = 1), with separate analyses performed for Archaeal and Bacterial data. PRIMER-E v6 (Clarke and Warwick, 1994) was used to calculate Bray-Curtis resemblance matrices for each dataset. These matrices were used to generate Principal Coordinate Analysis (PCoA) plots and hierarchical clustering dendrograms to visualize reproducibility of replicates and accuracy of community representation (based upon Q-PCR) for each amplicon region and sequencing strategy.

Comparison of bias due to GC content

The bias in metagenomic coverage on 454 and Illumina platforms was calculated across the range of genomic GC content of synthetic community constituents (R v2.14; stats package). For each genome, the Illumina accuracy factor was subtracted from the corresponding 454 values, and the resulting difference was regressed against the genomic GC content. Matched, pairwise t-tests were used to compare these accuracy differences between the sequencing platforms across the GC spectrum in three window intervals (27-40%, 40-60% and 60-70%). To determine the coverage bias across the metagenome, we analyzed the 454 and Illumina reads coverage for each genome in the community separately. We did not observe coverage fluctuations linked to genome size. However, the intra-genome sequence coverage matches what we observed at the level of the community with local GC content having a strong influence on the number of reads depending on the sequencing platform.

MEGAN Analysis

The available genomes of all Archaea and Bacteria were downloaded from the NCBI ftp site (`ftp://ftp.ncbi.nih.gov/genomes/`). We created three different blastable datasets with those genomes. First, a database that exclusively contained genomes of the synthetic community organisms (REF); second, a database that contains the genome of all organisms, including the genomes of synthetic community organisms (ALL); third, a database that excludes the members of the synthetic community, but includes all other organisms (X Reference). We used megablast (-v 1 -b 1 -a 10 -m 7) with either 454 or Illumina metagenomic sequences against these three databases. Additionally, we also used the less stringent blastn (-v 1 -b 1 -a 10 -m 7) against the X Reference database. We analyzed and quantified the taxonomic abundance of the community of each blast results using MEGAN (MEtaGenome ANalyzer) (Huson et al., 2007). For combined visualization of the result, ratios between the known composition and those determined for each database type and blast approach were displayed as a heat map and included species, genus and family taxonomic levels (Fig 2.10). The sequences known to belong to individual genomes (based on CLC-Bio genome mapping) were identified in terms of their predicted taxonomic affiliation

by blast-MEGAN and the distribution was projected as histograms for each individual genome (Fig 2.10 B, C) or globally (Fig 2.10 D).

IMG-M and MG-RAST analysis

To submit the 454 data into IMG/M we used a fasta format file containing the CLC-Bio quality filtered reads. Data processing used default parameters including gene prediction and functional annotation. For MG-RAST v3 analysis we uploaded both the 454 sff file and the Illumina fastq data files. Quality filtering and sequence analysis followed the default MG-RAST pipeline flow. To analyze the taxonomic composition of the community based on IMG-M and MG-RAST we extracted the inferred abundance at phylum level for reach dataset and also the number of taxonomic units predicted by both systems. MG-RAST enables changing cutoff parameters for the taxonomic mapping and we explored the effect those changes have on the types and numbers of predicted taxa (Table 2.3). To evaluate the community composition based on SSU rRNA sequences present in the metagenomes we extracted the sequences assigned to that gene from both IMG-M and MG-RAST and analyzed their affiliation using the RDP Classifier. The metagenomes are publicly available in those systems for further analyses. The raw data files have been deposited in the NCBI Sequence Read Archive (Accession # SRA059004). Sequences from various filtering stages are also available from the authors upon request.

PCR amplification and 454 sequencing of SSU rRNA amplicons

Sets of amplification primers were chosen to cover most of the hypervariable regions of SSU rRNA (Table 2.3) (Lane, 1991; Weisburg et al., 1991; Muyzer et al., 1996; Nübel et al., 1996; Suzuki and Giovannoni, 1996; Ovreås et al., 1997; Takai and Horikoshi, 2000; Watanabe et al., 2001; Baker et al., 2003; McCutcheon and Moran, 2007; Frank et al., 2008; Bates et al., 2011). Some of the primers were modified from their original published sequence or additional variants were added to broaden their taxonomic coverage. Still, some primer-rRNA gene mismatches to some of the species represented in the synthetic communities remained, allowing estimation of their effects on amplification efficiency (Figures 2.11 and 2.12). Amplification primers targeting bacterial V4, V12 and archaeal V4

regions were designed with FLX adapters and rest of the primers were designed with FLX Titanium adapters. To allow multiplexing, the sequencing primers contained 6-8 nt long barcodes. The primers were synthesized by IDT (Coralville, IA) and Eurofins MWG Operon (Huntsville, AL) and were HPLC or HPSF purified. Polymerase chain reaction (PCR) was performed in 50μl reactions with 1X High Fidelity PCR buffer (Invitrogen, Carlsbad CA), 2 mM MgSO4, 300 nM of each primer, 200 mM dNTPs, and 1 unit of Platinum Taq DNA Polymerase High Fidelity (Invitrogen; Carlsbad, CA). Between 2.5-10 ng template gDNA was used for the different synthetic communities. All reactions were performed in duplicate or triplicate, and separate reactions with different number of cycles, annealing temperature, and different polymerases were also conducted. The range of amplicon lengths obtained with each primer pair, based on the genomic sequences, is shown in Table 2.4 and a summary of amplification parameters is shown in Table S5. For the Bacteria-Archaea community, three different polymerases, TaqHiFi (Invitrogen), High GC (Roche Diagnostics, Indianapolis, IN) and Accuprime Pfx (Invitrogen) were used to compare effects of polymerase fidelity and annealing specificity on resulting sequences. Amplicons were purified using AMPure paramagnetic beads (Agencourt Biosciences Corporation, Beverly, MA) followed by concentration and size analysis using DNA 1000 chips on an Agilent 2100 Bioanalyzer (Agilent Technologies, Waldbronn, Germany). Amplicon libraries were then prepared for unidirectional sequencing using the emPCR Kit II (Roche) followed by sequencing on a 454 FLX Life Sciences Genome Sequencer (Roche Diagnostics, Indianapolis, IN). Pyrosequencing using the FLX chemistry and Titanium chemistry was done according to manufacturers instructions.

Amplicon Sequence Processing

For SSU rRNA amplicon sequence data processing we used primarily the software packages mothur (v1.1639)(Schloss et al., 2009), QIIME (v1.3.040)(Caporaso et al., 2010a), RDP(Cole et al., 2009), and AmpliconNoise v1.25(Quince et al., 2011). Sequences were processed using different filtering parameters for respective analyses. For Low Quality filtering (LQ), sequences were removed from the analysis if they were <200 nt, had ambiguous bases, had a non-exact barcode match, or showed more than two mismatches

for the amplification primer. Quality score was not used for this filtering. Remaining sequences were assigned to samples based on the barcode matches, trimmed and reads that were sequenced from the reverse end were reverse complemented so that all sequences begin with the 5 end of the amplicon. Potential chimeras were identified using mothur implementation of ChimeraSlayer. Reference sequences were aligned against a bacterial or archaeal SILVA database using the Needleman-Wunsch algorithm in mothur. The aligned reference sequence was then used as the template for flagging chimeric sequences. For High Quality filtering (HQ), sequences were removed from the analysis if they were <200 nt, had ambiguous bases, had a non-exact barcode and primer match, and had a homopolymer >9 nt. If a sequence quality score fell below 20 for a 50-nt window, then it was trimmed at previous position where the average quality score >20. Similarly, sequences were binned to corresponding sample based on corresponding barcodes and reverse complemented. Potential chimeras were identified using ChimeraSlayer. Reference sequences were aligned against the Greengenes database (greengenes.lbl.gov) using Pynast. The aligned reference sequences were then used as template for flagging chimerical sequences. For sequences that were filtered using AmpliconNoise (AN), all samples were run through the AmpliconNoise pipeline, which consists of removal of both sequencing and PCR errors and removal of chimeras using its in built Perseus algorithm(Quince et al., 2011). AmpliconNoise analysis consists of two stages, PyroNoise removal of 454 errors, and SeqNoise removal of PCR single base misincorporations. We have, therefore, estimated the proportion of errors attributable to these two sources by calculating the reduction in error rate after applying each algorithm (Table 2.5). Raw, per-base error rates varied from 0.1-0.25% for FLX and 0.15-0.9% for Titanium chemistry. For FLX, the V12 region (0.25%) was associated with a higher raw rate than V4 (0.1%). For Titanium, higher error rates are associated with V13 (0.8%) region, mostly due to PCR chimeras. Both 454 sequencing errors and PCR errors are responsible for around 0.05% of the overall error rate each, but there is some variation between regions. The V6 region appears particularly prone to PCR noise with a 0.1% error rate attributable to this source, and the V13 region has a higher rate of 454 errors (0.1%). Frequencies of chimeras were also calculated as a percentage of unique types of sequences following noise removal by AmpliconNoise. This method (Quince et al., 2011) was used to classify sequences as good, chimeric, or trimeric by direct comparison with databases composed of

the corresponding region extracted from genome sequences. These results confirmed those from ChimeraSlayer, and all the shorter FLX amplicons chimera frequencies were low among erroneous reads (<7%). However, Titanium sequences showed higher frequencies for V13 (>60%) than V35 (10%) or V69 (<5%), and a twofold reduction in frequency for V13 was observed when the cycle number was reduced to 24. For downstream OTU analysis, except for the sequences that were denoised using AmpliconNoise, sequences were then trimmed so that all sequences began and ended at the same coordinates. Sequences were aligned in mothur against the SILVA database and trimmed at the same alignment position. The position for trimming was manually selected to conserve number of sequences per sample and also have an approximate average length of 200 nt for FLX and 400 nt for Titanium amplicons (Table 2.4). An example of mothur batch file for each amplicon that was used to trim sequences is included. A summary of number of raw reads and processed reads is shown in Table 2.5.

OTU Diversity Analysis

On both the LQ and HQ datasets we applied three different clustering algorithms as implemented by RDP, mothur, and ESPRIT/SLP. For the RDP-based analysis, sequences were aligned using the secondary-structure-aware Infernal aligner and clustered using complete-linkage clustering. In mothur based OTU analysis, trimmed sequences were aligned against the SILVA database using Needleman-Wunsch alignment, pre-clustered using the mothur implementation of single-linkage pre-clustering algorithm from Huse et al. (2010) and clustered using average linkage clustering. Batch files that list the commands that were used to cluster the sequences are provided. For the SLP-PW/AL analysis, trimmed sequences were aligned using the pairwise alignment algorithm in ESPRIT, pre-clustered using the single linkage script from Huse et al. (2010) and clustered based on pairwise distances using average-linkage clustering in mothur. A shell script was used to generate pairwise distance and cluster sequences (`http://alrlab.research.pdx.edu/aquificales/pyrosequencing.html`). To identify which of the clusters at different distance levels corresponded to which taxa (strain, species, genus etc.) in the synthetic community, the trimmed reference sequences were also clustered with pyrosequence

data. Sequences that did not co-cluster with reference sequences were analyzed for potential mutations, chimeras and by taxonomic affiliation to identify potential unexpected contaminants. Sequences that were denoised using the AmpliconNoise pipeline were clustered based on the distance matrices generated as a result of pairwise alignment similar to ESPRIT package. Average linkage clustering was implemented in this case, essentially as described (Quince et al., 2011).

Taxonomic diversity analysis

Because each sequence in the dataset should correspond to a SSU rRNA gene sequence from the represented genomes, accuracy of SSU rDNA-based diversity estimation was also investigated directly by matching pyrosequence data to references and comparing observed diversity and abundance with those known based on the assembly of each synthetic community. Each processed amplicon dataset was top hit matched to a corresponding reference database by Megablast. As few as single nucleotide differences were sufficient for accurate matching to the corresponding reference sequence, as determined empirically. For some closely related strains or species, however some of the SSU rRNA region amplicons were 100% identical (Figure 2.13) and assignment to a specific organism in the community was not possible. In those cases, the numbers of hits to the group were assigned to the organisms based on their Q-PCR-based representation. Two pairs of organisms could not be discriminated with any amplicon (*Pyrococcus furiosus - P. horikoshii and Sulfitobacter sp. EE-36 - Sulfitobacter sp. NAS-14.1*), pointing to limitations of the SSU rRNA gene for comprehensive diversity estimation. For each amplicon dataset, we calculated a ratio between the observed reads-based abundance of each organism and that known based on qPCR-guided community assembly.

2.4 Results and discussion

Synthetic archaeal and bacterial community characteristics

By combining known amounts of purified gDNAs we constructed two diverse synthetic communities representing the domains Archaea and Bacteria, respectively. These communities included most phyla with cultured representatives, as well as contained closely related species and strain pairs. All included organisms have complete or high quality draft genomic sequences. Sixteen members of *Crenarchaeota, Euryarchaeota and Nanoarchaeota* represented Archaea, while Bacteria included 48 organisms from 18 phyla (Table 2.1 and Table 2.2). The organisms covered a wide variety of metabolic strategies and adaptations to the human body, marine and terrestrial aquatic environments, soils and the subsurface, and extreme physical or chemical conditions. Unlike environmental communities, each gDNA was individually purified and quantified prior to being mixed with others, thus true community composition was known, and extraction-based biases were eliminated. The genomes span a broad range of GC content (27-70%), sizes (0.5-10 Mbp) and rRNA operon number (1-10). Based on these known parameters and quantification by Q-PCR, we validated the representation (cell equivalents) of each species. The Archaea community contained one dominant species (*Nanoarchaeum equitans*, 30% of genome copies), with the others present at abundances between 1-10%. Due to differences in genome size and rRNA gene copy number, the actual contribution of individual organisms to the metagenome complexity spanned a 20-fold range (e.g. *N. equitans*, with a genome of 0.5 Mbp represented 10% of the metagenome). The Bacteria community contained no single dominant organism however there was a 25-fold variation in the gDNA abundance among the individual taxa. For several tests, we also combined them into a 64-member Archaea-Bacteria (AB) community in which the genomic abundance of taxa spanned a 200-fold range, from 0.05-1% (36 taxa), 1-5 (23 taxa) to 5-8% (6 taxa). These communities were not aimed at reproducing any specific type of natural diversity, but to broadly represent the phylogenetic and genomic heterogeneity within Bacteria and Archaea that is often encountered in complex community assessments. Many communities contain a vastly greater number of taxa at all levels (e.g. soil communities) or are scarcer in number of high taxa but much more diverse at genus level and below (e.g. human gut microbiota). However, the synthetic community used here,

by combining taxa adapted to many different types of environments, should provide a good initial assessment of the power of taxonomic and quantitative determinations to be expected when using a broad range of natural samples.

Metagenomic characterization of the synthetic Archaea-Bacteria community

For metagenomic 454 and Illumina sequencing we used the AB community as it contained the broadest variation in phylogenetic diversity and genomic characteristics. These two platforms differ in output and data characteristics but both have been widely used for environmental sequencing (Hess et al., 2011; McCarren et al., 2010)

The 454 sequencing library was generated using Nextera^{TM} in vitro transposition (Caruccio, 2011), which has low DNA, input requirement compared to physical shearing methods (50 ng vs. >1 μg). Analyses using single organism libraries have shown relatively similar coverage in such libraries compared to ones obtained by shearing (Adey et al., 2010). Because in many environmental studies the availability of DNA is limited, determining the accuracy of community composition inferences using metagenomic sequencing of such samples is important. A recent study demonstrated that extensive, phi29 polymerase amplification significantly alters the composition of metagenomic libraries (Yilmaz et al., 2010). While the transposition-generated library requires only mild amplification, a bias risk remains and was therefore investigated. The Nextera^{TM} AB library was sequenced on one quarter of a 454 FLX Titanium plate and generated 2.9 $\times$ 10^5 reads (85.5 Mbp of sequence). For the Illumina platform, a standard sheared library was constructed and bidirectional sequencing on one lane resulted in 107 million reads (>10 Gbp of sequence).

Using local alignment, >97% of the 454 data (83.5 Mbp of sequence) was mapped to the 64 reference genomes (205.6 Mbp total length) and 2% to plasmids of those organisms (32 plasmids, totaling 4.2 Mbp), reaching average metagenome coverage of 0.39 fold. Similarly, over 92% of the Illumina reads were mapped to the reference genomes and plasmids, achieving 46-50 fold metagenome coverage. The combination of a 200-fold range variation in individual genome abundance with the 20-fold variation in genome size generated distinct sequence frequency distributions for the different members of the community. Consequently,

the observed average fold coverage of individual genomes by mapped reads ranged from 0.01-1.3 for the 454 data to 6-300 for Illumina (Table 2.1 and Table 2.2). With Illumina, 53 of the 64 organisms had coverage overlaying >95% of their determined chromosomal sequences.

To determine how accurately the metagenomic data described the actual community composition, we compared the expected representation of each species based on qPCR quantification with the observed coverage. Overall, both metagenomic sequence sets described the community genomic composition remarkably well, with 70% (454) and 78% (Illumina) of the individual species/strains estimated within a factor of two-fold or less from their actual abundance Figure 2.1. Importantly, both sequencing platforms appear relatively unaffected by the abundance of individual genomes spanning two orders of magnitude variation. However, the 454 metagenome showed a measurable bias towards oversampling of genomes with low GC (<40%) and under sampling for those with high GC values (>60%), including in-depth of coverage across the individual genomes (Figure 2.2 and Figure 2.3). Potentially, such bias could be attributed to enzymatic steps in library construction. In comparison, the Illumina metagenome was less influenced by GC content, with abundances of most of the individual genomes <2-fold of their expected levels and with GC-based coverage better tracking the actual metagenome composition than 454. Pairwise t-tests confirmed a higher differential GC-linked bias in low-GC organisms (27-40% GC; p=0.027) and lower in high-GC organism (40-70% GC; p=2.7e-05) for 454 data, while representation of mid-range GC organisms (40-60%) was not significantly different between 454 and Illumina (p=0.170). The abundance of some thermophilic taxa *(Pyrococcus, Dictyoglomus, Sulfurihydrogenibium)* was overestimated by both platforms (2-5 fold). This may be linked to extensive regions of very low GC in their genomes, which displayed an inflated coverage with both platforms, although other sources of bias could be involved. However, when we analyzed each individual genome in the community in terms of sequence coverage, the only bias detected was linked to local GC content. A comparison of the intra- genome coverage with both 454 and Illumina for representative genomes with different GC content is presented in Figure 2.9 (similar plots were obtained for the other genomes of the community and also for genomes we sequenced independently).

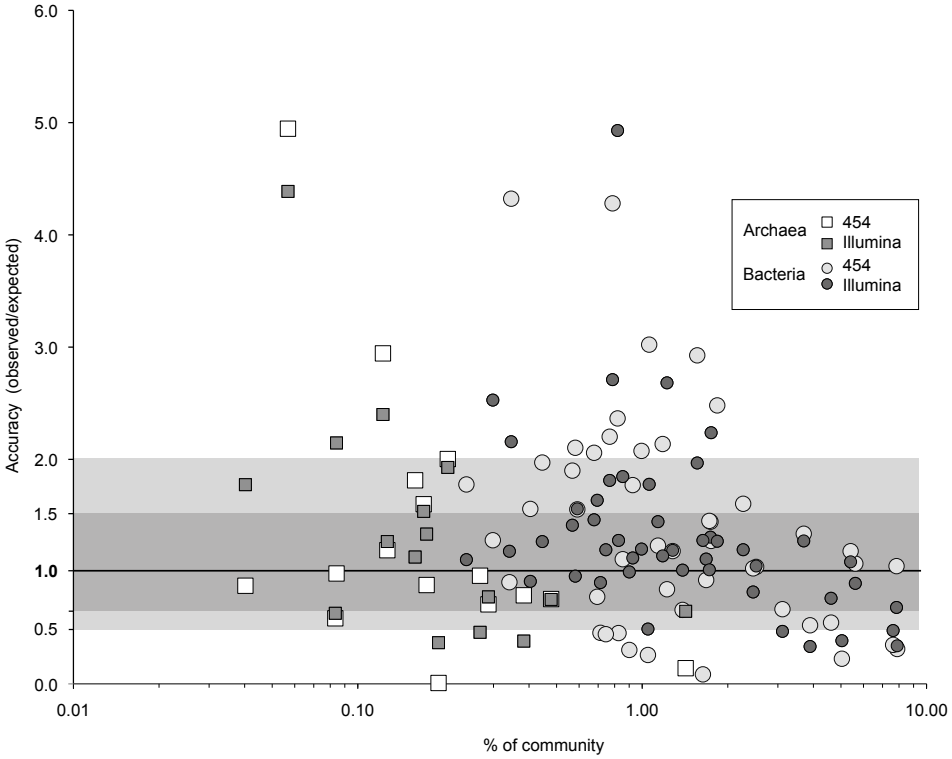
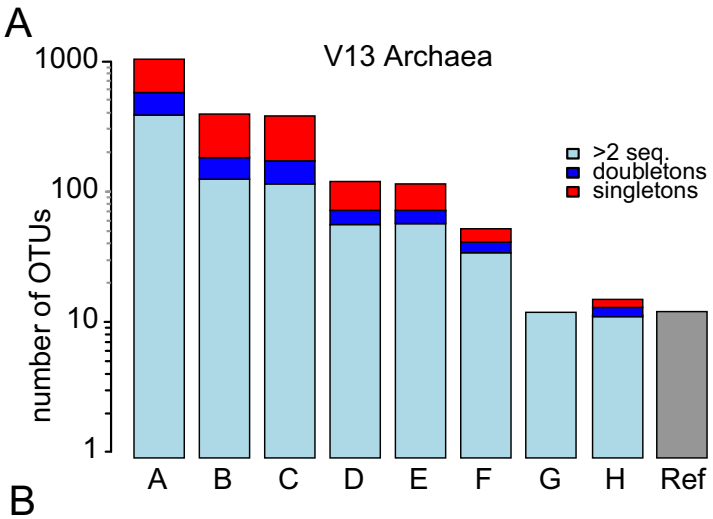


Figure 2.1: Characterization of the Archaea-Bacteria community by 454-FLX-T (A) and Illumina-HighSeq (B) metagenomic sequencing. The accuracy of retrieving the known composition of the metagenome is indicated for each organism as a ratio of the observed genomic coverage to the known genome abundance in the community and is plotted against its known abundance in the community. Shading zones indicate a low level of bias (dark: <1.5 fold; light: 1.5-2 fold) from the 12 perfect value of 1

The sequences were uploaded into two widely used metagenomic analysis systems, IMG/M (Markowitz et al., 2012) (454 data) and MG-RAST (Glass et al., 2010) (454 and Illumina data). Because all individual genomes of the synthetic community are integrated in these systems, we evaluated IMG/M and MG-RAST for accuracy in predicting and quantifying the genomic diversity of the synthetic metagenome. Although corrections for individual genome size and coverage are difficult to apply on metagenomic datasets, both systems recovered the bacterial phyla representation quite well, with most taxa estimated to within a factor of two of their actual abundance (Fig 2.9). Archaea were less accurately quantified, some being either under (*N. equitans*), or over estimated (*Crenarchaeota* and *Euryarchaeota*). Inferences based on the Illumina data were generally consistent with those based on 454, with some discrepancies potentially due to GC coverage differences between platforms. However, both IMG/M and MG-RAST predicted a higher diversity than actually



	A	B	C	D	E	F	G	H	Reference
PRIMERS	2	2	0	0	0	0	0	0	N/A
Q SCORE	0	0	0	25	25	20	20	N/A	N/A
HOMPOL	N/A	N/A	N/A	6	6	9	9	N/A	N/A
QWINDOW	N/A	N/A	N/A	N/A	N/A	50	50	N/A	N/A
SEQ TRIM	>200	~400	~400	~400	~400	~400	~400	~400	~400
CHIMERA	N/A	N/A	N/A	N/A	CS	CS	CS	Perseus	N/A
SEQS (Bac)	2982	2845	2833	2446	2296	1728	1728	2337	153
SEQS (Arc)	9496	9418	9020	6780	6774	6218	6218	6836	25
CLUSTER	MSA-CL	MSA-CL	MSA-CL	MSA-CL	MSA-CL	MSA-CL	Silva-SLP-AL	AN PW-AL	Silva-SLP-AL

Figure 2.2: Effect of sequence processing parameters on OTUs. Sequenced amplicons from V13 region of SSU rRNA of both Archaea (A) and Bacteria (B) were filtered, trimmed, and clustered using the parameters specified in a table form (A-H). Sequences were trimmed to the same coordinates after alignment against the SILVA database and clustered using either complete linkage clustering (CLC) or average linkage clustering (AL) with distances based on Infernal alignment or SILVA based alignment in mothur, respectively. The numbers of OTUs at 97% sequence similarity (distance 0.03) are shown with distinguished contribution from OTUs consisting of one, two or more sequences.

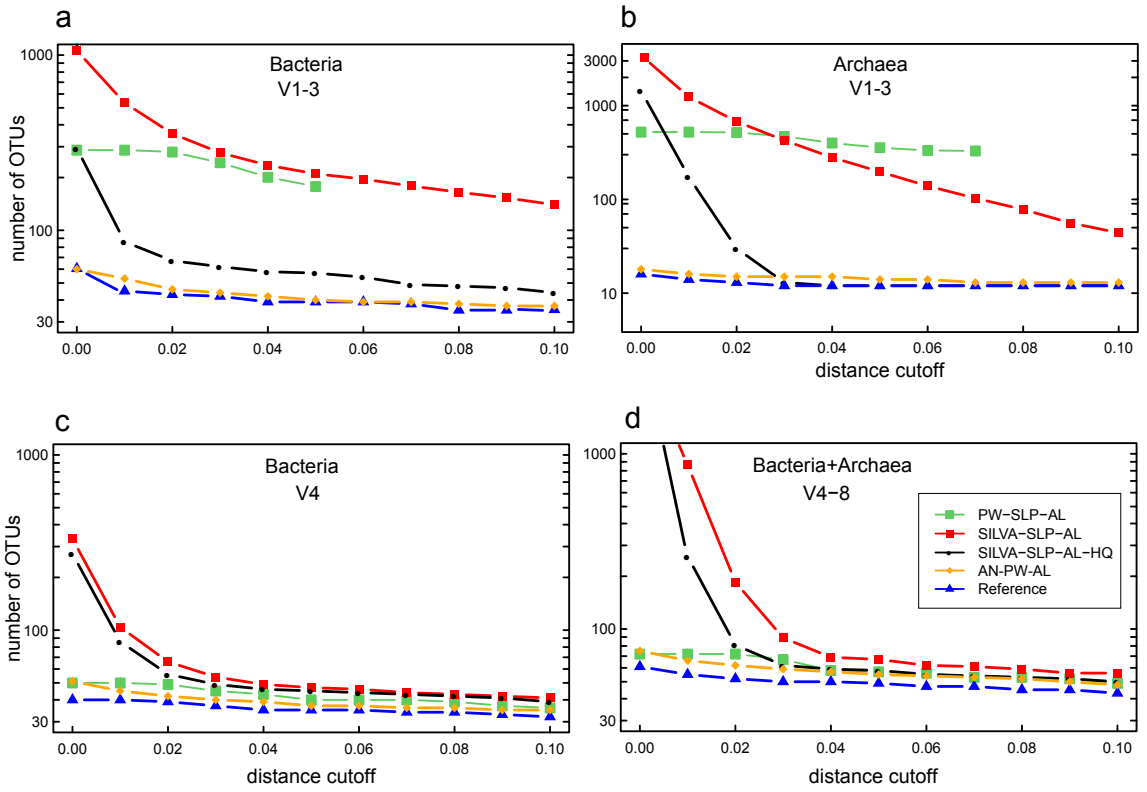


Figure 2.3: OTU-based diversity estimation as a function of genetic distance and analytical approach relative to the reference genomic SSU rRNA sequences. Bacterial V13, V4, archaeal V13 and the combined archaeal-bacterial V48 amplicon datasets are shown. The results for the other amplicons are shown in the Figure 2.14. Silva-SLP-AL (red) and Silva-SLP-AL-HQ (black): single-linkage pre-clustering 2% and average linkage clustering of SILVA alignment of sequences not purged of errors and on sequences with the chimeras removed (parameters B and G in Figure 2.2, respectively). PW-SLP-AL (green): single-linkage pre-clustering 2% and average linkage clustering of Needleman-Wunsch (NW) pairwise alignment of sequences not purged of errors. PW-AN-AL (orange): average linkage clustering of pairwise alignment of sequences after denoising and chimera removal using AmpliconNoise/Perseus. For comparison, OTUs obtained by clustering the reference sequences using Silva-SLP-AL (blue) are shown. Note that the y-axis in (A) is scaled logarithmically.

present, at most taxonomic levels (MG-RAST alpha diversity was overestimated by >6 fold for both datasets)(Table 2.3) . Most of the spurious groups of organisms at high taxonomic levels (some bacterial and archaeal phyla as well as fungal, plant and metazoan lineages) were based however on few sequences and had relatively low confidence values but numerous sequences were also incorrectly assigned to taxa closely related to those present in the community. This appears to be due to a combined effect of the short read data and the variable analysis stringency, limiting phylogenetic resolution and leading to incorrect

assignments between closely related organisms and, for conserved genes, even across high-rank taxa. Such, overestimations are an important factor to consider, especially in studies that aim to identify rare organisms. Default predictions by the current versions of these two widely used analysis platforms reported organisms and taxa (e.g. Eukarya, several bacterial and archaeal phyla) that cannot be linked to the community we used here. Some issues may be addressable by sequence assembly, which should improve gene prediction and functional annotation in addition to taxonomic assignments, especially at high sequence coverage (Pignatelli and Moya, 2011). At present, neither IMG-M nor MG-RAST provides Illumina sequence assembly options although MG-RAST allows upload and assignment of raw sequence reads. It is important to note, however, that the Illumina short reads provided a very good estimates of taxonomic distribution above the species level, with only a 2-3 fold overestimation of the actual number of genera and orders. For the 454 data, however, the use of the default parameters severely overestimated higher level diversity (20 fold for bacterial genera and identified >100 spurious eukaryotes). Increasing the stringency of the analysis produced much more accurate results, inline with the Illumina output (Table 2.3). In analyzing environmental metagenomic datasets selection of the various cutoff parameters is therefore an important consideration and, the results presented here may serve as initial guidance in developing such procedures.

We also evaluated the accuracy of taxonomic assignments under the hypothesis that none of the exact genomes of the community are represented in the database. While in natural communities one often times identifies closely related organisms to those that have a genome sequence determined, a large fraction of the Bacteria and Archaea still have poor genomic coverage, even often at phylum level. Therefore, determining where the metagenomic sequence data maps and how accurate the assignments are to higher taxon level (e.g. family, order, phylum) is of interest to expand results obtained with synthetic communities to natural ones. Because the genomes of all organisms we included in the synthetic community are part of the public databases used by IMG-M and MG-RAST, with no option to be excluded from the analysis, we performed a local metagenomic analysis with MEGAN (Huson et al., 2007). We compared taxonomic assignments to a Bacteria-Archaea database containing all available sequenced genomes, and to a version of that from which we removed the genomes represented in the community. In the first case, the

accuracy was verified to species, genus and family levels. When the reference organism was excluded from the community, the accuracy was analyzed to genus and family levels. When taxa were poorly represented in the genomic database (e.g. the reference genome in the community was the single sequenced representative at genus, order or even phylum level, such as *Ignicoccus, Nanoarchaeum, Gemmatimonas*), eliminating the reference genome from the database affected the assignment of those sequences, most having no match, especially using the stringent megablast algorithm. As a result, the abundance of those taxa were underestimated in the community. The more permissive blastn-based analysis produced a more accurate representation, especially at family level for both 454 and Illumina data. Figure 2.10 summarizes the result of analyses using the two different blast-mapping criteria in comparison to the known taxonomic composition of the community. In addition, we analyzed the use of a single marker gene (SSU rRNA) for explaining the taxonomic and quantitative composition of the community using the 454 and Illumina metagenomes. The reads corresponding to that gene were identified and analyzed using the RDP Classifier (Cole et al., 2009). While many of the taxa were identified to genus level, the quantitative recovery of the relative community composition was very poor, especially with the Illumina data, and there was a severe overestimation of the Archaea (Figure 2.10). We explain this by a combination of low taxonomic resolution of the short reads that randomly cover the rRNA sequence (and therefore carry different phylogenetic signal depending on the degree of variability within the gene) and by the GC bias present in the rRNA operons relative to the average genome content, especially in the many hyperthermophilic archaea present in the community. In addition, because single gene coverage by 454 data in complex metagenome is generally low, taxon identification and quantification is statistically weak. The result of this analysis indicates that a single gene marker such as rRNA is a poor determinant of the community structure in metagenomic sequence data from complex communities, especially when one desires quantitative estimates. Synthetic metagenomes therefore provide important controls in selecting algorithms and parameters used for interpretation of actual environmental data on various platforms and software and should be explored in conjunction with in-situ benchmark studies (Pignatelli and Moya, 2011; Mavromatis et al., 2007).

SSU rRNA gene amplification, pyrosequencing and data processing

For rRNA gene-based taxonomic characterization of the three synthetic communities, multiple, variable-length fragments of the SSU rRNA genes spanning most hypervariable regions were amplified and sequenced using the 454 platform. The selection of primers was based on their use in prior Sanger and 454 sequencing studies and included five pairs for Bacteria, three pairs for Archaea and a pair that we developed to simultaneously capture both domains (Frank et al., 2008; Engelbrektson et al., 2010; Wu et al., 2009; Porat et al., 2010; Bates et al., 2011; Haas et al., 2011; Kan et al., 2011). Because some were limited in taxonomic coverage, we introduced modifications or supplemental variants employed in primer mixtures, to expand their breadth (Figure 2.11 and Figure 2.12 and Table 2.4). While degenerate positions in primers were predicted to enable annealing to almost all taxa included in the simulated communities, mismatches to some of the target sequences existed. Such mismatches allowed us to identify their effects in detecting those taxa, and reflect the important reality that there are no truly universal primer sets. Effects of polymerase fidelity, amplification cycle number and amplicon length on inferred taxonomic diversity as well as the variability between replicate amplifications were also tested. Amplicons were sequenced using either FLX or Titanium chemistry and the resulting data processed for barcode-based de-multiplexing, removal of amplification or sequencing artifacts, and diversity analyses using a combination of software packages (mothur, ChimeraSlayer, AmpliconNoise, RDP, ESPRIT, QIIME). AmpliconNoise analysis involved PyroNoise removal of 454 errors and SeqNoise removal of PCR single base misincorporations (Quince et al., 2011). We estimated the proportion of errors attributable to these two sources by calculating the reduction in error rate after applying each algorithm (Table 2.5). Raw per-base error rates varied from 0.1% to 0.25% for FLX and 0.15% to 0.9% for Titanium, with some differences noted between the various amplicons. The error rate following noise and chimera removal was remarkably low, at less than 0.1% for most regions.

A commonly reported artifact in SSU rRNA amplicon analyses is the formation of chimeras during PCR (Haas et al., 2011; Quince et al., 2011; Suzuki and Giovannoni, 1996), which inflates the inferred richness. Overall, the frequency of chimeras detected by ChimeraSlayer or Perseus in AmpliconNoise was very low (<1% of reads) with the exception

of the bacterial V13 dataset for which it ranged between 7-10%. The rate decreased (2-3 fold) with fewer PCR cycles (<3% at 24 cycles for V13) and also when using highly accurate enzymes with additives for increasing PCR specificity (High-GC mix, Accuprime-Pfx). Whilst the proportion of chimeric reads was generally low, they could form a large proportion of the unique sequences present following noise removal, implying that their contribution to the over estimation of diversity is significant

Community diversity analysis using sequence similarity

Because SSU rRNA gene sequence similarity decreases with increasing phylogenetic distance between organisms, quantifying the differences between individual sequences in microbial community datasets provides a metric of phylogenetic diversity that can be standardized and applied in an ecological and statistical framework. Though approximate and not without pitfalls (Stackebrandt and Ebers, 2006), pairwise similarity values have been adopted in comparing distance-based classifications to phylogenetically defined taxonomic ranks (e.g. 97% similarity corresponding to species level). For the synthetic communities analyzed here, we determined the actual sequence similarity level for each sequenced region of the SSU rRNA gene and each pair of species and strains from the same genus of Archaea and Bacteria (Figure 2.13). These values were used to determine the maximum resolution of the sequence analysis step and, in conjunction with the pairwise distances between all the members of the community, to calculate the actual number of taxonomic units at various levels of sequence similarity. For some genera the 97% value holds relatively well and is uniform between the various regions. For other genera however (e.g. *Thermotoga, Sulfurihydrogenibium, Salinispora*) inter-species similarity values were significantly higher (>99%), which limited the taxonomic resolution and underestimated diversity. However, as OTU similarity cutoffs approach 100%, effective resolution of species and strains in natural communities is confounded by rare sequence errors. Parameters and methods used for sequence processing and clustering into OTUs can additionally impact the inferred diversity OTUs (Huse et al., 2010; Kunin et al., 2009; Schloss et al., 2011; Sun et al., 2012). To exemplify these effects the number of OTUs at 97% similarity (3% distance) is shown in Figure 2.2 for bacterial and archaeal V1-V3 region. The frequency of OTUs with only one

or two sequences is compared with those consisting of multiple sequences as well as with the actual OTUs determined by clustering of reference sequences. Less stringent sequence processing leads to severe diversity overestimation, primarily by singletons. Sequence trimming to common coordinates and quality filtering eliminated a large proportion of the singletons and reduced the number of spurious OTUs. However, even after OTU calculation using the mothur implementation of SLP-AL (Huse et al., 2010; Schloss et al., 2011), some 30% (19-21 out of 61-63) of the bacterial OTUs were still attributable to noise although only one contained more than two sequences. This could be reduced to just 6% (2-3 out of 44-47) if AmpliconNoise is used instead of single linkage pre-clustering for noise removal. This effect is even more dramatic at lower similarity cut-offs. A summary of the number of OTUs at progressive distances for each SSU rRNA gene amplicon is shown in Figures 2.3 and 2.14. However, combined removal of sequencing/PCR noise and chimeras by AmpliconNoise and Perseus followed by pairwise alignments and average linking clustering, eliminated most spurious OTUs at virtually all distance settings and best represented the community diversity.

Community diversity analysis using taxonomic mapping

In addition to diversity estimation using similarity clustering, relating SSU rRNA gene sequences to taxonomically classified organisms provides important information about the composition and, to some extent, potential physiological and ecological characteristics of a community. Because the actual composition of natural communities is not known a priori, the accuracy and resolution of sequence based taxonomic inferences remains undetermined, and most often, is not verified by independent measurements/techniques in ecological studies. Using the synthetic Archaea and Bacteria communities we analyzed how the different SSU rRNA regions reflect the known quantitative taxonomic composition and in comparison to the frequencies obtained by metagenomics. For each sequence dataset and each organism, an accuracy ratio (observed versus predicted sequence frequency) was determined and the average of three replicates is represented as a heat map in Figure 2.4, with a value of one corresponding to perfect agreement. The technical reproducibility between replicates for each primer set ranged from an average of 2.5 fold variation for

the bacterial V4 amplicon to 1.5 fold for V13 amplicon. A higher variation in inferred abundance between the replicates was correlated with decreasing actual organism abundance in the synthetic community, especially at levels below 1%. This closely followed the expected patterns associated with Poisson distribution noise and as such may be mostly be attributable to under-sampling. However, stochastic variation in PCR amplification efficiencies may play a role as well (Figure 2.15). In general, over or underestimating the abundance of the different taxa in a community by up to two fold may be considered a resolution limit of these approaches, however these would likely be greater in the absence of averaging independent sequencing data or pooling of PCR products before sequencing. Such noise can be expected to be even more pronounced in natural communities with higher diversity and many low abundance organisms (Zhou et al., 2011). Our analysis indicates that although for many organisms the inferred abundance is within a factor of two or less from the actual value, no primer set was ideal for quantitatively representing the entire diversity of even our relatively simple community and biases did occur. Some taxon underestimation could be explained by mismatches between primers and the target sequence, as no primers are universal, especially at species level. Primer alignments for all tested organisms and identification of mismatches likely associated with underestimation or lack of detection are shown in Figures 2.11 and 2.12. Surprisingly, some phylum-level detection problems in several amplicon regions could not be directly attributed to primer mismatches. The most apparent discrepancies were underestimation of *Bacteroidetes* and *Actinobacteria* by the V4 amplicons and the lack of detection of most thermophilic *Aquificales* and *Thermotogales* by V69 amplicons. Because these group are important members of specific communities (e.g. mammalian gut, soils, hydrothermal environments), the choice of primers can significantly impact diversity estimation in ways not always predictable by primer sequence analysis (Morales and Holben, 2009). Therefore, caution should be applied when analyzing diversity with novel sets of primers and, if feasible, testing using a synthetic community of gDNA or SSU rRNA plasmid clones from that environment should be considered.

Similarly for the Archaea community, significant differences occurred between primer sets and no combination quantitatively reproduced the composition of the synthetic community. The V13 region amplicon performed best for most of the Euryarchaeota lineages but failed to detect two of the *Pyrobaculum* species that have an intron in that region,

and underestimated other *Crenarchaeota* as well as *N. equitans*. Conversely, combinations of primers that amplify the V4 or V48 region tend to overestimate *Crenarchaeota* and underestimate *Euryarchaeota*, including the methanogens. The explanation for these biases is unclear as, with the exception of *Methanopyrus kandleri*, no clear mismatches occur for any primer combinations with their target species. One potential reason for these fluctuations could be the high GC content of the SSU rRNA sequence of these mostly thermophilic and hyperthermophilic organisms, that in some cases contrasts sharply with that of the overall gDNA. Differences in local melting kinetics in such genomes combined with PCR competition between primers may be one explanation for such bias. Species with extreme genome GC composition (*N. equitans* and *H. volcanii*) were indeed most affected, both in amplicon and metagenomic sequencing. We did not observe any correlation between the degree of bias in either *Bacteria* or *Archaea* communities that can be traced back to the number of rRNA operons in individual genomes. One has to consider nevertheless that accurate quantification of an organisms presence is dependent on rRNA copy number estimation. Because in natural communities the actual number of rRNA associated with each organism is unknown, inferences have to rely on using genome sequence of related species.

To evaluate the reproducibility of replicates and accuracy of community representation between each rDNA amplicon region, and each metagenomic sequencing approach, relative to the expected community structure we calculated Bray-Curtis similarity matrices using the species detection ratios for each dataset. Principal coordinate analysis and hierarchical clustering derived from these matrices indicated that both Illumina and 454 metagenomic data closely recovered the known taxonomic and quantitative composition of both Bacteria and Archaea communities (Figure 2.5). Among the rRNA amplicons, V13 and V35 for Bacteria and V13 and V4a for Archaea best represented the overall composition of the two communities and displayed lowest variability among replicates. The amplicon that captures Bacteria and Archaea (V48) also appears to be a viable option for diversity surveys that target both domains simultaneously.

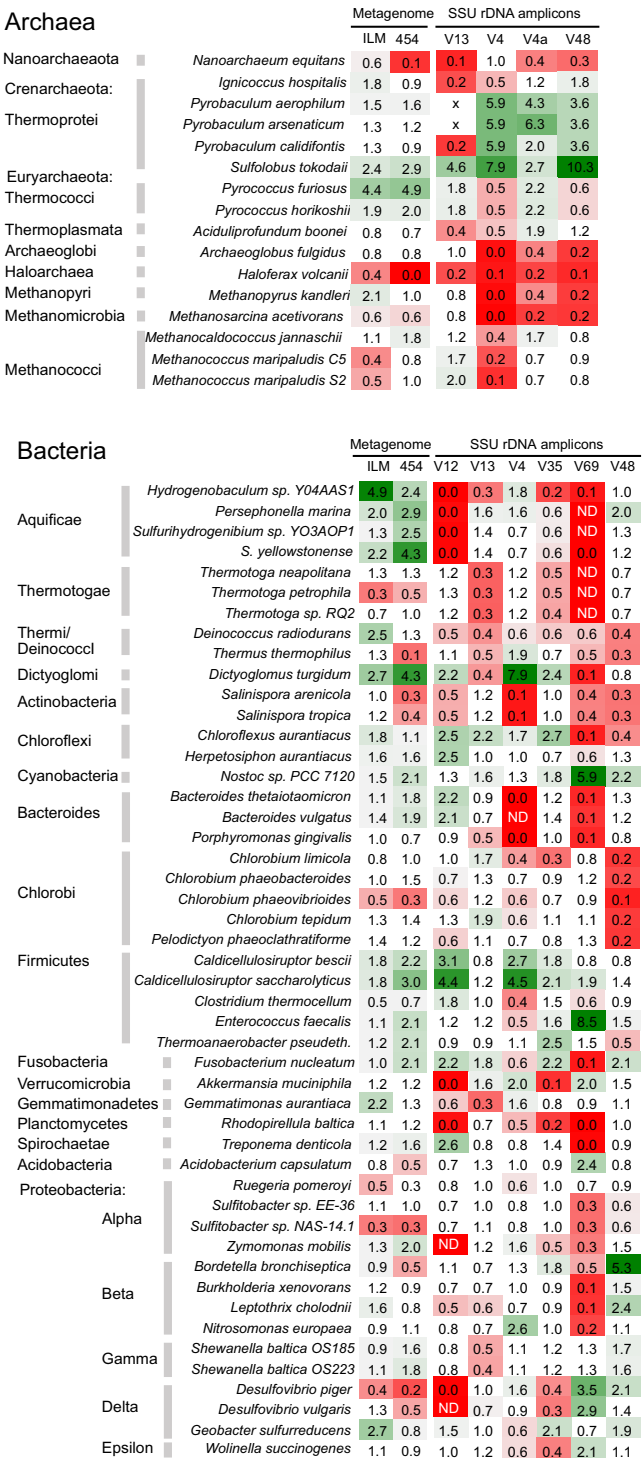


Figure 2.4: Taxonomic diversity and abundance inferences based on shotgun metagenomic and amplicon sequencing. The accuracy ratio (observed abundance/expected abundance) is represented as a heat map diagram with values for each organism and data set. Bias values of >1.5-fold are represented as a heat map of increasing color intensity (red for underestimated and green for overestimated abundance). A value of 0.0 indicates >10 fold underestimated abundance, but detection at low levels. ND indicates that no sequence for that organism was identified in that amplicon dataset. Values are averages of three independent amplification and sequencing replicates.

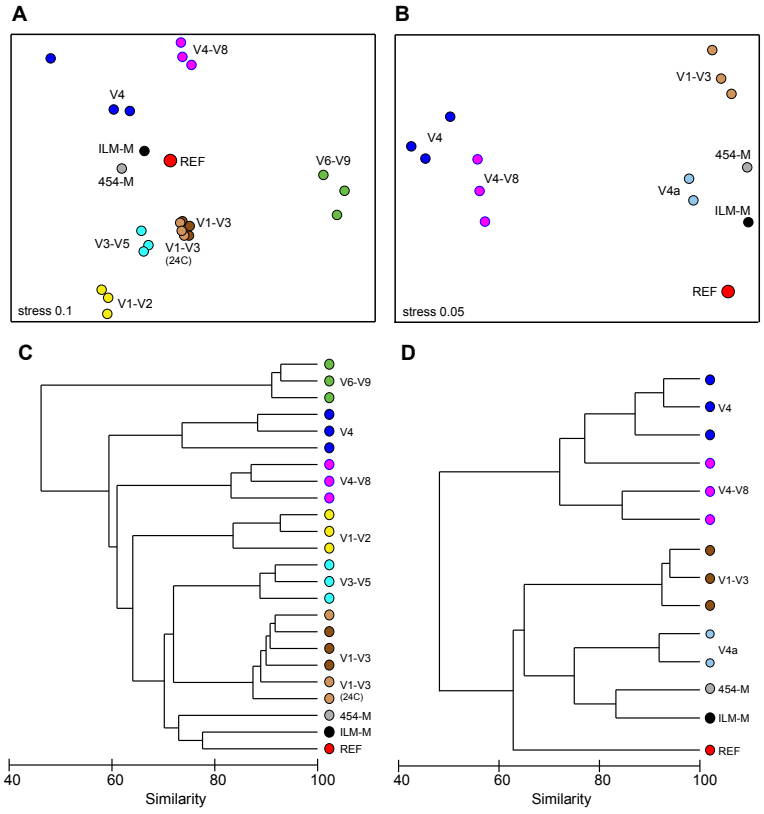


Figure 2.5: (A) Principal Coordinate Analysis (Bray-Curtis similarity) of Bacteria and Archaea community composition inferred using metagenomics (454-M and ILM-M) and SSU rDNA amplicon sequencing relative to the known composition based on community assembly (REF). Replicates for each amplicon are presented, with closer grouping indicating less variability. The V48 data is presented separately for Archaea and Bacteria in those respective panels but was obtained using the combined AB community. (B) Hierarchical clustering (Bray-Curtis similarity) of community composition accuracy indexes for each amplicon region and sequencing strategy.

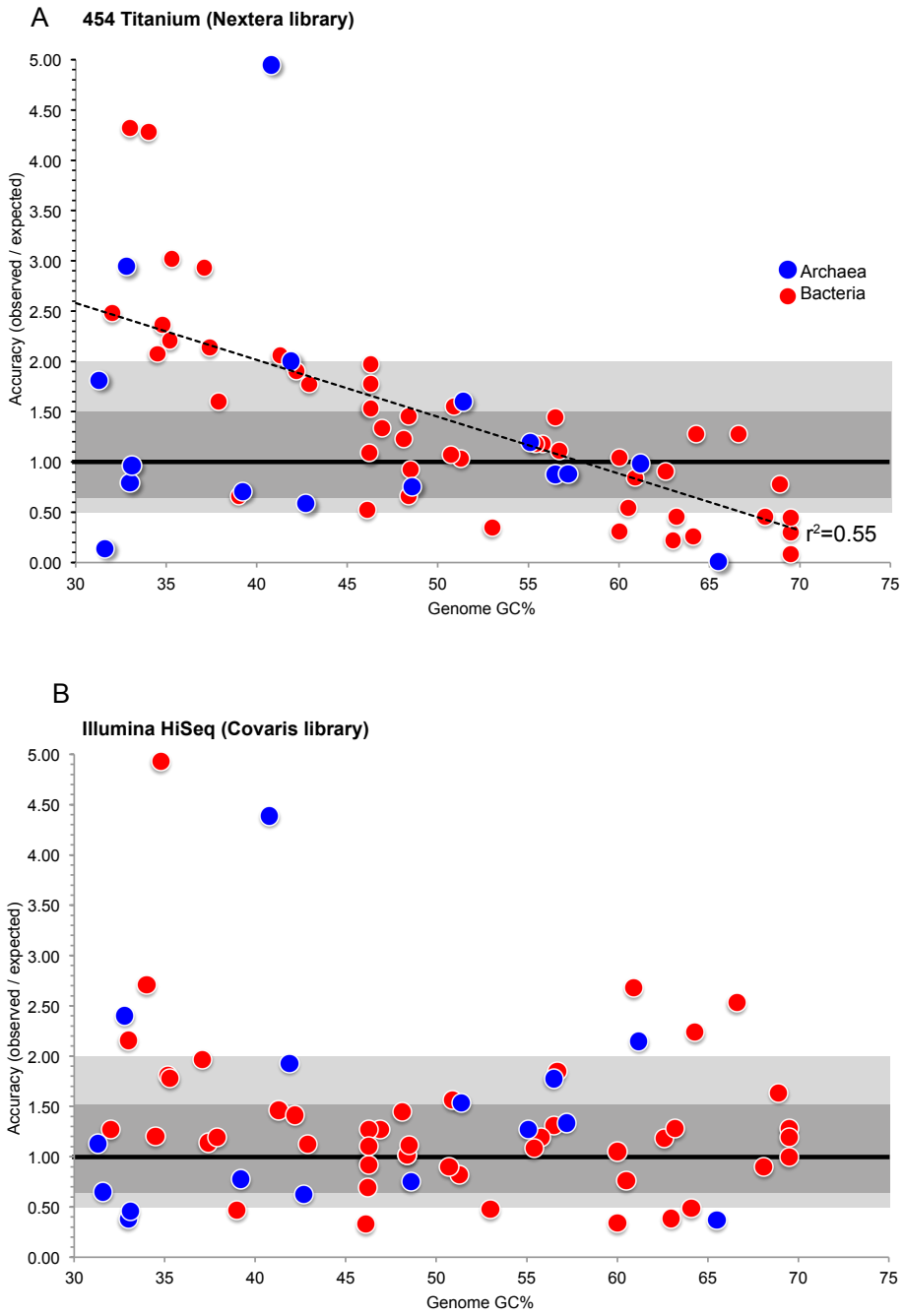


Figure 2.6: The relationship between accuracy of metagenomic abundance estimates and genomic G+C%. The ratio between observed genomic coverage and known genome abundance in the community is plotted relative to the GC contents of that genome for each of the sequencing platforms. Shading zones indicate a low level of bias (dark <1.5 fold, light 1.5-2 fold) from the perfect agreement value of 1. Genomes above or below those zones display an increased bias, correlated with low or high GC content.

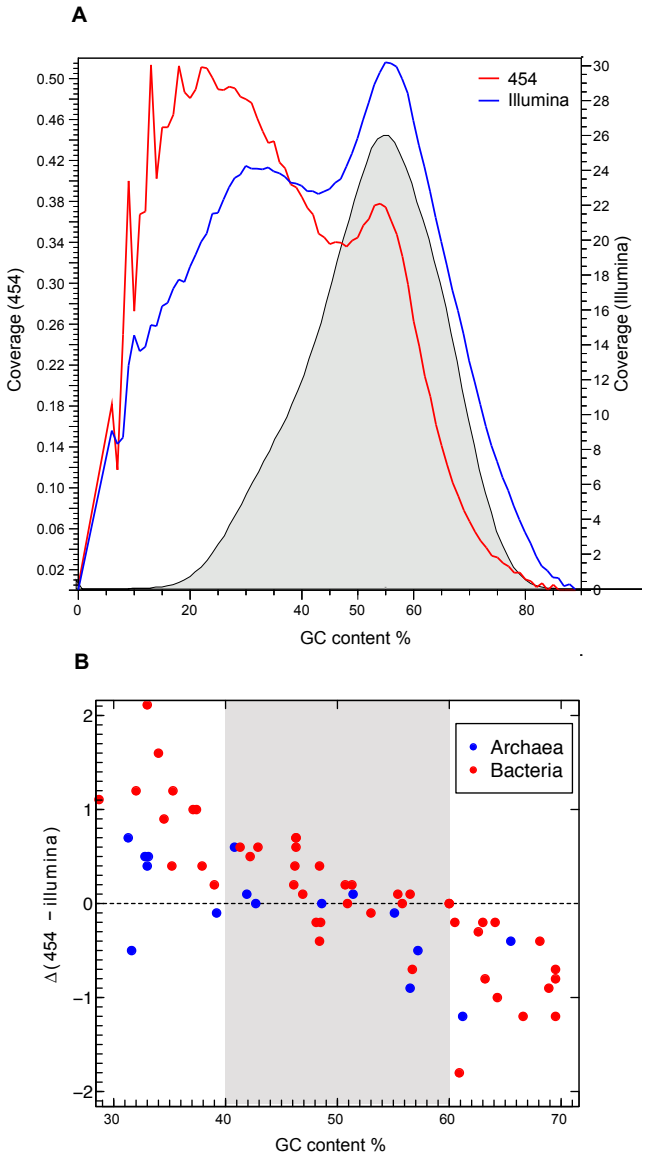


Figure 2.7: (A). Depth of coverage by 454 and Illumina sequence reads on the Archaea-Bacteria metagenome. Genomes of all included organisms served as the reference and the plot displays, for each GC content level, the mean read coverage of 100-bp reference segments with that GC content. The overlapping shaded area represents the quantitative GC content distribution in the reference metagenome (moving 100-bp sequence segments), not scaled to y-axis and included only for distribution shape comparison with the 454 and Illumina data. (B) Differential GC bias in metagenomic quantitative inferences between 454 and Illumina platforms. The three GC window intervals (27-40%,40-60% and 60-70%) were used for pairwise t-test comparisons

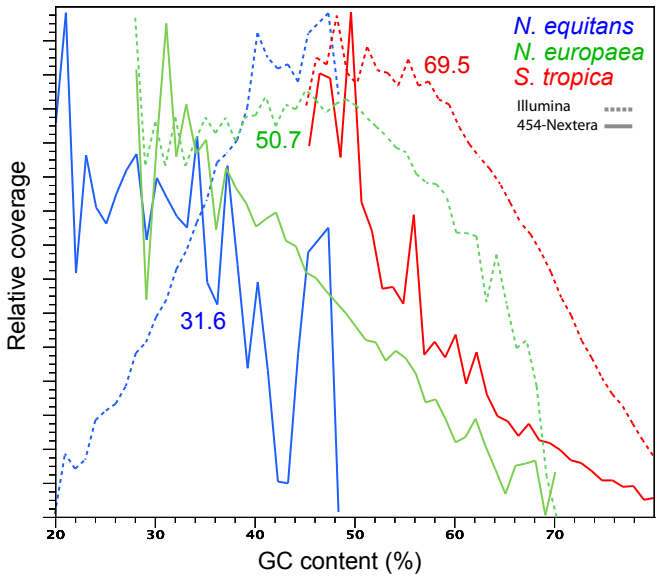


Figure 2.8: Depth of coverage by 454 and Illumina sequence reads of three genomes with low (*Nanoarchaeum equitans*, 31.6%), medium (*Nitrosomonas europaea*, 50.7%) and high (*Salinispora tropica*, 69.5%) average GC content. Each genome contains regions of GC content that depart significantly from the average value (e.g. in ribosomal RNA genes, in non-coding or repetitive regions). To enable overlapped representation of the coverage bias, the Y-axis scale is in relative units, the absolute values being different depending on genomes and sequencing platform.

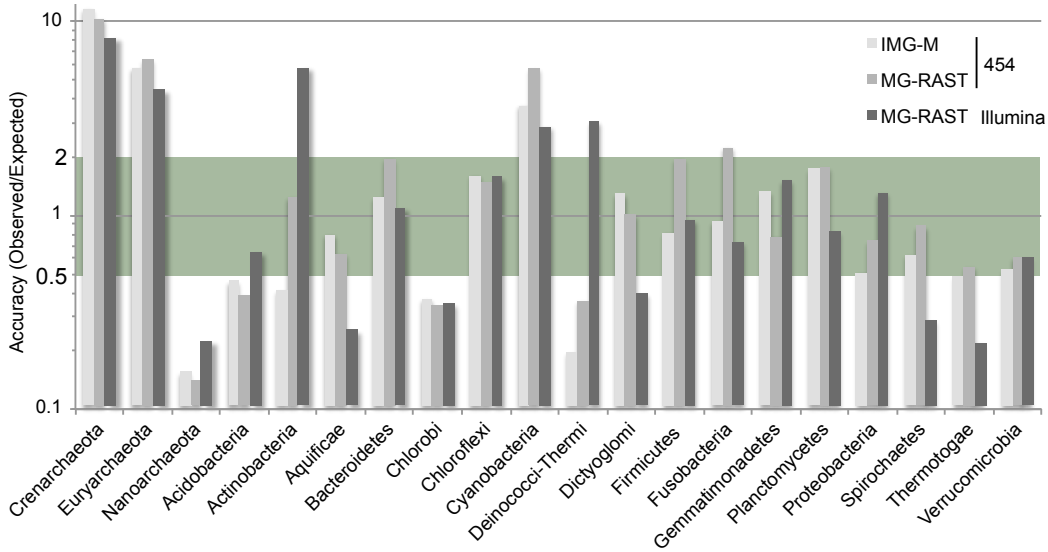


Figure 2.9: Taxonomic diversity composition of the Archaea-Bacteria community inferred by IMG/M and MG-RAST. The accuracy ratio was calculated between the percentage of sequences (454 or Illumina) assigned to individual phyla by the two analysis systems and the known quantitative distribution of those taxa in the community. The shaded region indicates a two-fold accuracy window. Sequences with no assignments or assigned to non-present phyla were not taken into account.

Figure 2.10: MEGAN-based analysis of taxonomic accuracy for 454 and Illumina metagenomes. (A) Heatmap of the accuracy ratio (observed abundance/ expected abundance) at species(s), genus (g), and family (f) level. The heatmap uses the megablast and blastn output against three different databases: (Ref) only genomes of the synthetic community organisms; (All) all microbial genomes; and (XRef) microbial genomes excluding the synthetic community members. Illumina sequence distribution (domain level) for each organism when sequences that mapped to the reference sequences were megablasted or blasted against the XRef database, shown for each organism (B,C) or globally for each blast output (D). No hit indicates the percentage of sequences that were not mapped to any genome. Bacteria and Archaea represent the percentage of sequence that were correctly mapped to corresponding genomes. Other Bacteria and Other Archaea represents the fraction of sequences incorrectly mapped to genomes not present in the synthetic community.

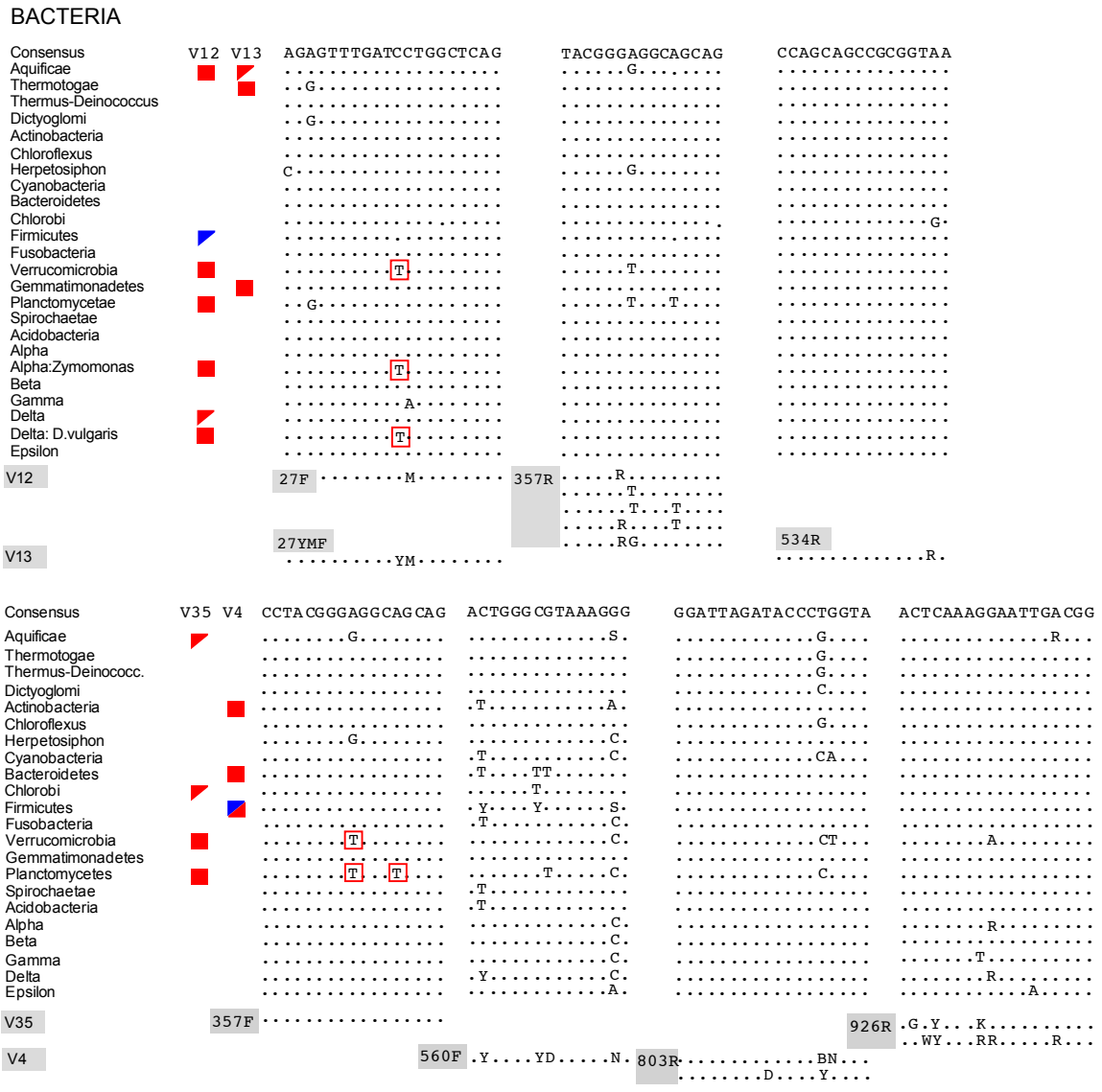


Figure 2.11: SSU rDNA primer pair sequence coverage map. The consensus for all the sequences in the synthetic community and the occasional differences observed for some taxa are illustrated. Nucleotide differences that correlate with observed sequencing bias are in red rectangles. For the various taxa and rRNA amplicons, filled squares indicate a level of >2 fold over (blue) or under estimation (red) for most or all species from that taxon. Isolated cases of bias are indicated by a triangle.

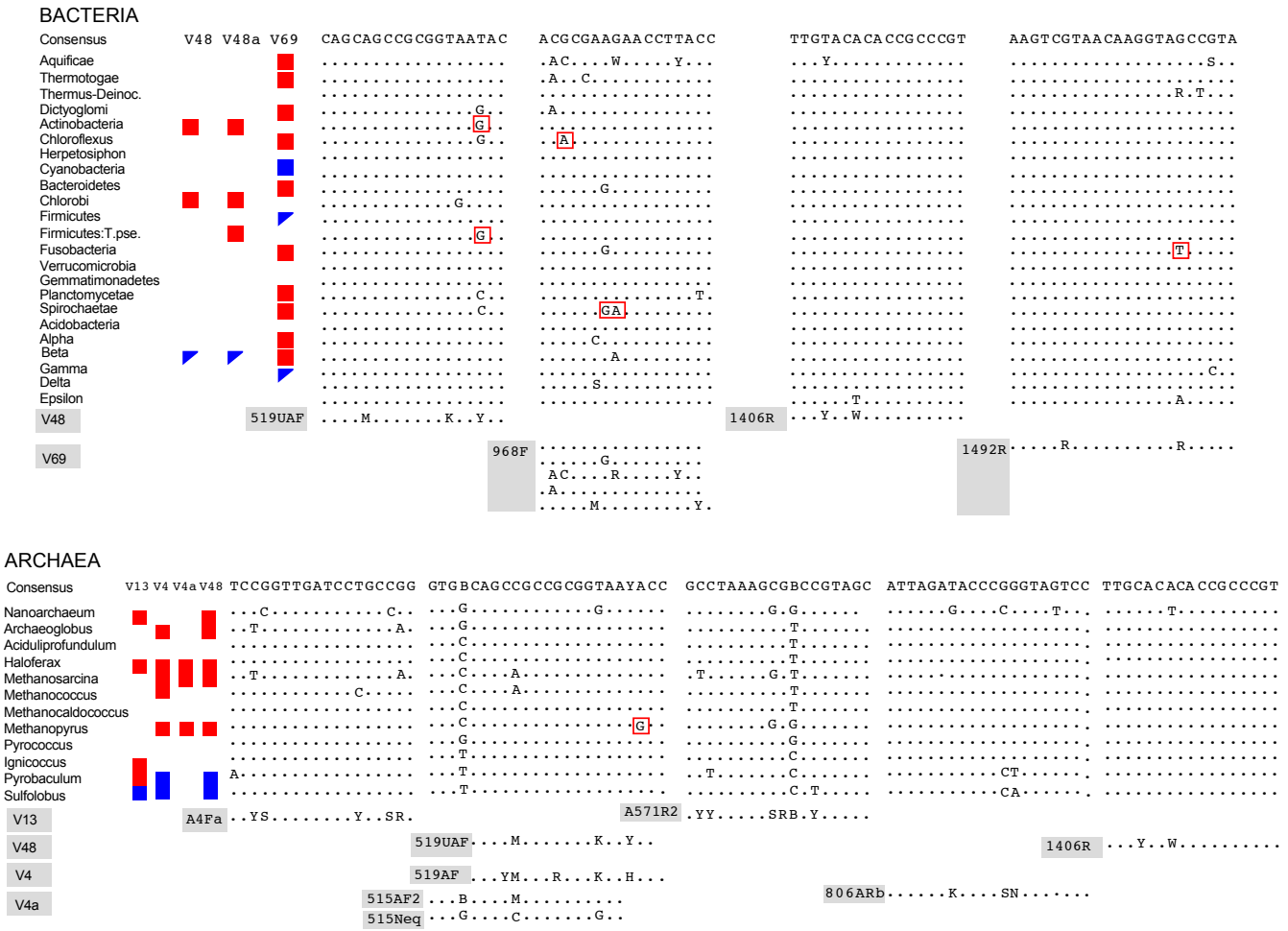


Figure 2.12: SSU rDNA primer pair sequence coverage map. The consensus for all the sequences in the synthetic community and the occasional differences observed for some taxa are illustrated. Nucleotide differences that correlate with observed sequencing bias are in red rectangles. For the various taxa and rRNA amplicons, filled squares indicate a level of >2 fold over (blue) or under estimation (red) for most or all species from that taxon. Isolated cases of bias are indicated by a triangle.

	16S	V12	V13	V35	V4	V69	V48
BACTERIA							
Aquificae							
Sulfurihydrogenibium yellowstonense	**99.4**	99.3	99.3	99.8	100	99.7	99.6
Sulfurihydrogenibium sp. YO3AOP1							
Thermotogae							
Thermotoga naphthophila	**99.5**	99.5	99	99	99	99	99
Thermotoga petrophila	**99.1**	99.4	99	99	99.5	99	99
Thermotoga neapolitana	**99.5**	99	99	99	100	99.8	99.6
Thermotoga sp. RQ2							
Firmicutes							
Caldicellulosiruptor bescii	**96**	97	97	95	89	97	94
Caldicellulosiruptor saccharolyticus							
Actinobacteria							
Salinispora arenicola	**99.5**	98	99	99.8	100	99.8	99.8
Salinispora tropica							
Bacteroidetes							
Bacteroides vulgatus	**91**	86	85	92	91	90.5	93
Bacteroides thetaiotaomicron							
Chlorobi							
Chlorobium phaeovibrioides	**96.8**	95	96	98	97.6	95.5	97.5
Chlorobium limicola	**96.2**	95.5	95	97	98	95.5	97
Chlorobium phaeobacteroides	**94**	92	90	94	96	92.5	96
Chlorobium tepidum							
Proteobacteria							
Sulfitobacter sp. EE-36	**99.9**	100	100	100	100	100	100
Sulfitobacter sp. NAS-14.1							
Shewanella baltica OS185	**>99**	>99	98.7-99.5	99 - 100	100	98 - 100	98-100
Shewanella baltica OS223							
Desulfovibrio piger	**91**	86	82	91	81	91	92
Desulfovibrio vulgaris DP4							
ARCHAEA							
Euryarchaeota							
Methanococcus maripaludis C5	**99.5**	-	99.2	99.3	99.1	-	99.4
Methanococcus maripaludis S2							
Pyrococcus furiosus	**99.1**	-	100	100	100	-	100
Pyrococcus horikoshii							
Crenarchaeota							
Pyrobaculum calidifontis	**98.5**	-	X	99.7	98.8	-	99
Pyrobaculum aerophilum	**98.9**	-		99	99.6	-	99.8
Pyrobaculum arsenaticum							

Figure 2.13: Pairwise sequence identity levels between species/strains in different amplicons

2.5 Conclusions

With the dramatic decline in cost and increase in output, NGS technologies have changed the scale of microbial ecological studies and have made deep metagenomic sequencing much more feasible, affordable and enabled statistically replicated designs that can be quantitatively robust. As 454 and Illumina sequencing probe deeper into the structure of complex communities, determining the real diversity and distinguishing novel or rare organisms from experimental and computational artifacts continues to be a challenge, even though methods and algorithms are continuously improving. Results presented here allow direct and quantitative comparisons within a defined taxonomic space of two complementary and widely used approaches in microbial ecology, shotgun metagenomics and SSU rRNA gene-based diversity characterization. Both metagenomic strategies recovered the quantitative distribution of the various archaeal and bacterial taxa remarkably well even though organisms spanned two orders of magnitude in abundance. A certain degree of bias was observed for genomes with extreme genomic GC content in transposon based library sequencing but because that method enables analysis of samples with reduced biomass, such potential bias may be acceptable and could be accounted for when required by sample/environmental constraints. Additional challenges in analyses of actual environmental metagenomic datasets remain, such as taxonomic assignments for sequences that belong to uncultured taxa, distinguishing closely related organisms, and genome scale assemblies for low abundance species. Advances in taxonomic binning and assembly algorithms (Koren et al., 2011; Liu et al., 2011; Patil et al., 2011), expanding the repertoire of genome sequences to understudied taxa and uncultured single cells (Wu et al., 2009) and very deep sequencing using the Illumina platform (Hess et al., 2011) indicate that even more complex communities are becoming amenable to comprehensive metagenomic characterization. Although metagenomic sequencing outperformed most SSU rRNA gene primer sets used in this study, diversity characterization using this traditional phylogenetic marker is an important approach to compare complex communities in ecological studies where large numbers of samples are required. With the decline in cost and the development of Illumina amplicon sequencing (Caporaso et al., 2011), deeper coverage and more extensive technical replication has become feasible even for highly diverse communities (Prosser, 2010;

Zhou et al., 2011). Among the bacterial primer sets for rRNA gene regions, V13 recovered most accurately the composition of the synthetic community. None of the archaeal sets tested performed comparably to the bacterial ones and presented biases that generally occurred at high taxonomic levels but a modified set of V4 primers (V4a) produced good results. The universal Archaea-Bacteria primer sets (V48) that we tested here, although suboptimal for several taxa, allowed simultaneous comparisons of the representation of the two of the three domains of cellular life in environmental samples. In addition, this was the longest amplicon tested and could provide increased taxonomic resolution with future improvements in read lengths. Each of the primer sets presented phylum-specific biases, not all of which were easily predictable computationally even within the known genomic context of this synthetic community. In particular, the suboptimal detection of *Bacteroidetes* and *Actinobacteria* by the V4 primer set can impact analysis of human microbiota and some soil samples, while V12, V13 and V35 each has difficulties in recovering phyla that are often times highly abundant in some free living communities (e.g. *Aquificae, Thermotogae, Planctomycetes*) or in some human microbiota samples (e.g. *Verrucomicrobia*). Concerted use of two distinct primer pairs for different rDNA regions is therefore important for revealing such biases or even missed detection that may occur for certain taxa (Campbell et al., 2012; Gomez-Alvarez et al., 2009). Since many natural communities contain a much higher taxonomic richness and include uncultured taxa not represented here separate primer sets can provide an independent measure of the accuracy of diversity inferences. An important aspect in microbial ecology studies is richness and evenness estimation and its comparison between communities (alpha and beta diversity). Severe alpha diversity over estimation, especially at low divergence levels (<0.03), can result from sequence errors and from clustering artifacts that are unaccounted for in QA/QC procedures (Huse et al., 2010; Kunin et al., 2009; Quince et al., 2011, 2009; Reeder and Knight, 2010; Schloss et al., 2011). At the same time, the high sequence similarity of SSU rRNA genes between clearly distinct organisms indicates that a component of the diversity may be lost if sequence data is analyzed at distances above the generally applied 0.03 threshold. The results presented here demonstrate that the use of quality-filtered data can nearly eliminate diversity artifacts in SSU rRNA amplicon data. Addressing diversity overestimation in metagenomic analyses is more computationally difficult and may

require simultaneous advances in sequence assembly combined with sequence composition analysis and classification improvements. In environmental datasets, distinguishing rare but real OTUs or metagenomic signatures of uncultured taxa from experimental and computational artifacts remains a challenge. As more and more microbial groups are being sequenced based on pure cultures or single cell genomic DNA, the uncertainty in recognizing and quantifying currently uncultured organisms in metagenomic data is diminishing. In addition, metagenomic sequence binning and assembly is becoming an effective method to identify uncultured taxa and reconstitute their metabolic capabilities (Iverson et al., 2012; Wrighton et al., 2012). Diverse synthetic communities and validation datasets such as the ones presented here enable direct comparison of sequencing, data processing accuracy and effectiveness in sequence binning and assembly for representing the environmental microbial composition and genomic information. Tailored to more closely resemble the expected taxonomic diversity from a specific environment, additional synthetic communities could provide important analytical controls, weather for single gene-type or, increasingly, for metagenomic studies.

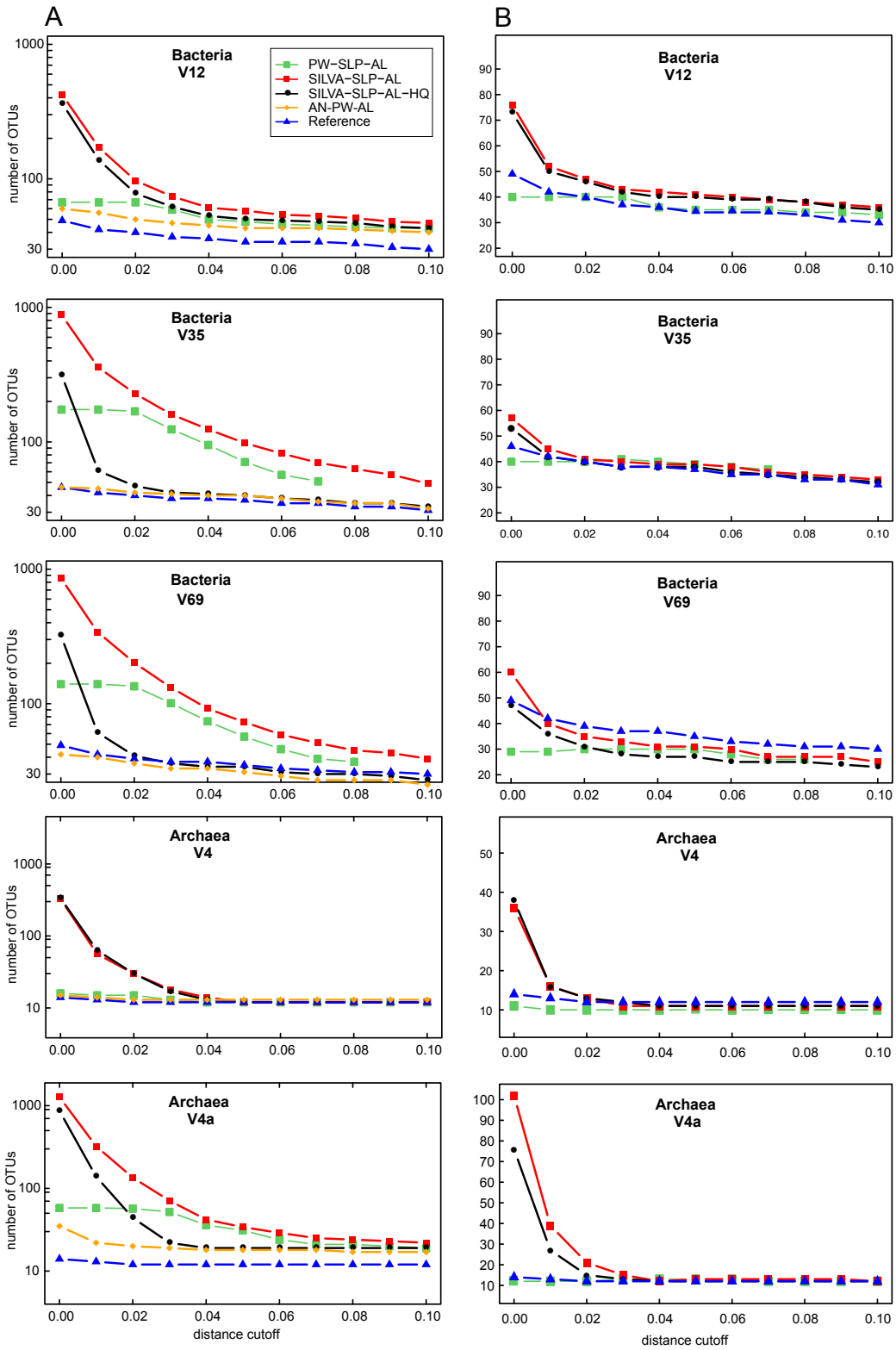


Figure 2.14: OTU-based diversity estimation as a function of genetic distance and analytical approach relative to the reference genomic SSU rRNA sequences

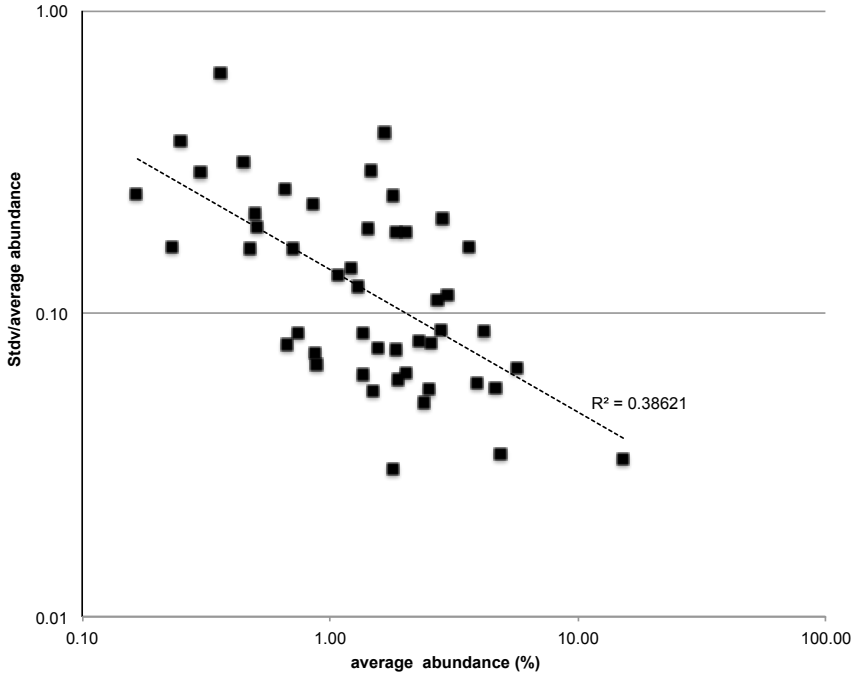


Figure 2.15: Log-linear representation of increased technical replicates variability (standard deviation) as a function of measured organism abundance in the synthetic community, based on bacterial V13 amplicon data.

Table 2.1: List of Archaea used for the synthetic community and their sources. General genomic parameters used to calculate abundance in the communities are presented, as well as the achieved coverage and inferred accuracy values based on metagenomic and amplicon

Ref. Seq	Organism Name	Phylum	Cells/DNA source	Gen.Size (Mbp)	GC %	# 16S	rRNA GC%	Gen. MW
58365	*Ignicoccus hospitalis*	Crenarchaeota	Shakya et al, ORNL	1.30	56.5	1	67.5	8.6E+08
57727	*Pyrobaculum aerophilum* IM2	Crenarchaeota	Lowe T, UCSC	2.22	51.4	1	60.7	1.5E+09
58409	*Pyrobaculum arsenaticum*	Crenarchaeota	Shakya et al, ORNL	2.12	55.1	1	60.7	1.4E+09
58787	*Pyrobaculum calidifontis*	Crenarchaeota	Lowe T, UCSC	2.00	57.2	1	67.8	1.3E+09
57807	*Sulfolobus tokodaii*	Crenarchaeota	Stedman K, PSU	2.70	32.8	1	64	1.8E+09
57873	*Pyrococcus furiosus*	Euryarchaeota	Shakya et al, ORNL	1.90	40.8	1	66.3	1.3E+09
57753	*Pyrococcus horikoshii*	Euryarchaeota	Shakya et al, ORNL	1.74	41.9	1	66.4	1.1E+09
43333	*Aciduliprofundum boonei*	Euryarchaeota	Reysenbach AL, PSU	1.49	39.2	1	61.7	9.8E+08
57717	*Archaeoglobus fulgidus*	Euryarchaeota	Shakya et al, ORNL	2.18	48.6	1	63.9	1.4E+09
46845	*Haloferax volcanii*	Euryarchaeota	Maupin-Furlow J, UF	2.85	65.5	2	56.9	1.9E+09
57883	*Methanopyrus kandleri*	Euryarchaeota	Shakya et al, ORNL	1.69	61.2	1	68.1	1.1E+09
57879	*Methanosarcina acetivorans* C2A	Euryarchaeota	Metcalf B, UIUC	5.75	42.7	3	56.6	3.8E+09
57713	*Methanocaldococcus jannaschii*	Euryarchaeota	Shakya et al, ORNL	1.74	31.3	2	64.3	1.1E+09
58741	*Methanococcus maripaludis* C5	Euryarchaeota	Whitman B, UGA	1.81	33	3	57.6	1.2E+09
58035	*Methanococcus maripaludis* S2	Euryarchaeota	Whitman B, UGA	1.66	33.1	3	57.4	1.1E+09
58009	*Nanoarchaeum equitans*	Nanoarchaeota	Huber H, U. Regensburg	0.49	31.6	1	67.7	3.2E+08

Table 2.2: List of organisms used for the synthetic community and their sources. General genomic parameters used to calculate abundance in the communities are presented, as well as the achieved coverage and inferred accuracy values based on metagenomic and amplicon

Ref. Seq	Organism Name	Phylum	Cells/DNA source	Gen.Size (Mbp)	GC %	# 16S	rRNA GC%	Gen. MW
58857	*Hydrogenobaculum sp. Y04AAS1*	Aquificae	Reysenbach AL, PSU	1.56	34.8	2	54.8	1.0E+09
58119	*Persephonella marina* EX-H1	Aquificae	Reysenbach AL, PSU	1.98	37.1	2	60.8	1.3E+09
58855	*Sulfurihydrogenibium sp. YO3AOP1*	Aquificae	Reysenbach AL, PSU	1.84	32	3	57	1.2E+09
54637	*Sulfurihydrogenibium yellowstonense* SS-5	Aquificae	Reysenbach AL, PSU	1.53	33	4	56.9	1.0E+09
59065	*Thermotoga neapolitana* DSM 4359	Thermotogae	Kelly RM, NCSU	1.88	46.9	1	64	1.2E+09
58655	*Thermotoga petrophila* RKU-1	Thermotogae	Kelly RM, NCSU	1.82	46.1	1	64.1	1.2E+09
58935	*Thermotoga sp. RQ2*	Thermotogae	Kelly RM, NCSU	1.88	46.2	1	64	1.2E+09
57665	*Deinococcus radiodurans* R1	Deinococci	Shakya et al, ORNL	3.28	66.6	3	55.4	2.2E+09
58223	*Thermus thermophilus* HB8	Thermi	Shakya et al, ORNL	1.85	69.5	2	64.2	1.2E+09
59177	*Dictyoglomus turgidum*	Dictyoglomi	Shakya et al, ORNL	1.86	34	2	59.5	1.2E+09
58659	*Salinispora arenicola*	Actinobacteria	Jensen P. UCSD	5.79	69.5	3	60.1	3.8E+09
58565	*Salinispora tropica*	Actinobacteria	Jensen P. UCSD	5.18	69.5	3	60	3.4E+09
57657	*Chloroflexus aurantiacus* J-10-fl	Chloroflexi	Bryant D, PennSU	5.26	56.7	3	63.5	3.5E+09
58599	*Herpetosiphon aurantiacus*	Chloroflexi	Bryant D, PennSU	6.79	50.9	5	58.7	4.5E+09
57803	*Nostoc sp.* PCC 7120	Cyanobacteria	Shakya et al, ORNL	7.20	41.3	4	54	4.8E+09
399	*Bacteroides thetaiotaomicron*	Bacteroidetes	Leys E, OSU	6.29	42.9	5	50.3	4.2E+09
58253	*Bacteroides vulgatus*	Bacteroidetes	Shakya et al, ORNL	5.16	42.2	7	52.4	3.4E+09
58879	*Porphyromonas gingivalis*	Bacteroidetes	Leys E, OSU	2.35	48.4	4	53.3	1.6E+09
58127	*Chlorobium limicola*	Chlorobi	Bryant D, PennSU	2.76	51.3	2	51.8	1.8E+09
58133	*Chlorobium phaeobacteroides*	Chlorobi	Bryant D, PennSU	3.13	48.4	2	51.7	2.1E+09
58129	*Chlorobium phaeovibrioides*	Chlorobi	Bryant D, PennSU	1.97	53	1	52.2	1.3E+09

Table 2.2 continued ...

Ref. Seq	Organism Name	Phylum	Cells/DNA source	Gen.Size (Mbp)	GC %	# 16S	rRNA GC%	Gen. MW
57897	*Chlorobium tepidum*	Chlorobi	Bryant D, PennSU	2.15	56.5	2	52.8	1.4E+09
58173	*Pelodictyon phaeoclathratiforme*	Chlorobi	Bryant D, PennSU	3.02	48.1	3	52.4	2.0E+09
59201	*Caldicellulosiruptor bescii*	Firmicutes	Hamilton-B S, ORNL	2.91	35.2	3	58.8	1.9E+09
58289	*Caldicellulosiruptor saccharolyticus*	Firmicutes	Kelly RM, NCSU	2.97	35.3	3	58.7	2.0E+09
57917	*Clostridium thermocellum*	Firmicutes	Raman B, ORNL	3.84	39	4	55.3	2.5E+09
57669	*Enterococcus faecalis*	Firmicutes	Shakya et al, ORNL	3.34	37.4	4	54	2.2E+09
58339	*Thermoanaerobacter pseudethanolicus*	Firmicutes	Shakya et al, ORNL	2.36	34.5	4	58.5	1.6E+09
57885	*Fusobacterium nucleatum nucleatum*	Fusobacteria	Leys E, OSU	2.17	27.2	5	50.5	1.4E+09
58985	*Akkermansia muciniphila*	Verrucomicrobia	Shakya et al, ORNL	2.66	55.8	3	55.8	1.8E+09
58813	*Gemmatimonas aurantiaca*	Gemmatimonadetes	Shakya et al, ORNL	4.64	64.3	1	60	3.1E+09
61589	*Rhodopirellula baltica*	Planctomycetes	Shakya et al, ORNL	7.15	55.4	1	54.4	4.7E+09
57583	*Treponema denticola*	Spirochaetes	Shakya et al, ORNL	2.84	37.9	2	52.6	1.9E+09
59127	*Acidobacterium capsulatum*	Acidobacteria	Shakya et al, ORNL	4.13	60.5	1	56	2.7E+09
57863	*Ruegeria pomeroyi*	Proteobacteria (a)	Buchan A, UTK	4.59	64.1	3	56.1	3.0E+09
54191	*Sulfitobacter sp. EE-36*	Proteobacteria (a)	Buchan A, UTK	3.60	60	4	54.5	2.4E+09
54259	*Sulfitobacter sp. NAS-14.1*	Proteobacteria (a)	Buchan A, UTK	4.03	60	4	54.4	2.7E+09
58095	*Zymomonas mobilis*	Proteobacteria (a)	Brown S, ORNL	2.06	46.3	3	53.3	1.4E+09
57613	*Bordetella bronchiseptica*	Proteobacteria (b)	Leys E, OSU	5.34	68.1	3	55.8	3.5E+09
57823	*Burkholderia xenovorans* LB400	Proteobacteria (b)	Tiedje J, MSU	9.74	62.6	6	56.2	6.4E+09
58971	*Leptothrix cholodnii*	Proteobacteria (b)	Emerson D, Bigelow Lab	4.91	68.9	2	55.4	3.2E+09
57647	*Nitrosomonas europaea*	Proteobacteria (b)	Shakya et al, ORNL	2.81	50.7	1	53.1	1.9E+09
58743	*Shewanella baltica* OS185	Proteobacteria (g)	Shakya et al, ORNL	5.31	46.3	10	54.8	3.5E+09
58775	*Shewanella baltica* OS223	Proteobacteria (g)	Shakya et al, ORNL	5.36	46.3	10	55	3.5E+09

Table 2.2 continued ...

Ref. Seq	Organism Name	Phylum	Cells/DNA source	Gen.Size (Mbp)	GC %	# 16S	rRNA GC%	Gen. MW
54519	*Desulfovibrio piger*	Proteobacteria (d)	Shakya et al, ORNL	2.90	63	2	55	1.9E+09
58679	*Desulfovibrio vulgaris* DP4	Proteobacteria (d)	Shakya et al, ORNL	3.66	63.2	6	57	2.4E+09
57743	*Geobacter sulfurreducens* PCA	Proteobacteria (d)	Shakya et al, ORNL	3.81	60.9	2	56	2.5E+09
61591	*Wolinella succinogenes*	Proteobacteria (e)	Shakya et al, ORNL	2.11	48.5	3	51.8	1.4E+09

Table 2.3: Taxonomic distribution of Archaea-Bacteria community sequences based on MG-RAST and IMG-M analysis. The AB Ref column indicates the number of actual taxonomic levels (phyla to genus) represented by the organisms included in the synthetic community. The 454 sequence data was analyzed using both MG-RAST (RAST) and IMG-M. For MG-RAST, three cutoff parameters were used: (a, default) max e-value: 1e-5, min% identity: 60%, min alignment length: 15; (b) max e-value: 1e-10, min% identity: 60%, min alignment length: 15; (c) max e-value: 1e-20, min% identity: 80%, min alignment length: 50; For IMG-M analysis there are parameter options and the taxonomic richness was calculated based on the database output (x indicates taxonomic levels not available in the output). The Illumina reads were analyzed using the default parameters in MG-RAST (max e-value: 1e-5, min% identity: 60%, min alignment length: 15;).

	Ref	454				*Illum*
	AB	*a*	*b*	*c*	*M*	*RAST*
Archaea						
phylum	3	5	5	3	3	3
class	9	12	12	9	9	7
order	10	20	20	10	0	8
family	12	30	29	11	18	8
genus	12	72	71	17	0	9
Bacteria						
phylum	17	28	28	20	23	19
class	22	43	43	26	33	26
order	29	91	91	48	0	51
family	34	215	206	65	182	72
genus	35	648	635	115	0	102
Eukarya						
phylum	0	18	13	2	5	1
class	0	52	35	2	0	2
order	0	89	59	2	0	3
genus	0	119	76	2	10	3
family	0	170	103	2	0	3
Viruses						
family	0	5	4	2	3	2

Table 2.4: List of SSU rRNA primers used for amplicon-based diversity characterization of the synthetic communities.

Region	E.coli numbering	Platform	Primer Name	Forward Primer Sequence (5'-3')	Reverse Primer Sequence (5'-3')	Amplicon Length*	Read Length**	Reference
Bacteria								
V1-V2	27-357	FLX	27F-357 (R1-5)	GTTTGATCMTGGCTCAG	CTGCTGCCTYCCGTA (80) CTGCTGCCCYCCGTA (7.5) CTGCTGCCACCCGTA (7.5) CTGCAGCCACCCGTA (2.5) CTGCAGCCTYCCGTA (2.5)	281-400	152-222	(Weisburg et al., 1991)
V1-V3	27-534	Tita.	27YMF-534R	AGAGTTTGATYMTGGCTCAG	TYACCGCGGCTGCTGG	431-550	292-400	(Frank et al., 2008)
V3-V5	357-926	Tita.	357F-926R	CCTACGGGAGGCAGCAG	CCGTCAATTCMTTTRAGT (80)	525-553	385-399	(Muyzer et al., 1996)
V4	560-803R	FLX	560F-803R	AYTGGGYDTAAAGNG	CCGYCAATTYYTTTRWGT (20) TACCRGGGTHTCTAATCC (30) TACCAGAGTATCTAATTC (5)	206-207	200-213	`pyro.cme.msu.edu`
V6-V9	968-1492	Tita.	F96(1-5) 1492R	ACGCGAAGAACCTTAC (70) ACGCGAGGAACCTTAC (10) ACCGAARAACCTYAC(10) AAGCGAAGAACCTTAC(5) ACGCGMAGAACCTTAYC (5)	CTACDSRGGTMTCTAATC (10) TACNVGGGTATCTAATCC (55) TACGGYTACCTTGTTAYGACTT	485-687	347-385	(Weisburg et al., 1991)
Universal								
V4-V8	519-1406	Tita.	519UAF-1406UAR	CAGCMGCCGCGGKAAYAC	ACGGGCGGTGWGTRCAA	850-863(A) 829-972(B)	372-400	(Ovreås et al., 1997)
Archaea								
V4	519-806	FLX	519ArcF-Arc806Rb	CAGYMGCCRCGGKAAHACC	GGACTACNSGGGTMTCTAAT	248-250	210-256	(Suzuki and Giovannoni, 1996)
V4a	515-806	Tita.	515ArcF-Arc806Rb	GTGBCAGCMGCCGCGGTAA	GGACTACNSGGGTMTCTAAT	248-250	241-255	(Bates et al., 2011)
V1-V3	2-571	Tita.	A2FA-571R	GTGGCAGYCGCCRCGGGAA TCYSGTTGATCCYGCSRG	GCTACRGVYSCTTTARRC	479-1221	321-366	(Baker et al., 2003)

Table 2.5: Overview of sequences, percentage per base error rates, and chimeras in pyrosequencing reads before and after QA/QC algorithms.

Region	Synthetic Community	Replicate	Raw reads	Raw error rate	removal of 454 errors	removal of PCR errors	removal of chimeras	454 error rate	PCR error rate	Chimera rate (AN)	# chimeras (CS)	QC pass reads
		1	9197	0.244	0.187	0.147	0.146	0.057	0.041	0.001	1	9114
V1-V2	Bacteria	2	6977	0.271	0.212	0.191	0.185	0.060	0.020	0.006	3	6868
		3	8293	0.236	0.161	0.113	0.111	0.075	0.048	0.002	8	8228
		1	2906	0.861	0.755	0.717	0.093	0.106	0.038	0.624	206	1724
V1-V3	Bacteria	2	2824	0.811	0.716	0.661	0.094	0.095	0.055	0.566	212	1629
		3	3930	0.893	0.781	0.746	0.085	0.112	0.035	0.661	377	2211
		1	2817	0.167	0.120	0.062	0.041	0.047	0.058	0.021	19	1984
V3-V5	Bacteria	2	2001	0.217	0.176	0.125	0.036	0.040	0.051	0.089	22	1451
		3	1649	0.206	0.161	0.113	0.029	0.044	0.049	0.083	10	1195
		1	5209	0.112	0.067	0.019	0.016	0.045	0.049	0.002	1	5178
V4	Bacteria	2	7042	0.125	0.082	0.032	0.032	0.043	0.051	0.000	1	7026
		3	6624	0.115	0.076	0.027	0.027	0.039	0.049	0.000	0	6571
		1	1644	0.287	0.207	0.093	0.072	0.080	0.114	0.021	4	978
V6-V9	Bacteria	2	1944	0.292	0.227	0.084	0.080	0.066	0.142	0.005	6	1103
		3	1724	0.257	0.198	0.086	0.072	0.058	0.113	0.013	2	1016
		1	3608	0.115	0.086	0.055	0.053	0.029	0.031	0.002	1	3442
V4 (A)	Archaea	2	8344	0.088	0.057	0.026	0.026	0.031	0.031	0.000	0	8105
		3	7461	0.107	0.069	0.030	0.030	0.038	0.039	0.000	0	7457
V4a (A)	Archaea	1	11396	0.259	0.237	0.234	0.229	0.022	0.003	0.005	10	11283
		2	12656	0.263	0.244	0.218	0.213	0.019	0.026	0.05	17	12508
		1	8236	0.492	0.307	0.064	0.045	0.185	0.243	0.019	7	6217
V1-V3(A)	Archaea	2	7351	0.506	0.324	0.067	0.064	0.182	0.257	0.003	1	5604
		3	13242	0.489	0.289	0.072	0.069	0.2	0.217	0.003	4	10516

Chapter 3

A multifactor analysis of fungal and bacterial community structure of the root microbiome of mature *Populus deltoides* trees

Disclosure: This chapter has been adapted from text that has been submitted for publication to the journal PLOS One. Migun Shakya was responsible for the majority of data analysis and writing. It will be published as:

Shakya, M., Gottel, N., Castro, H., Yang, Z. K., Gunter L., Labbe J., Muchero W., Bonito G., Vigalys R., Tuskan G., Podar, M. and Schadt, C. W., (2013). A multifactor analysis of fungal and bacterial community structure of the root microbiome of mature *Populus deltoides* trees

3.1 Abstract

Bacterial and fungal communities associated with plant roots are central to the host-health, survival and growth. However, a robust understanding of root-microbiome and the factors that drive host associated microbial community structure have remained elusive, especially in mature perennial plants from natural settings. Here, we investigated relationships of bacterial and fungal communities in the rhizosphere and root endosphere of the riparian tree species *Populus deltoides*, and the influence of soil parameters, environmental properties (host phenotype and aboveground environmental settings), host plant genotype (Simple Sequence Repeat (SSR) markers), season (Spring vs. Fall) and geographic setting (at scales from regional watersheds to local riparian zones) on microbial community structure. Each of the trees sampled displayed unique aspects to its associated community structure with high numbers of Operational Taxonomic Units (OTUs) specific to an individual trees (bacteria >90%, fungi >60%). Over the diverse conditions surveyed only a small number of OTUs were common to all samples within rhizosphere (35 bacterial and 4 fungal) and endosphere (1 bacterial and 1 fungal) microbiomes. As expected, *Proteobacteria* and *Ascomycota* were dominant in root communities (>50%) while other higher-level phylogenetic groups (*Chytridiomycota*, *Acidobacteria*) displayed greatly reduced abundance in endosphere compared to the rhizosphere. Variance partitioning partially explained differences in microbiome composition between all sampled roots on the basis of seasonal and soil properties (4% to 23%). While most variation remains unattributed, we observed significant differences in the microbiota between watersheds (Tennessee vs. North Carolina) and seasons (Spring vs. Fall). SSR markers clearly delineated two host populations associated with the samples taken in TN vs. NC, but overall genotypic distances did not have a significant effect on corresponding communities that could be separated from other measured effects.

3.2 Introduction

Terrestrial plants experience complex interactions with microbes found immediately surrounding the root (rhizosphere) and inside of root tissues (endosphere). This is

particularly true of perennial land plants where inter annual climatic variability and extensive and long-lived root systems, that invade and occupy large volumes of soil, may increase the complexity of rhizospheric interactions. The microbiomes in these root-associated environments are comprised of bacteria, fungi, and to a lesser extent archaea, each with potential beneficial, neutral or detrimental effects on hosts growth and development (van der Lelie et al., 2009; Rodriguez et al., 2009; Berendsen et al., 2012; Danielsen et al., 2012; Mendes et al., 2013; Turner et al., 2013) . A thorough understanding of these complex relationships requires knowledge of resident microbes and factors shaping their abundance and community structure. Few studies have simultaneously examined bacterial and fungal root communities from the same host or genotype over time and even fewer have simultaneously and thoroughly measured the other associated physical, chemical, spatial and temporal factors that may affect these communities. Thus, a deeper analysis of root microbiome as a function of host and environmental factors is pivotal for expanding understanding of the nature and function of these relationships.

Native, woody perennial plant environments, such as those of cottonwood trees (*Populus spp.*), provide an ideal opportunity to understand these associations within relevant environmental settings. The importance of *Populus spp.* in the pulp and paper industry and their potential for future use in production of cellulose-derived biofuels, contributes incentive to increasing our understanding of the effects of microbial relationships on their growth and development. Additionally, *P. trichocarpa* was the first tree species to have a complete genome sequence (Tuskan et al., 2006) and several *Populus* species have become important plant model organisms for understanding the biology and ecology of woody perennials. Moreover, the possibility to study Populus in greenhouses, plantation agroecosytems, as well as in natural ecosystems where they can be dominant keystone species (especially in riparian zones) together make them a powerful and relevant system for providing a better understanding of plant-microbe relationships.

The rhizosphere and endosphere microbiome of *Populus* is important to its overall health and development. Populus associated bacteria are known to promote plant growth and development, increase disease resistance and improve phytoremediation potential (Weston et al., 2012; Doty et al., 2009; Taghavi et al., 2009; Graff and Conrad, 2005). Ectomycorrhizal (ECM) and arbuscular mycorrhizal (AM) relationships also are known

to occur within Populus and influence plant growth and fitness (Lu and Koide, 1994), structure and composition of surrounding plants (Bever, 2003), and overall ecosystem functions (van der Heijden et al., 1998). Thus, characterizing the complex interactions between these trees and their microbiomes are an important step in understanding the overall properties of plants.

Several studies have focused on effects of either bacterial or fungal communities on *Populus* through sequencing clones and cultured representatives of the most abundant organisms (Taghavi et al., 2009; Doty et al., 2009; Graff and Conrad, 2005). We previously used high throughput sequencing to characterize microbes associated with roots of *P. deltoides* and identified a clear distinction between communities in and on the roots (e.g., endosphere vs. rhizosphere) (Gottel et al., 2011). However, that study was limited to only a few individuals within two stands and did not address potential host or environmental factors that may structure microbial communities, or how these communities change over space and time. In other studies, mostly with agriculturally important plants and in greenhouse settings, developmental stage, growing season, genotype/cultivar effects and soil properties have been shown to influence microbial community structure (Lottmann et al., 2010; Aira et al., 2010; Bulgarelli et al., 2012; Lundberg et al., 2012; Hannula et al., 2012; Moore et al., 2006). Deep-sequencing efforts that allow multiplexing of many samples simultaneously, such as the ones used in this study, present an opportunity to scale up these types of analyses and to potentially unravel the links between the *Populus* root microbiome and a wide variety of environmental and host factors that may shape them.

In this study two naturally occurring riparian populations of *Populus deltoides* occurring in Tennessee (TN) and North Carolina (NC) were investigated. We focused on examining the ecological and host factors that could lead to variation in the microbial diversity in and around natural root systems. Specifically, we correlated measures of root microbiome composition and structure with soil physical and chemical factors, host phenotypic factors and genotypic patterns (i.e., SSR-based genetic distances). We sampled roots over two seasons to discern the potential for seasonal variation within these communities. Finally, we described the distribution of OTUs among sampled trees and between rhizosphere and endosphere niches, and identified a core set of both fungal and bacterial OTUs in these two habitats that may play important roles within the plant-microbe-soil interface.

3.3 Methods

Study site and sampling.

We collected native *P. deltoides* samples in two campaigns conducted in spring (May) and fall (September) of 2010. These samples were collected from multiple sites in two watersheds of North Carolina and Tennessee. A total of 24 samples were collected with eleven from North Carolina and thirteen from Tennessee. At each sampling point, we recorded the GPS coordinate and compass intersection of each tree with a handheld GPS. Three soil cores were taken from the adjacent area to each tree in spring sampling campaign only. These soils were refrigerated until soil characterization. Soil characterizations were performed at the Agricultural and Environmental Services Laboratory (AESL) of University of Georgia (`http://aesl.ces.uga.edu/`) on the sieved (4mm) composited samples. The soil characteristics of each tree and surrounding soil are presented in Table 3.3. We collected root samples by carefully excavating and tracing the roots back to *P. deltoides* to ensure identity. The collected root samples were stored in ice and processed next day in lab. Tertiary fine roots were removed, and loosely adhered soils were removed by shaking and then washed with 100ml of 10mM NaCl solution to remove the adhering rhizosphere soil. The resultant wash was collected in 50mL tubes, which made up the rhizosphere samples. For endophyte samples, surface of root samples were sterilized by rinsing root 5 times with sterile distilled water. Then the roots with diameter 2mm or less were transferred to 50ml centrifuge tubes and then washed using 3% of H_2O_2 for 30s, 100% ethanol for 30s, 6.15% of NaOCl with 2 to 3 drops of Tween 20 per 100 ml for 3 min, and again with 3% of H_2O_2 for 30s. These surface sterilized roots were then washed for 3 times with sterilized distilled water. The sterility of the surface was assessed by plating the subsample of surface sterilized root into Luria Broth (LB) plates and incubating the plate overnight at 30°C . These surface sterilized root samples constituent endophyte samples.

Detection of microsatellite polymorphism.

Twenty microsatellites that previously showed clear polymorphisms in *P. trichocarpa* (Tuskan et al., 2004) and tested *P. deltoides* clones, were pre-selected for use in this study from a set of over 200. The PCR and SSR analytic protocols were as follows: reaction

mixtures contained 25 ng of DNA, 50 ng of each SSR primer, 0.2 mM dNTPs, 0.5 U Taq DNA polymerase (Promega Corp., Madison, WI), 10 mM Tris-HCl (pH 8.3), 50 mM KCl, 2.0 mM $MgCl_2$, 0.01% gelatin, and 0.1 mg bovine serum albumin/mL. Amplification conditions on a GeneAMP 9700 thermocycler (Applied Biosystems) included an initial denaturation step at 94°C for 45s followed by 30 cycles of 94°C for 15s, 50–55°C for 15 s, and 72°C for 1 min and concluded with a 5-min extension at 72°C . Reaction products were diluted up to 1:200, denatured in HiDi form amide containing a 400-bp ROX standard (Applied Biosystems), and processed on the ABI Prism 3700 DNA analyzer. GeneScan version 3.5 was used for size calling of raw alleles based on the internal standard and Genotype version 3.5 was used to visualize and assign alleles to categories for scoring purposes (Tuskan et al., 2004).

Microbial DNA extraction and 454 pyrosequencing.

For rhizosphere samples, 2.0 ml of rhizosphere material were pelleted via centrifugation. The resultant pellet was then used for extractions using a PowerSoil DNA extraction kit (MoBio, Carlsbad, CA). For endophyte samples, the surface sterilized roots were chopped into 1 mm sections, divided into 50 mg subsamples, and total DNA was extracted using PowerPlant DNA isolation kit (MoBio, Carlsbad, CA) with the following modifications relative to manufacturers instruction. We added 50 μl of 10% cetyltrimethylammonium bromide to each lysis tube containing the lysis solution and roots to enhance plant cell lysis, followed by three freeze-thaw cycles (80°C /65°C ; 10 min each) and homogenization in a mixer mill for 20 min at 30 Hz (model MM400; Retsch Inc., Newtown, PA). Three subsamples were then concentrated and combined into a single 50 μl extraction. PCR amplification of bacterial 16S rRNA gene from the genomic DNA of 96 (23 trees X 2 seasons X 2 environments) samples was conducted using a pair of primer that targets the V6-V9 region of 16S. The fusion F1070F (5‘-TCAGCTCGTGTYGTGARA-3’) and 1492R primers (5‘- TACCTTGTTACGACTT-3’) were employed with modification for use with the GS FLX Titanium platform (454 Life Sciences, Branford, CT). These primers discriminate against plastid DNA and surrounded an $\sim$ 200bp mitochondrial insert in Populus. Thus we excised and gel purified the bacterial enriched band prior to emulsion PCR. For each

sample, the fusion forward primer was preceded by a unique 8 bp barcode, which was in turn preceded by the 454 A/B primers. For each sample, a 50μl PCR reaction was conducted using the High Fidelity PCR system (Invitrogen, Carlsbad, CA), 0.2 mM of deoxyribonucleotide triphosphates (dNTPs), 2 mM MgSO4, and otherwise carried out as in Gottel et al. (2011). Fungal primers and conditions were identical to those used by Gottel et al. (2011).

Sequence Analysis.

We denoised the pyrosequencing data using AmpliconNoise (Quince et al., 2011), which corrects for both PCR and sequencing error, through QIIME 1.4.0/1.5.0 (Caporaso et al., 2010a). The resultant sequences that were less than 300bp long and didn't align well with the aligned Silva database in mothur (Schloss et al., 2009) were removed. The resulting high quality sequences were then trimmed at around 300bp and binned to respective samples based on unique barcode. For bacterial samples, the sequences were then clustered using uclust (Edgar, 2010) to representative Operational Taxonomic Units (OTUs) at a sequence similarity of 97%. The representative sequences from OTUs were then checked for chimeras using ChimeraSlayer against the gold database provided with the software package. OTUs were assigned a taxonomic unit using RDP classifier 2.2 (Liu et al., 2012) implementation of QIIME 1.4.0, and OTUs that were classified as chloroplast and archaea were removed from further analysis. A phylogeny of the representative sequence was built using the FastTree (Price et al., 2010) algorithm in QIIME v1.4.0/1.5.0 after aligning with Pynast (Caporaso et al., 2010b) algorithm against the GreenGenes (DeSantis et al., 2006) database. Further downstream analyses for Unifrac phylogenetic distance metric; bray Curtis similarity metrics were all conducted in QIIME using a rarefied OTUs table to control for unequal sampling between samples. A principal coordinate analysis (PCoA) ordination based on Unifrac distance matrix and bray Curtis metric was also generated. For fungal sequences, the sequences were checked for chimeras using implementation of UCHIME (Edgar et al., 2011) in mothur without any reference sequences. It detects chimera de novo with an assumption that chimeras are less abundant than there parent sequence. The sequences that were flagged as chimeras were then removed from further analysis. Any sequences that were

less than 200bp were also removed and the resultant sequences were then clustered into OTUs using uclust at sequence similarity of 97%. The representative sequences from OTUs were then assigned to taxonomic unit using RDP classifier 2.4 (Liu et al., 2012). Raw sequence data and analysis files are available from the NCBI-BioProject data archive `http://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA209455` and our PMI project website `http://pmi.ornl.gov` respectively.

Data Analysis

To investigate the relationship between microbial community composition and environmental factors, tree genotype, seasonal variations, and geographic distance, we used *capscale* function of vegan 2.0-5 (Oksanen et al., 2007) for variance partitioning and ranked partial mantel test in the *ecodist* package (Goslee and Urban, 2007) of R statistical software (Team et al., 2008). The species by sample or OTU table that was used in the study was rarified to 1000 sequences per sample for bacteria and 400 for fungi. In order to conduct partial mantel test with soil properties and tree properties listed in Table 3.2 and 3.3, we built a separate composite distance matrix from variables that were selected using the forward and backward selection against the corresponding distance matrix. The distance metric was then generated based on Euclidean distance metrics using dist function in R. For genotype data, we generated a distance metrics of Euclidean based genotype data using a program called GGT 2.0 (van Berloo, 2008). For seasonal variation, which is a categorical data, we generated a distance matrix using the daisy command of R package cluster. We created the geographical distance matrix between each tree using the location and compass direction (degree, minutes, and seconds (DMS)) that were collected using a handheld GPS. The DMS format was converted to decimal degree (DD) using an online tool at `http://transition.fcc.gov/mb/audio/bickel/DDDMMSS-decimal.html`. The DD coordinates were then used to generate a geographic distance matrix by platform free java based software Geographic Distance Matrix Generator. For variance partitioning, distance matrix was converted to principal coordinates using PCNM function, and only the significant coordinates were included in the model. The following normalizing transformations of the variables were done before performing multivariate analyses: Canopy, Basal Areas, River

Distance, P, Ca, Mn, Zn, K, Mg, N, DBH, LBC (ppm CaCO3 / pH) were log10 transformed; percentages of clay, silt, sand, C, OM, basal area dominance of *Populus* spp./hectare were arcsine transformed; count values of # of proximal trees (Prism) and # of proximal *Populus* spp. trees (Prism) were square root transformed, and pH values were left unchanged.

3.4 Results

We sampled roots [ca. 2 mm or less in dia.] from twenty-three *P. deltoides* individuals along watersheds in Yadkin River, North Carolina (NC) and Caney Fork River, Tennessee (TN) over spring and fall seasons (May and September 2010) (Figure 3.1). During the May sampling, we also collected bulk soil from three adjacent locations around each tree to characterize their physical and chemical properties. Geographic coordinates of the sites and the physical and silvicultural properties of host and surrounding environment were also assessed. For this study, host properties are comprised of measurements associated with host phenotype and its surrounding silvicultural setting (including size and distance to nearest neighbor, distance to river, etc.). A comprehensive list of all the host and soil data that was recorded is listed in Tables 3.2 and 3.3.

Soils between the two watersheds differed significantly ($p<0.05$) in numerous properties including Ca^{2+}, $CaCO_3$, K^+, organic matter (OM), phosphate content, and pH. A hierarchical cluster analysis of the measured host and environmental variables revealed high correlation between several of these measured factors. For instance, C and N (spearman $\rho^2 = 0.94$), CaCO3 and OM ($\rho^2 = 0.89$), % sand and % clay ($\rho^2 = 0.75$) and basal area and DBH (Diameter at Breast Height) (ρ^2=0.97) showed high correlation between pairwise combinations (Figures 3.5: (S1) and (S2)). Thus, several of these highly correlated factors (Soil: N, $CaCO_3$, Sand and Host: DBH) were removed from downstream analysis to minimize redundancy within variance partitioning models employed.

Host genotype analyses based on twenty pairs of simple sequence repeat (SSR) primers resulted in two distinct genetic groups, each comprised of individuals originating from either NC or TN with no overlap (Figure 3.1). A total of forty-eight alleles were observed, with an average of 2.4 alleles per primer pair. A phylogenetic tree showing relationships between the individuals is shown in Figure 3.1. The 20 microsatellites uniquely discriminated 11 of

the 23 individuals evaluated. Fourteen individuals that could not be uniquely genotyped fell into three putatively clonal groups, two in TN population and 1 the NC population. Group I of TN was represented by eight individuals, group II consisted of two TN individuals, and group XI was comprised of four individuals from NC. The overall geographic distance between trees from NC and TN significantly correlated with pairwise genetic distance between the trees (Mantel test:ρ=0.726, p=0.0001) (Goslee and Urban, 2007; Team et al., 2008). However, within each local population these associations were much weaker and only the geographic distance between tree locations from the watershed in TN significantly correlated with genetic distance (ρ=0.390, p=0.0214).

Barcoded 454 pyrosequencing of bacterial 16S rRNA and fungal 28S rRNA gene amplicons from 185 rhizosphere and endosphere samples resulted 946,354 high-quality reads after removing sequencing and PCR artifacts using AmpliconNoise (Quince et al., 2011) and ChimeraSlayer (Haas et al., 2011). These sequences grouped into 24,435 bacterial OTUs ($\geq$ 97% similarity) and 2,999 fungal OTUs. Table 3.4 and 3.5 summarize the sequencing reads and OTUs obtained for each rhizosphere and endosphere sample from bacteria and fungi along with number of OTUs. Unlike our previous efforts targeting the bacterial V4 region (Gottel et al., 2011) the V6-V8 primer sets and gel separation procedures employed in this study were able to reduce the amount of host plastid and mitochondrial sequence coincidentally contained in bacterial endosphere samples to an average of $\sim$ 8%, from $\sim$ 85% on average in our previous study.

Taxonomic distribution

Across all samples, we detected a total of forty bacterial phyla from the rhizospheric and endospheric samples, but only nine had an average abundance greater than 1%. The phyla that made up most of*P. deltoides* overall root (rhizosphere and endosphere) microbiome were *Proteobacteria* (56.1%), *Actinobacteria* (17.5%), *Acidobacteria* (10.0%), *Firmicutes* (2.1%), *Planctomycetes* (3.0%), *Verrucomicrobia* (2.8%), *TM7* (1.8%), *Chloroflexi* (1.1%) and *Gemmatimonadetes* (1.0%) (Figure 3.2 (a)). Differential phyla level trends were observed in the rhizosphere and endosphere bacterial communities. In all rhizosphere samples,

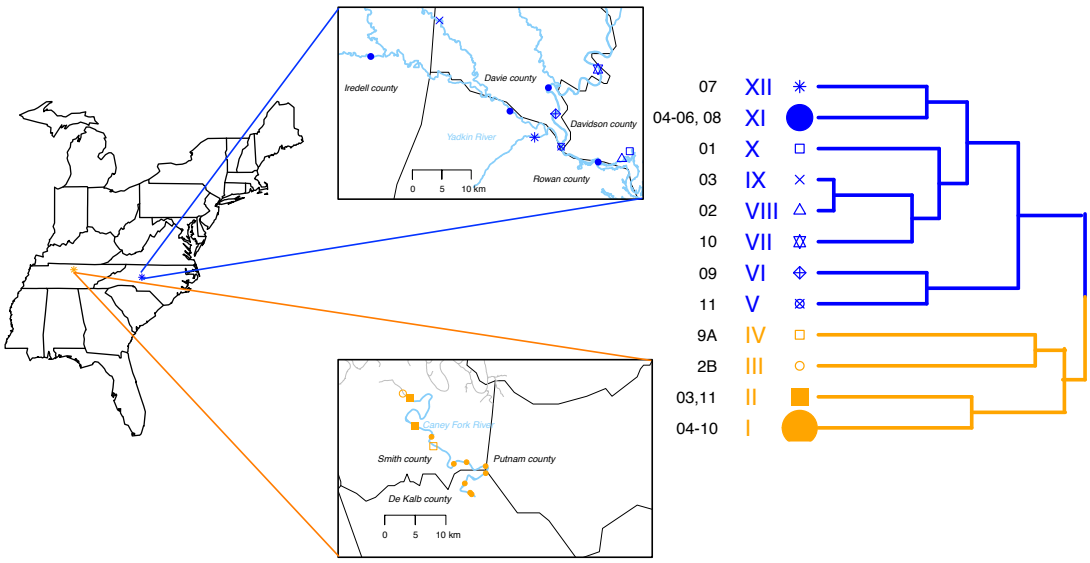


Figure 3.1: Map of sample locations along Caney Fork River in TN and Yadkin River in NC along with phylogenetic tree for twenty-three individuals of *P. deltoides* from twenty simple sequences repeat markers. Each point in the map represents the location of sample along the river and the corresponding point in the phylogenetic tree represents its genotype position compared to each other. The size of the point corresponds to the number of tree in that clonal group.

regardless of watershed or seasonal origin, *Proteobacteria* (51%) was the most abundant phylum followed by either *Actinobacteria* (12.1%) or *Acidobacteria* (14.6%). The remainder of the phyla showed high variability in abundance from sample to sample. For instance, relative abundance of *TM7* in a rhizosphere sample was as high as 19.1%, but its average abundance was only 1.6%. Whereas in the endosphere samples, *Proteobacteria* (62.4%), *Actinobacteria* (23.9%) were enriched, largely at the expense of *Acidobacteria* (4.3%), and members of the *Chloroflexi* (1.0%), *Planctomycetes* (1.1%), *TM7* (2%) and *Verrucomicrobia* (1.3%) were among the less abundant phyla. Endosphere sampled exhibited much greater variability from sample to sample than those from the rhizosphere (Figure 3.9). Also unlike in rhizosphere samples, *Proteobacteria* were not always the most abundant phyla as in the endosphere, as *Actinobacteria* were dominant in $\sim$ 10% of samples.

We detected a total of eight fungal phyla in the rhizospheric and endospheric samples from *P. deltoides*. Across all samples, *Ascomycota* (52%) dominated the overall fungal communities - both rhizosphere and endosphere - followed by *Basidiomycota* (26.9%), *Chytridiomycota* (7.8%) and others of the largely unresolved basal lineages in the former

Zygomycota (now *Mucorales*, *Mortierellales*, etc) that are reported here as *Fungi incertae sedis* (11.4%) (Figure 3.2 (b)). A similar trend was observed in the rhizosphere with *Ascomycota* (50%) as the most dominant phylum, followed by *Basidiomycota*, (20.5%), *Fungi incertae sedis* (11.7%) and *Chytridiomycota* (14.7%) (Figure 3.2 (b)). In contrast, overall endosphere communities consisted primarily of *Ascomycota* (55%), *Basidiomycota* (33%) *Fungi incertae sedis* (11%), while *Chytridiomycota* were largely absent (<1%).

At the higher taxonomic levels we observed more moderate differentiation over geographic space and season compared to differences between the rhizosphere and endosphere. For instance, seven out of the nine major bacterial phyla differed in abundance significantly between rhizosphere and endosphere ($p<0.05$) (Figure 3.6 (a)). Similarly, *Chytridiomycota* were completely absent from 50% of endosphere samples, and across all spatial and temporal samples, only reached a total of 0.7% of endosphere sequences, yet was one of the dominant phyla in the rhizosphere samples. Other fungal phyla including *Basidiomycota*, *Blastocladiomycota*, *Neocallimastigomycota* also differed significantly between rhizosphere and endosphere communities (Figure 3.7), however only few phyla differed in their composition over space (i.e. watershed) and time (i.e. season) (Figure 3.6). Compared to differences between rhizosphere and endosphere, we only observed moderate and often inconsistent differences within these communities, regardless their location or season of sample (Figure 3.9). Over space, *Chloroflexi* and *Ascomycota* from endosphere communities and *Blastocladiomycota*, *Acidobacteria*, and *Chloroflexi* from rhizosphere communities were significantly different between trees from watersheds in NC and TN. Over the two seasons, *Glomeromycota* from endosphere of TN trees was the only fungal phyla that changed significantly from one season to another. In contrast, 10 bacterial phyla showed significant changes between the two seasons, however these seasonal patterns were often inconsistent between watersheds. For example, a significant shift between dominance of *Proteobacteria* (dominant in Spring) and *Actinobacteria* (dominant in fall) in the root endosphere was observed between season in the Tennessee samples but not the North Carolina samples. (Figure 3.6).

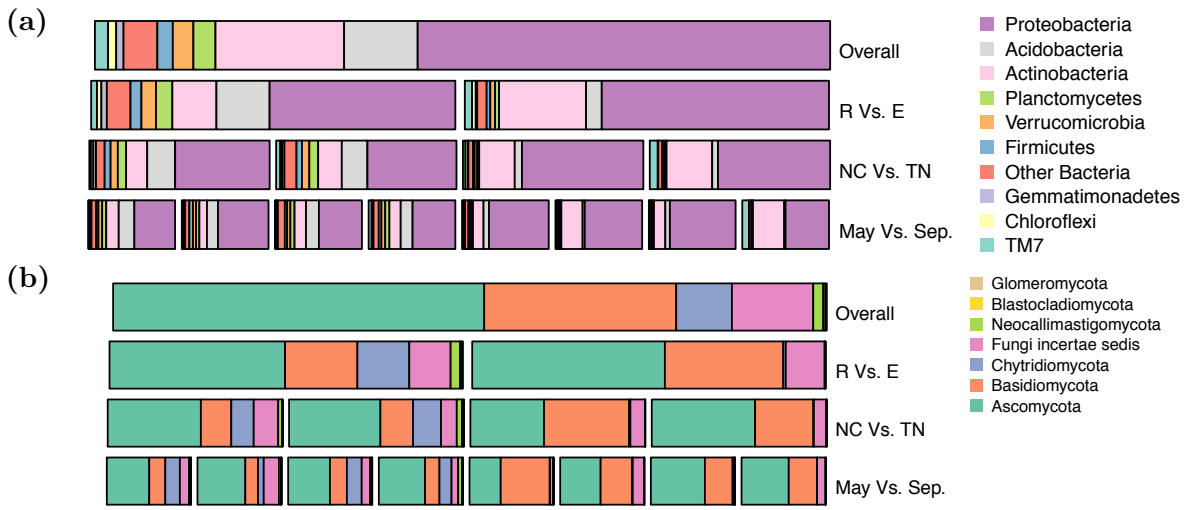


Figure 3.2: : Taxonomic distribution of **(a)** bacterial and **(b)** fungal communities from roots of *P. deltoides*. The first row of stacked bar represents the overall relative abundance across endosphere and rhizosphere; the second row represents endosphere and rhizosphere, third row represents the relative abundance in each watershed, and the fourth row represents the relative abundance in May and September.

Factors related to microbial community patterns of the rhizosphere and endosphere.

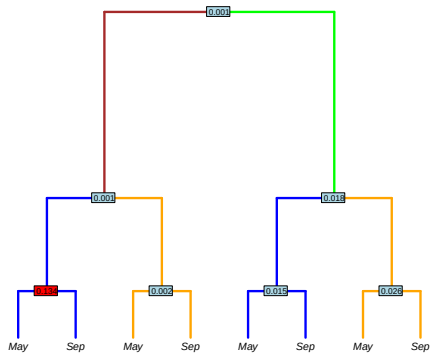
Pairwise UniFrac distances (Lozupone and Knight, 2005) between each sample indicated that bacterial and fungal communities from roots of *P. deltoides* varied significantly ($p<0.05$, Figure 3.8) between rhizosphere and endosphere (Figures 3.3 and 3.9). Though the rhizosphere and endosphere from a common root sample were only millimeters apart, they displayed significant differences in major phyla (Figure 3.6), number of OTUs (Table 3.4 and 3.5) and UniFrac distance (Figure 3.8). To further characterize these communities we separated rhizosphere and endosphere data in recognition that that these likely represent separate habitats or niches that may have differing drivers of their community structure. To identify these drivers of microbial community structure we tested the relationships between community structure and various measurements that included host genotypes and phenotype, soil physical and chemical parameters, geographic distance between samples, season, the characteristics of the reciprocal community associates (bacterial vs. fungal) and the interactions of these variables. Using variance partitioning with distance-based redundancy analysis (db-RDA) of UniFrac inter-sample distances, we determined which host and environmental factors best explained the community structure (Borcard and Legendre, 2002; Legendre and Anderson, 1999). In our results, most of the variation in the community

structure for both communities in rhizosphere and endosphere is statistically unexplained (>40%) with only few of the factors contributing significantly to the variance (∼20%, $p<0.05$). Figure 3.4 represents the proportion of community variance explained by variation of individual factors (effects of all others are neutralized), interaction among factors, and the unexplained variance for both bacterial and fungal communities in rhizosphere and endosphere.

The most important factors that are directly or indirectly affecting the bacterial community in rhizosphere are soil and season. Based on variance partitioning, seasonal change ($p<0.05$; $\sim 4\%$) and soil properties (9.1%, $p<0.05$) explained significant proportions of variance in pairwise UniFrac distances between samples. To elaborate on the component of soil properties, we plotted Canonical Correspondence Analysis (CCA) with a subset of best factors that makes up the composite soil variable. The bacterial rhizosphere communities are influenced by pH as per the heavily weighted arrow and its high correlation with first axis (CCA1: the main explainable variation in the relative abundance of OTUs) (Figure 3.10). Interestingly, the fact that the differences in the rhizosphere bacterial communities are best explained by the variance in local soil properties like pH despite significant differences between communities from two populations (Figure 3.3) suggest that the bacterial rhizosphere communities in *P. deltoides* are structured by changes in the local environment and not by the geographical settings and differences in genotype. The difference observed between two populations is likely due to difference in the local soil properties. In endosphere bacterial communities, only seasonal change significantly explained the variance ($p<0.05$; $\sim 4\%$), suggesting that both local environment and host properties have lesser influence on endosphere bacterial communities compared to changes due to season.

Unlike bacterial communities, significant proportions of variance in (14.0%, $p<0.005$) UniFrac distance between fungal rhizosphere communities was explained by inter-tree distances. Furthermore, soil properties explained 9.83% ($p<0.05$) of variance in UniFrac distance between communities (Figure 3.4 (b)). A CCA plot of fungal communities from rhizosphere consisting of relatively heavily weighted arrows indicate a relationship between

(a)



(b)

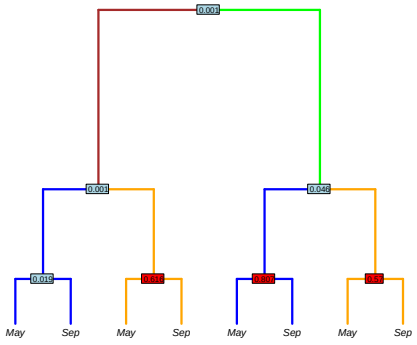


Figure 3.3: A phylogram based illustration of the experimental design and difference in phylogenetic based community structure between two rhizosphere, endosphere, watersheds and seasons. Rhizosphere is represented by brown edges, and endosphere is represented by green edges. Similarly, two watersheds are represented by orange and blue edges for Tennessee and North Carolina respectively. The end node represents two seasons of sample collection. The number at the node represents the p-value (red for insignificant >0.05) generated by comparing the unweighted UniFrac distance metrics between two conditions (left and right nodes) using adonis function of vegan package in R. Phylograms representing **(a)** bacterial and **(b)** fungal communities.

soil properties like Ca, Mn, and moisture content on these communities (Figure 3.11). In contrast to bacterial rhizosphere communities, both the local environment and geographical settings, but not the host genotype, influenced fungal rhizosphere communities. The importance of geographical setting in structuring fungal communities can be attributed to dispersal limitation of fungi, which is less likely to be dispersed between two isolated

locations compared to smaller bacteria (Peay et al., 2007). For endosphere fungal communities, none of the factors that we measured explained significant proportion of variance.

Correlation between fungal and bacterial communities.

Over regional scale (both NC and TN watershed in unison), fungal communities in rhizosphere appear to influence corresponding bacterial communities. Here, partial Mantel tests (Goslee and Urban, 2007) revealed that the Unifrac distances between fungal communities from rhizosphere are significantly correlated with Unifrac distances between bacterial communities (ρ=0.24, p=0.004). Similar to variance partitioning, the test accounted for all other measured variables. Furthermore, to negate the effect of separate watershed location, we conducted the test separately within local NC and TN population; the significant correlation was only maintained in TN population (ρ=0.28, p=0.03). The endosphere bacterial and fungal communities did correlate with each other at regional scale, but the correlation was observed in the endosphere of trees from TN population (ρ=0.26, p=0.03).

OTUs distributions and the core microbiome.

OTUs from roots can be divided into three categories based on their distribution: rhizosphere-specific: 1) OTUs that are only found in rhizosphere, 2) shared OTUs that are found in rhizosphere and endosphere and 3) endosphere specific OTUs. Most of the OTUs (bacterial and fungal) in the roots were rhizosphere specific, with a few shared between the two habitats, and even fewer being endosphere specific (Table 3.1). However, while the number of unique rhizosphere OTUs is high, shared OTUs comprised of most of the sequences (82%), indicating greater dominance often enrichment in the endosphere compartment. A deeper analysis into the distribution of the OTUs among each category showed that 77.8% of rhizosphere-specific OTUs and 90% of endosphere-specific OTUs were unique to one host sample and only few OTUs were present in all sampled trees. Similarly, there were shared OTUs that were only found in one host, but most shared

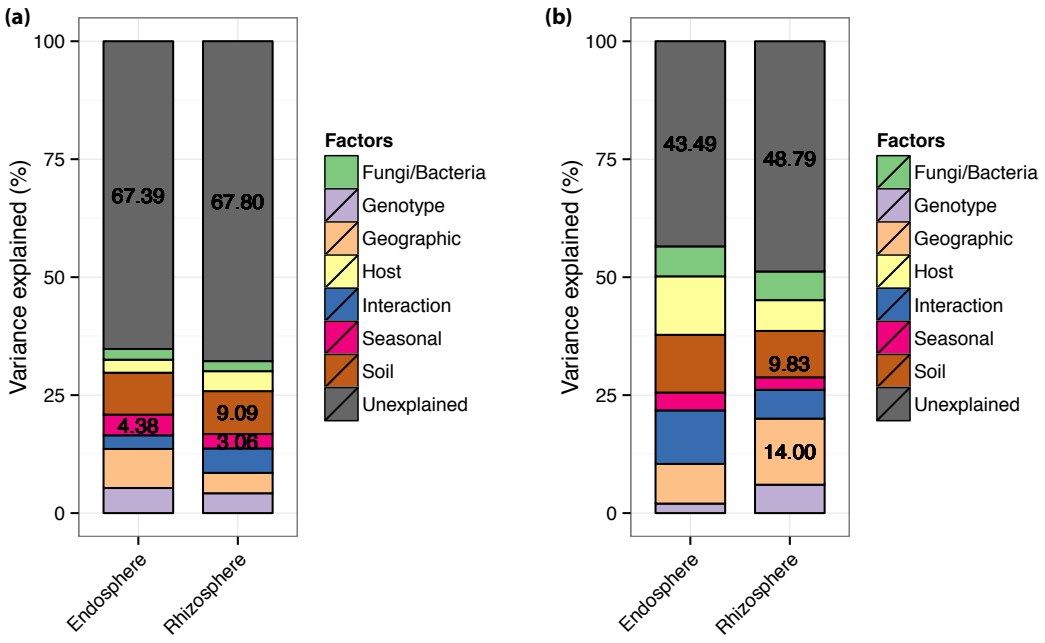


Figure 3.4: Variance partitioning of Bacterial and Fungal communities from the roots of *P. deltoides* into soil properties, host properties, spatial, host genotype, seasonal, and beta diversity of corresponding bacterial or fungal community. Each bar represents total variance, partitioned into pure effect or interaction of two or all factors. Only Variance proportions that were statistically significant are listed in the chart ($p<0.05$). Variables for host and soil properties were selected based on stepwise selection (forward and backward) to remove non-significant terms from the model. **(a)** Variance partitioning of bacterial community from rhizosphere and endosphere. **(b)** Variance partitioning of fungal community from rhizosphere and endosphere.

OTUs were present in multiple host samples. For instance, among shared OTUs, 85% and 53% were found in multiple tree rhizosphere and endosphere samples, respectively. We also detected shared OTUs that were present in rhizosphere or endosphere samples of all trees. A set of 34 OTUs that were shared and one rhizosphere-specific OTU constituted putative 'core' rhizosphere microbiome of P. deltoides from two populations. One of the core OTUs from the rhizosphere was detected in endosphere of every sampled tree. A table with the list of core OTUs along with their top BLAST hits against reference genomic sequences database in NCBI website is listed in Table 3.6. These core OTUs in rhizosphere mostly consist of *Proteobacteria*, and among others there were *Actinobacteria*, *Acidobacteria*, *Verrucomicrobia*, and *Chloroflexi*.

We observed similar distribution of fungal OTUs between rhizosphere and endosphere and among trees. Here 70% of rhizosphere specific and 81% of endosphere specific OTUs were only detected in one host (Table 3.1). We also found shared fungal OTUs that were specific to single tree, but most were found in multiple trees as 71% of shared OTUs in rhizosphere and 50% of endosphere were common to multiple trees. The shared OTUs also constituted a core set of rhizosphere OTUs that comprised of 4 OTUs. One of the four OTUs was also found in endosphere of all the sampled trees. Three out of four core OTUs from rhizosphere classified as *Ascomycota* and one of them classified as a *Mortierella spp.* (Table 3.7).

3.5 Discussion

Differences in rhizosphere and endosphere communities

We previously conducted a study of two locations near the Caney Fork River in Tennessee, USA that revealed that the rhizosphere and endosphere communities of *P. deltoides* were distinct across both their bacterial and fungal communities (Gottel et al., 2011). With the current study we show this pattern clearly holds true across a much wider range of soil types, seasonal transitions, host characteristics and across two regions in the southeastern USA. Additionally, the present study delineates the phyla that are contributing to these difference between rhizosphere and endosphere communities and recovers a greater range of microbes than was revealed in the previous study. At higher taxonomic levels, we observed *Acidobacteria* and *Chytridiomycota* were both more abundant in rhizosphere compared to the endosphere. This result is consistent with recent results reported for studies of the roots of *Arabidopsis*, which also reported low levels of *Acidobacteria* in the endosphere (Lundberg et al., 2012). These two reports suggest that members within *Chytridiomycota* and *Acidobacteria* phyla may lack properties essential for proliferation within endophytic environments.

Also, in contrast to our previous work, we found *Actinobacteria*, similar to the genus *Streptomyces*, were sometimes as dominant (or more so) within endophytic samples as *Pseudomonas*-like *Gamma-proteobacteria*. Our recent study using 'synthetic community mixtures of known composition have shown that the V4 primer set we used in our previous

study underrepresented *Actinobacteria* in community analyses (Shakya et al., 2013). In the current study we employed new primers targeted at the V6-9 region and additional methods to reduce host plastid and mitochondrial rRNA gene contamination. The primer set was tested against this 'synthetic community, which revealed that V6-9 set was able to better recover the overall bacterial diversity, including *Actinobacteria* (Figure 3.12) that the V4 primers had biased against (Shakya et al., 2013). Beyond reducing plant organelle sequence (averaging $\sim 85\%$ in our endosphere samples in the previous study to $\sim 8\%$ in the present), these methods also appear to have eliminated previous biases against *Actinobacteria*. This prominence of *Actinobacteria* in the endosphere is also consistent with other recent studies of plant root endophytes (Lundberg et al., 2012). Detailed follow-up studies with isolates of these phyla may provide valuable insights into deciphering genotypic and phenotypic properties of hosts and microbes that contribute to the entry, survival, growth and function within host habitats.

Factors governing rhizosphere and endosphere community composition.

Our analyses employed variance-partitioning methods to understand how combinatorial effects of host factors, soil properties; presence of other microbes and seasonal variation effect plant associated microbial communities. However, a quantitative understanding of the relative importance of each of factors remained elusive in our study, especially for endosphere communities, that exhibited low diversity (average OTUs: 154 (Bacteria), 39 (Fungi)), but high variability from sample to sample (ranging from 19-1079: bacterial OTUs and 8-169: fungal OTUs). Given the large amounts of unexplained variance within our study, despite considerable efforts to measure a diverse suite of host and environment associated variables, it is possible that unmeasured and/or stochastic factors may play large roles in formation of endosphere and rhizosphere communities. However, given the diverse nature of these communities compared to the relatively low sample sizes we employed (derived from 23 trees, tracked over two seasons in two watersheds), it is quite likely that more of this variation may be attributable with a more robust sample size. Additionally, our power to observe differences given these sample numbers was also likely limited by the significant amount of co-variation that occurred across the two watersheds/populations we sampled,

as a variety of soil properties and host genotype differed significantly between the TN and NC sample origin.

Despite these limitations several factors were found to have significant effects on the structure of rhizosphere and endosphere communities. Within rhizosphere communities the effects of several soil properties (but especially pH), while not large, were significant across both seasons and regions for both bacterial and fungal communities. Such results have also been observed in previous studies (Marschner et al., 2004; Lauber et al., 2009). Season of sampling also explained significant proportion of variance consistently in bacterial communities of rhizosphere, however was not consistently significant in explaining variation within fungal communities. Additionally, both bacterial and fungal community properties varied strongly within the region in which they were sampled (TN vs. NC). The overall observations agree well with previous studies that have identified soil type and season as important players in shaping the microbial community of plants(Hannula et al., 2012; Lottmann et al., 2010; Smalla et al., 2001). The importance of geographical setting in structuring fungal communities may also be due to greater dispersal limitations of fungi than for bacteria, leading to larger effects due to isolation by distance (Peay et al., 2007).

Both bacterial and fungal community structure within rhizosphere were shown to have influences upon each other in the TN population samples (e.g., bacterial community structure correlated with fungal community structure and vice versa). Such interactions, especially bacterial community structures being dependent on corresponding fungal diversity have been documented in other cases (Roesti et al., 2005; Singh et al., 2008; Vestergård et al., 2008). Bacterial influence on fungi, while well documented within studies conducted on Petri plates, are less well documented in natural systems (Kai et al., 2008). The correlations may be indicative of relationship across these groups through the production of secondary metabolites, anti-microbial compounds and/or physical contact (Bonfante and Anca, 2009). For instance, enzymatic activity of extracellular fungal enzymes in lignocellulose-rich soil environments that results in production of water-soluble sugars and phenolic compounds serve as growth substrates for bacteria (Boer et al., 2006).

Plant genotypic effects on microbial community in and around the roots have been documented in other host species (Aira et al., 2010; Caporaso et al., 2011)). Based on the twenty SSR markers that we employed across both natural populations in our study, these

influences were not significant. However, there was a large degree of covariance in our data sets, such that genetic relatedness measured with the SSR markers, as well as multiple soil properties, tended to co-vary between the two regional sampling areas. So while regional distinctions in both rhizosphere and endosphere microbiomes were clearly evident in our data sets (Figure 3.3(a) and 3.3(b)) the specific influence of host versus environmental drivers on these differences remain mostly unexplained, at least in part likely due to this high degree of covariance between variables across the two regions sampled. However, even within the three putative clonal types (ramet genotypes) identified in our SSR analysis, variation was not significantly different than between genotypes as measured by Unifrac distances (Figure 3.9). Host influences on the microbial assemblages in the rhizosphere are complex, but may occur through factors affecting soil properties such as the release of rhizodeposits and exudates (Buée et al., 2009; Broeckling et al., 2008; Shi et al., 2011), secondary metabolites and other factors that were also beyond the scope of variables tracked in this study.

The 'core' endosphere and rhizosphere microbiome of Populus deltoides.

A core microbiome is defined as members of the community that are found in all of the assemblages associated with a habitat (Shade and Handelsman, 2012; Turnbaugh and Gordon, 2009). Deciphering the core microbiome has been proposed to be fundamental to understand the ecology of a microbial community, as the groups of species that are commonly occurring in all habitats are likely to play important role towards communities function (Shade and Handelsman, 2012). We defined the core endosphere and rhizosphere microbial OTUs associated with all the sampled trees in the study (regardless of season, genotype, regional location, etc.) using rarified data sets that may exclude some common, but low abundance organisms compared to those that use overall (unrarified) distributions. These conservative approaches resulted in a rather narrow core microbiome in each habitat. Our core bacterial microbiome in rhizosphere was comprised of only 35 OTUs, one of which, a *Methylibium*-like OTU within *Burkholderiales*, was also the only member of the core endosphere microbiome (Table 3.6). Most of these core rhizosphere OTUs were within the order of *Burkholderiales* and *Rhizobiales* which are known to be important plant associated

organisms, as well as to contain diverse gene clusters encoding degradation pathways for an array of aromatic compounds including pollutants (Pérez-Pantoja et al., 2012).

The core fungal microbiome constituted four only rhizosphere OTUs and one endosphere OTU. Sequence analysis of fungal core OTUs in rhizosphere and endosphere revealed members likely represented the genera *Exophiala, Metarhizium,, Neonectria, and Mortierella*. Some of these organisms are known to have positive benefits to the plants by increasing plant growth, preventing oxidative damage, mitigating salt stress, transferring nitrogen from insect to plant and acting as entomopathogens (Behie et al., 2012; Khan et al., 2012b,a). *Neonectria* is known as an opportunistic plant pathogen in some environments, however their function within native rhizosphere habitats of *P. deltoides* remains undefined. Further genome sequencing of isolates, controlled inoculations and other experiments to test the molecular basis of these associations with host plants will be required to fully appreciate the roles and functions of these fungi.

3.6 Conclusions

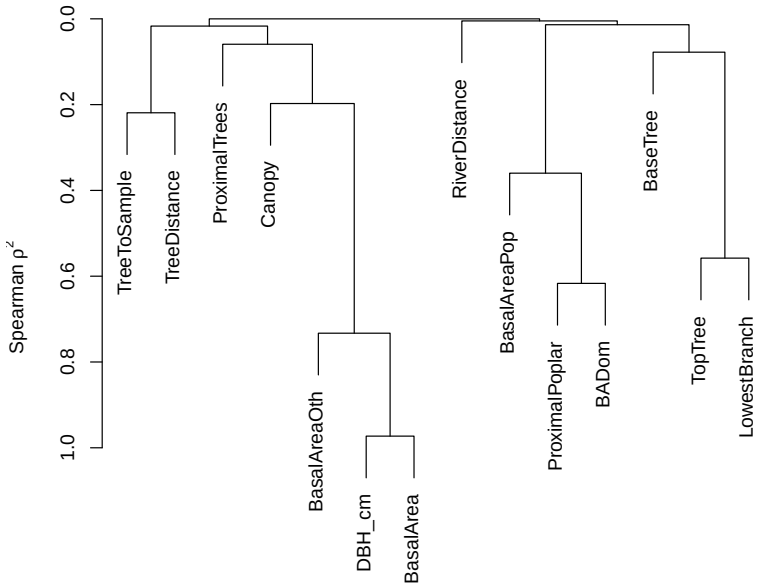
Analysis of rRNA gene amplicons pyrosequencing data from 23 *P. deltoides* host trees across two watersheds and over two seasons for fungal and bacterial community revealed new details about the microbes and microbial community structure in the roots of *P. deltoides*. At higher taxonomic levels (e.g., phyla) rhizosphere and endosphere communities were highly similar between two watersheds differing only in abundance of major phyla. However, at finer levels such as methods using OTUs or UniFrac distances that account for overall phylogenetic variation, clear distinctions were observed for communities from different watersheds suggesting that mature plants of the same species in different locations harbor distinct microbial communities in and on their roots. Also, we observed a seasonally dynamic bacterial community in both the rhizosphere and endosphere of *Populus*. The high degree of covariation within the host and environmental datasets likely limited the power to distinguish between many of the genotypic, geographic and environmental factors that may shape the *Populus* microbiome. Future studies with more extensive sampling and in depth host characterization should further elucidate the factors shaping community structure of both rhizosphere and endosphere communities in *Populus*. Fungal and bacterial

Table 3.1: Distribution of rhizosphere specific, shared, and endosphere specific bacterial and fungal OTUs among all sampled trees. Not all the endosphere samples amplified, so the NA represents the samples that were not sequenced.

	Bacteria				**Fungi**			
	Rhizosphere		**Endosphere**		**Rhizosphere**		**Endosphere**	
# of Trees	Specific	Shared	Specific	Shared	Specific	Shared	Specific	Shared
1	4920	127	328	410	769	72	101	126
2	683	102	24	178	174	30	11	35
3	255	84	9	84	59	21	3	23
4	152	64	3	50	31	17	3	10
5	88	53	2	27	19	21	2	9
6	60	45	0	18	11	10	3	9
7	37	41	0	22	10	14	1	4
8	28	33	1	8	6	6	0	12
9	19	30	1	10	5	11	0	5
10	20	32	0	6	2	8	0	5
11	13	19	0	5	3	5	0	3
12	11	16	0	6	1	3	0	2
13	11	11	0	3	1	5	0	3
14	2	17	0	3	0	3	0	3
15	5	18	0	3	0	3	0	0
16	4	17	0	4	0	5	0	1
17	1	21	0	2	0	3	0	0
18	4	17	0	3	0	4	0	1
19	2	13	0	8	0	3	0	1
20	2	18	0	6	0	2	0	1
21	0	22	0	1	0	3	0	0
22	1	23	NA	NA	0	1	0	0
23	1	34	NA	NA	0	4	0	1

OTU distribution across samples suggested a small set of OTUs that formed the core microbiome and should guide isolate studies that target the detailed mechanisms of host-microbe interactions in *Populus*.

(S1)



(S2)

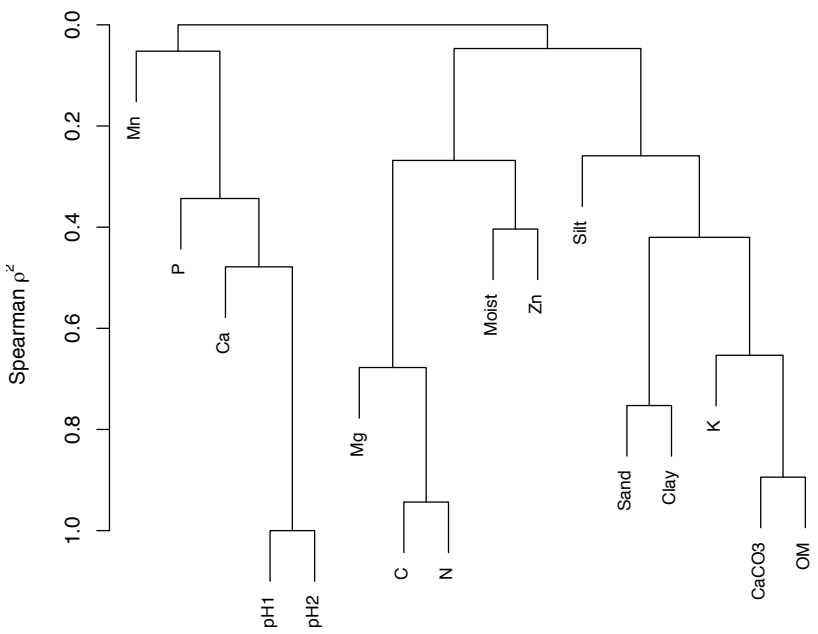


Figure 3.5: Cluster analysis of the measured environmental variables to remove redundant variables from the model. The analysis was done using *varclus* function of Hmisc package in R statistical software. **(S1)**: Tree and stand properties **(S2)**: Soil properties

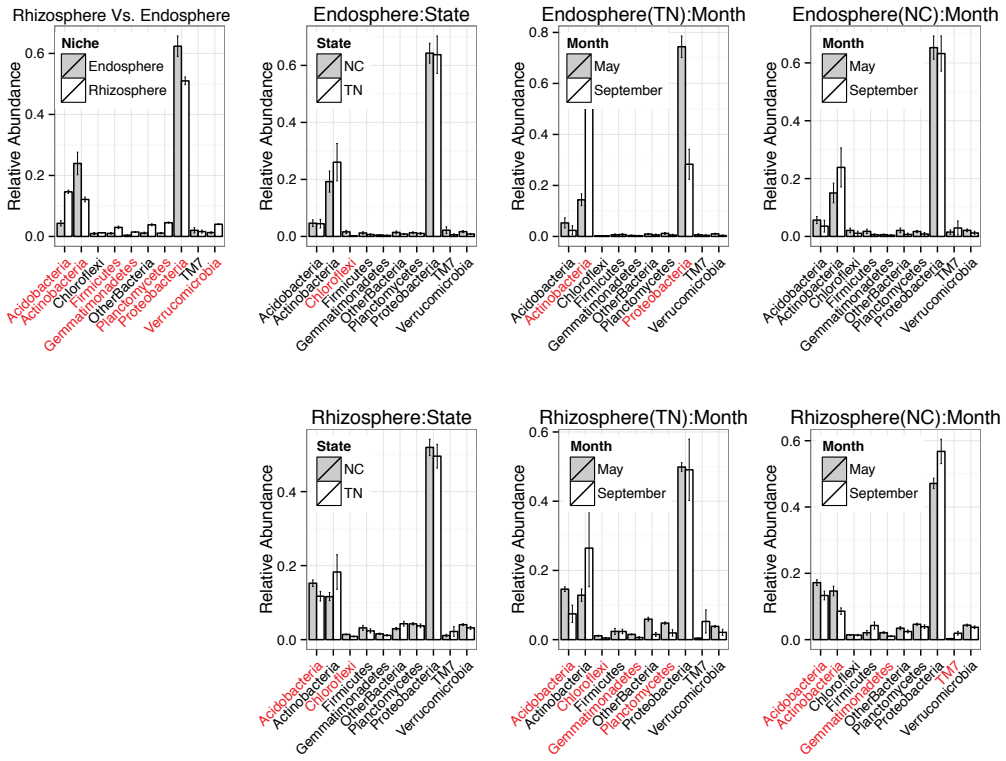


Figure 3.6: Comparative analysis of major bacterial phyla between rhizosphere and endosphere, two populations, and seasons. The significant difference is calculated using t-test between relative abundance of two. Each bar represents its relative abundance and the label colored red represents that the difference is significant.

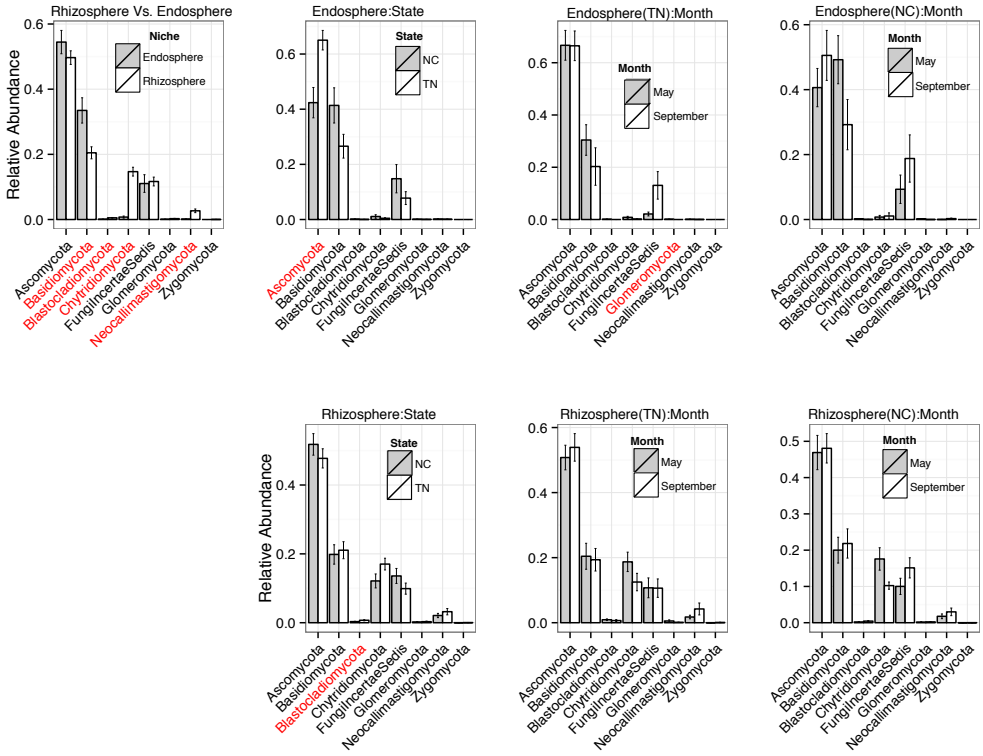


Figure 3.7: Comparative analysis of major fungal phyla between rhizosphere and endosphere, two populations, and seasons. The significant difference is calculated using t-test between relative abundance of two. Each bar represents its relative abundance and the label colored red represents that the difference is significant.

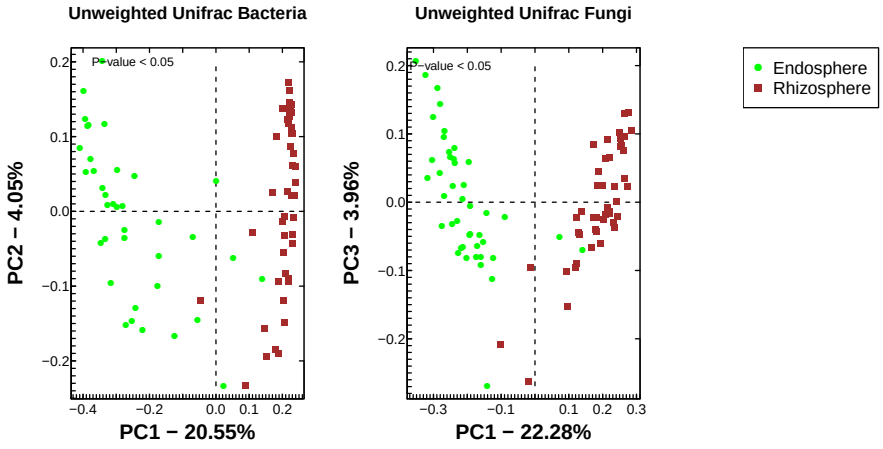


Figure 3.8: Principle Coordinate analysis of Unweighted UniFrac distance for bacterial (left) and fungal (right) communities. The plot indicates the rhizosphere and endosphere communities are distinct for both bacteria and fungi. Average Unweighted UniFrac distance matrix was calculated from 999 even rarefactions of 1000 sequences per sample for bacteria and 400 sequences per sample for fungi. The significance of the difference was calculated using adonis function of vegan package in R.

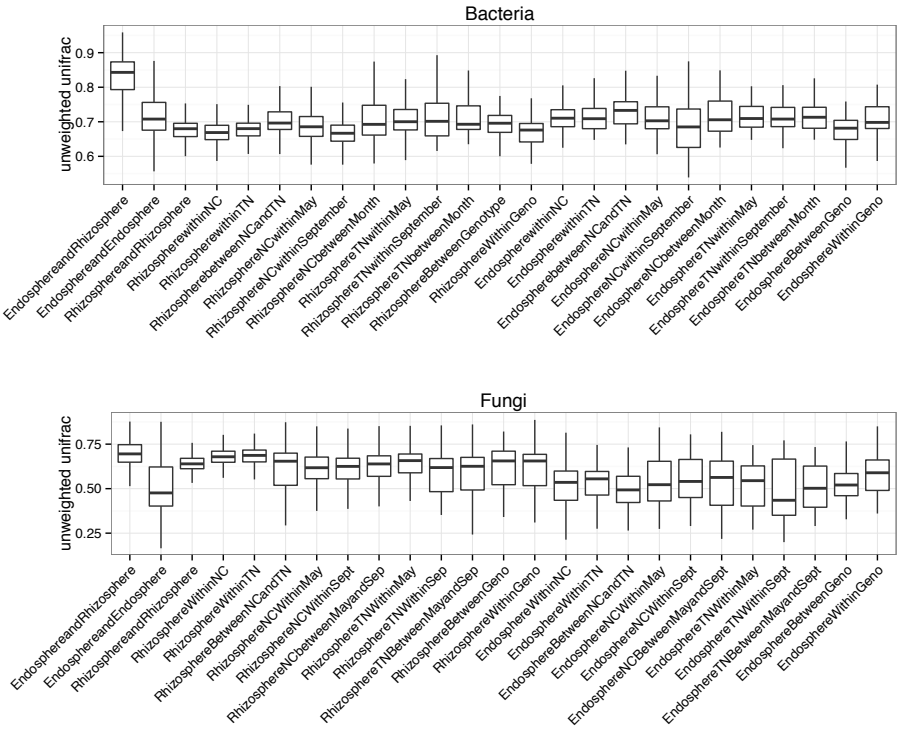


Figure 3.9: Box plot of UniFrac distances comparison between and within niche, population, and season.

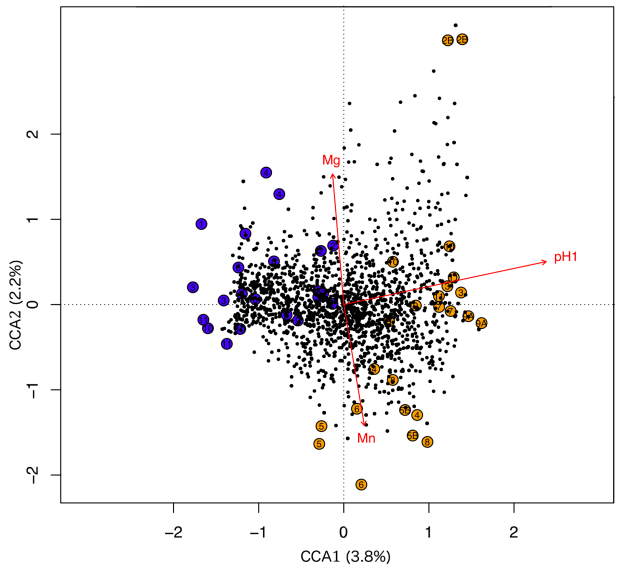


Figure 3.10: CCA of bacterial OTUs from rhizosphere and soil and host factors. Larger circles represent samples that are color-coded based on their location, and smaller dots represent species/OTUs.

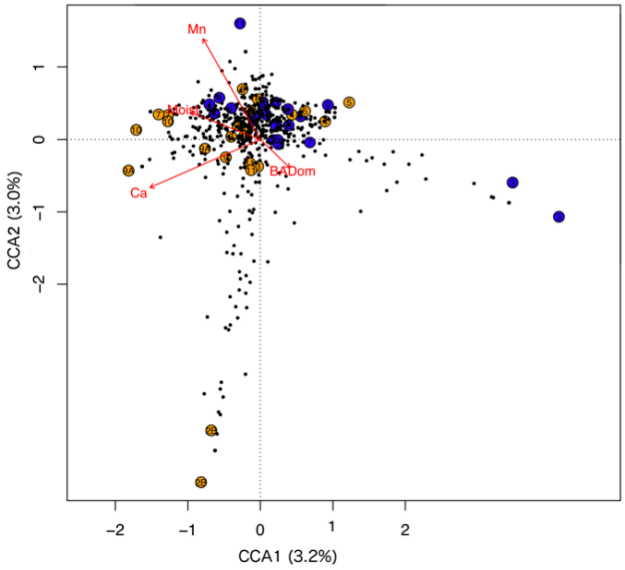


Figure 3.11: CCA of fungal OTUs from rhizosphere and soil and host factors. Larger circles represent samples that are color-coded based on their location, and smaller dots represent species/OTUs.

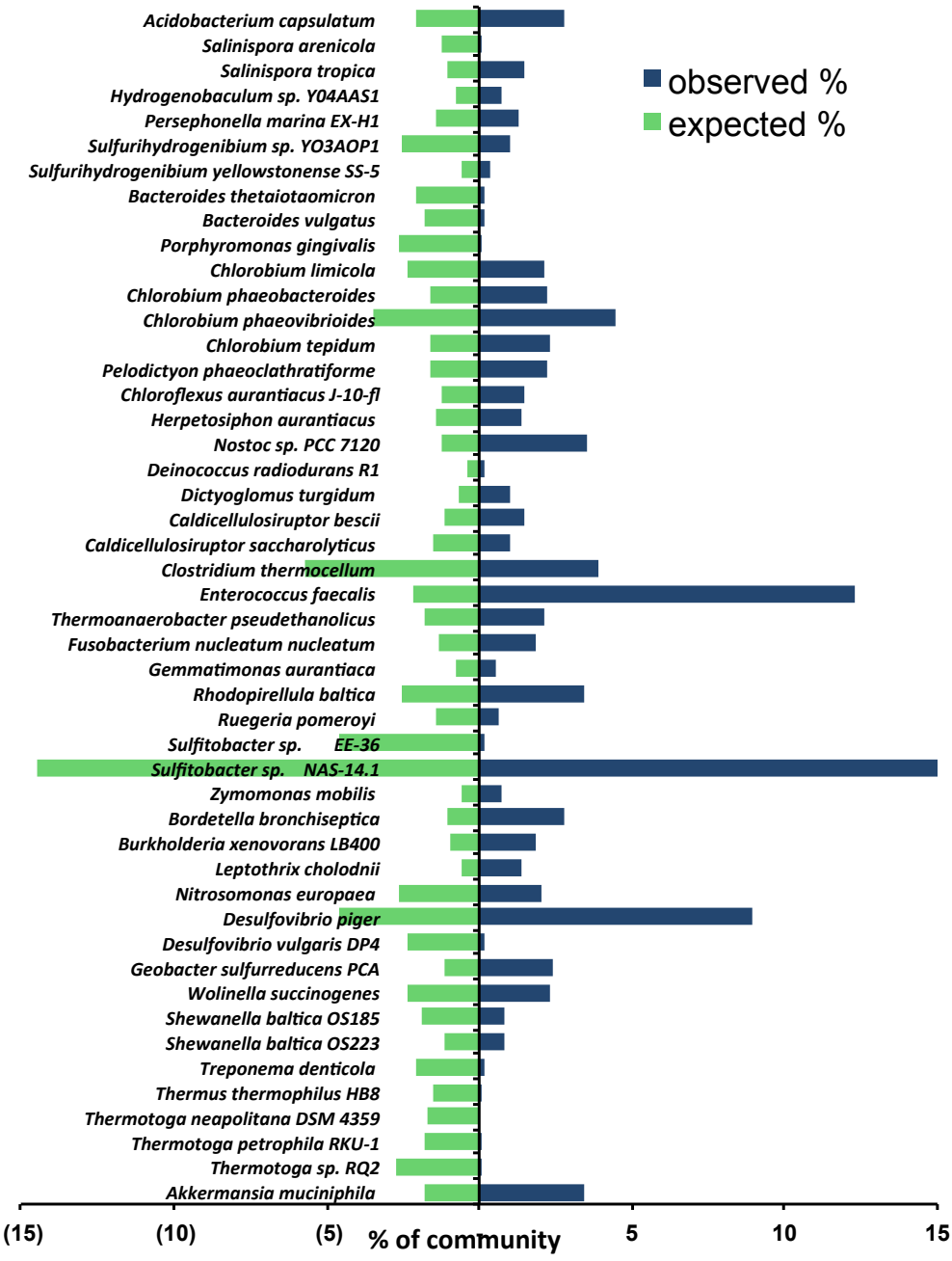


Figure 3.12: A representation of accuracy of V6-V9 primers from this study in characterizing synthetic community from our previous study (Shakya et al., 2013)

Table 3.2: Measurable physical features of *P. deltoides* and its surrounding environment.

TreeID	DBH (cm)	Dist. to river (m)	Dist. to closest Pop. (m)	Tree to sampled roots (m)	# of proximal trees (Prism)	# of proximal Poplar (Prism)	Sample distance to tree (m)	Angle to top of the tree	Angle to base of the tree	Angle to lowest live branch	Canopy width (m)	Basal area (m2)	Basal area per hectare	Basal area per hectare
TNP03	73.5	14.02	3.5	2	4	2	28.65	54.5	-8.75	10.4	16.61	0.4	2.1	1.3
TNP04	122.4	32.9	3.05	2.5	10	4	32.55	57.5	-3.35	21.75	23.85	1.2	12.9	5.9
TNP05	77	15.24	4	3.3	7	5	15.2	62.75	-4.75	36	12.25	0.5	3.7	2.8
TNP06	80	13.6	6.3	1.4	10	1	15	68.5	-1.45	18.25	16.3	0.5	5.5	1
TNP07	68.5	18.4	1.92	0.51	9	2	22.45	56.9	-16.75	21	12.51	0.4	3.7	1.1
TNP08	71	13	no Pop	0.67	8	0	10.9	68.2	-5.75	30.2	10.6	0.4	3.6	0.4
TNP09	68	26.4	5.8	0.46	3	1	18.6	61.5	-1.7	28.55	15.85	0.4	1.5	0.7
TNP10	56	11	no Pop	2.14	5	0	21	45.75	-18.5	15	52	0.2	1.5	0.2
TNP11	61	40	1.6	0.8	3	3	8.45	76.15	-7.7	36.1	9.8	0.3	1.2	1.2
TNP2B	119	25.2	4.7	4	3	1	37.1	38.8	-8.7	9	18.14	1.1	4.4	2.2
TNP5B	87.5	30.8	11	2.2	8	1	12.9	68	-3.9	39.5	15.9	0.6	5.4	1.2
TNP9A	84.3	30.48	14	0.2	6	0	11	39.85	-4.85	6.2	21.8	0.6	3.9	0.6
NCP01	66.3	0 *	20	2.5	4	0	17.98	55.9	-10.7	-7.4	9.41	0.3	1.7	0.3
NCP02	51.7	0	10	1	1	1	11.58	71.65	1.9	19.3	12.05	0.2	0.4	0.4
NCP03	66.4	6.4	1.52	0.5	9	4	17.2	68	-3.5	28	15.6	0.3	3.5	1.7
NCP04	47.25	106.7	6.01	1.98	12	4	12.3	76	-6	24	42.75	0.2	2.3	0.9
NCP05	62.8	55.5	no Pop	6.4	9	0	32.16	45.85	-4.2	22.75	13.1	0.3	3.1	0.3
NCP06	71.5	16.9	no Pop	7	7	0	36.27	44.6	-3.55	14.35	11.83	0.4	3.2	0.4
NCP07	37.7	53.5	0.23	1.59	8	3	17.19	53	-16.65	15	8.23	0.1	1	0.4
NCP08	119.5	70.1	5.48	6.19	11	1	30.48	65.5	-3	25.5	25.9	1.1	13.5	2.2
NCP09	95	2	50	15.24	10	7	25.5	58.1	4.45	22.9	21.89	0.7	7.8	5.7
NCP10	82.5	20	no Pop	6.1	9	0	12.5	72.15	-7.5	40.95	16.15	0.5	5.3	0.5
NCP11	69.6	2.13	no Pop	0.36	14	0	30.66	44.75	-4.2	15.15	14.88	0.4	5.7	0.4

*on stream/creek

Table 3.3: Physical and chemical properties of soil around *P. deltoides* of NC and TN.

TreeID	% Soil Moisture	LBC_1 (ppm)	pH $CaCl_2$	Equiv. H_2O pH	Sand (%)	Silt (%)	Clay (%)	C (%)	N (%)	OM (%)	Ca (ppm)	K (ppm)	Mg (ppm)	Mn (ppm)	P (ppm)	Zn (ppm)
TNP03	16.77	401	6.85	7.45	47.2	35.9	16.8	3	0.24	4.95	2762	88.9	221.3	56.5	148.5	18.2
TNP04	11.61	256	6.25	6.85	75.9	15.9	8.2	1.59	0.12	2.64	1584	44.4	98.2	44.9	151.9	5.5
TNP05	14.29	244	5.55	6.15	55.8	30	14.2	0.85	0.07	1.79	1329	61.9	90.8	46.6	231.2	2.3
TNP06	19.4	399	5.69	6.29	41.9	43	15.1	2	0.16	4.19	1768.5	61.5	97.2	66.1	125.1	5
TNP07	37.11	546	6.6	7.2	21.2	50	28.9	3.84	0.34	7.23	4361	72.1	285.9	74.2	390.7	29.1
TNP08	13.93	228	6.4	7	79.2	13.3	7.5	1.47	0.11	2.89	1650	34.7	93.2	60	205.1	3.7
TNP09	20.05	456	6.26	6.86	20.6	55.3	24.1	2.58	0.22	5.16	2585	109.3	189.2	49.4	132.4	5.5
TNP10	27.08	385	6.86	7.46	64	22.5	13.5	3.97	0.28	6.59	3727	80.1	336.3	61.4	198.2	25.3
TNP11	25.29	382	6.65	7.25	43.9	32.6	23.5	1.88	0.16	3.68	3455	63	182.5	42.9	473.5	16
TNP2B	18.61	596	7.42	8.02	25.2	41.2	33.5	7.38	0.33	7.26	7026	125.5	246.4	4.7	56.3	1.5
TNP5B	18.4	373	6.09	6.69	53.9	36.7	9.4	2.48	0.2	4.35	1972	49.4	112	52.6	70.4	8.5
TNP9A	18.55	504	6.97	7.57	22	50.5	27.5	3.83	0.31	6.67	4967	84.9	220	73.4	307.1	7.8
NCP01	22.97	701	5.12	5.72	35.3	43.2	21.5	3.84	0.24	7.75	2847	94.6	543.2	67.3	159.4	11.6
NCP02	17.53	636	5.88	6.48	35.9	31.9	32.2	5.31	0.38	9.65	3015	345.2	436.2	58.5	31	14.9
NCP03	5.01	186	5.05	5.65	89.9	7.3	2.8	0.81	0.05	1.63	409	48.6	52.1	8.7	9.4	3.9
NCP04	65.49	1303	5.23	5.83	25.9	55.8	18.3	32.65	0.62	17.03	3085	384.1	436.3	31.2	171.2	32.2
NCP05	19.44	455	5.51	6.11	52.6	32	15.4	2.24	0.18	4.73	1493	132.3	177.3	36	15.6	6.1
NCP06	24.24	646	5.25	5.85	18.6	52	29.4	3.1	0.23	7.38	1782	120.3	263.9	38.2	27.6	14.1
NCP07	21.15	507	5.73	6.33	35.2	33.3	31.4	2.51	0.19	5.87	1774	147.9	192.6	39.5	22.7	6.8
NCP08	31.01	579	5.24	5.84	21.2	41.3	37.5	2.72	0.23	7.65	1524	171.5	176	38.3	17.3	8
NCP09	24.32	643	5.14	5.74	21.3	44.7	34	3.14	0.25	7.3	1343	218	177.1	32.4	10.6	7.3
NCP10	24.55	793	4.81	5.41	16	50.6	33.4	3.43	0.27	7.47	1426	174.3	181.9	43.1	30.6	12
NCP11	22.12	831	4.86	5.46	18	47.9	34.1	3.53	0.26	8.21	1602	112.6	307.2	84.7	21.8	11.3

Table 3.6: A list of core bacterial OTUs and their closest relative organism.

OTU	Phylum	Order	Closest sequenced relative	% ID
1211	*Acidobacteria*	*Desulfuromonadales*	*Geobacter sulfurreducens* PCA	89%
3934	*Acidobacteria*	*Desulfuromonadales*	*Geobacter pelophilus* Dfr2	88%
5037	*Acidobacteria*	*Desulfuromonadales*	*Geobacter daltonii* FRC-32 strain FRC-32	92%
16727	*Acidobacteria*	*Solirubrobacterales*	*Solirubrobacter soli* strain Gsoil 355	100%
18613	*Acidobacteria*	*Solibacterales*	*Candidatus Solibacter usitatus* Ellin6076	99%
20548	*Acidobacteria*	*Desulfuromonadales*	*Geobacter psychrophilus* strain P35	90%
27680	*Acidobacteria*	*Acidobacteriales*	*Thermolithobacter ferrireducens* strain KA2	89%
3066	*Actinobacteria*	*Actinomycetales*	*Mycobacterium setense*	100%
4037	*Actinobacteria*	*Nitriliruptoridae*	*Nitriliruptor alkaliphilus* DSM 45188 ANL-iso2	93%
2637 †	*Actinobacteria*	*Actinomycetales*	*Microlunatus phosphovorus* NM-1 $$	99%
24338	*Chloroflexi*	*Anaerolineales*	*Longilinea arvoryzae*	86%
1413	*Proteobacteria*	*Rhizobiales*	*Zavarzinella formosa* strain : A10	87%
5154	*Proteobacteria*	*Rhodospirillales*	*Nisaea nitritireducens* strain DR41_18	93%
8886	*Proteobacteria*	*Xanthomonadales*	*Nevskia soli* strain GR15-1	100%
9201	*Proteobacteria*	*Xanthomonadales*	*Steroidobacter denitrificans* strain FS	94%
10210	*Proteobacteria*	*Rhizobiales*	*Hyphomicrobium sulfonivorans* strain S1	95%
12581	*Proteobacteria*	Methylophilales	*Methylovorus glucosetrophus* SIP3-4	100%
12723	*Proteobacteria*	*Gallionellales*	*Gallionella capsiferriformans* ES-2 strain ES-2	94%
12806	*Proteobacteria*	*Rhodobacterales*	Ponticaulis koreensis DSM 19734	94%
13148	*Proteobacteria*	*Rhizobiales*	*Rhodoplanes elegans* strain AS130	96%
14206	*Proteobacteria*	*Burkholderiales*	*Oxalicibacterium flavum* strain TA17	96%
15435	*Proteobacteria*	*Rhizobiales*	*Bradyrhizobium japonicum*	100%
17447	*Proteobacteria*	*Myxococcales*	*Kofleria flava* strain Pl vt1	92%
19176	*Proteobacteria*	*Syntrophobacterales*	*Desulfoglaeba alkanexedens* ALDC	94%
20376	*Proteobacteria*	*Rhizobiales*	*Agrobacterium radiobacter K84* strain K84	100%
22702	*Proteobacteria*	Caulobacterales	*Caulobacter sp.* strain FWC21	99%
23419	*Proteobacteria*	*Rhizobiales*	*Agrobacterium fabrum* str. C58 strain C58	100%
23459	*Proteobacteria*	*Burkholderiales*	*Variovorax paradoxus* EPS strain EPS	100%
24004	*Proteobacteria*	*Burkholderiales*	*Burkholderia phymatum* STM815 strain STM815	96%
24018	*Proteobacteria*	*Burkholderiales*	*Thiobacter subterraneus* strain C55	96%
24967	*Proteobacteria*	*Rhizobiales*	*Zoogloea oryzae* strain A-7	97%
25294‡	*Proteobacteria*	*Burkholderiales*	*Methylibium fulvum* strain Gsoil 322	100%
25950	*Proteobacteria*	*Burkholderiales*	*Duganella violaceinigra* strain YIM 31327	100%
11162	*Verrucomicrobia*	*Verrucomicrobiales*	*Prosthecobacter fluviatilis* strain HAQ-1	88%
15696	*Verrucomicrobia*	*Verrucomicrobiales*	*Haloferula phyci* strain AK18-024	89%

†rhizosphere specific

‡present in all samples

§All samples

Table 3.4: List of bacterial reads, OTUs, chao indices for all samples. The empty boxes represent samples that failed to amplify.

RHIZOSPHERE							
MAY				SEPTEMBER			
NORTH CAROLINA							
SAMPLE	READS	OTUs	Chao1	SAMPLE	READS	OTUs	Chao1
R.M.N.01	9,201	2,006	5,309	R.S.N.01	7,369	1,336	3,178
R.M.N.02	7,410	1,640	4,787	R.S.N.02	8,777	1,403	3,182
R.M.N.03	15,247	1,626	2,564	R.S.N.03	9,042	873	1,292
R.M.N.04	4,980	1,464	4,032	R.S.N.04	5,434	1,538	3,961
R.M.N.05	4,852	1,654	5,452	R.S.N.05	7,635	1,810	4,517
R.M.N.06	7,438	1,063	1,702	R.S.N.06	2,245	703	1,601
R.M.N.07	3,162	1,108	3,433	R.S.N.07	9,348	2,300	6,270
R.M.N.08	3,126	917	1,947	R.S.N.08	8,061	1,354	2,315
R.M.N.09	10,176	1,693	3,950	R.S.N.09	8,816	1,634	3,813
R.M.N.10	15,502	2,021	4,061	R.S.N.10	10,470	711	1,726
R.M.N.11	8,735	2,051	6,834	R.S.N.11	5,095	968	1,706
TENNESSEE							
R.M.T.03	2,882	1,047	3,508	R.S.T.03	11,210	1,962	5,164
R.M.T.04	5,459	1,837	6,417	R.S.T.04	4,148	342	405
R.M.T.05	6,045	1,825	5,978	R.S.T.05	7,613	1,694	4,370
R.M.T.06	2,253	971	3,219	R.S.T.06	5,722	1,327	3,227
R.M.T.07	3,359	1,218	4,302	R.S.T.07	6,095	1,688	5,059
R.M.T.08	10,057	1,812	4,524	R.S.T.08	8,560	2,289	7,541
R.M.T.09	7,567	2,211	8,235	R.S.T.09	7,355	1,309	2,947
R.M.T.10	2,660	1,124	4,529	R.S.T.10	6,454	1,206	2,982
R.M.T.11	4,065	1,327	3,938	R.S.T.11	9,957	2,445	7,818
R.M.T.2B	5,913	1,099	1,944	R.S.T.2B	9,890	1,689	3,786
R.M.T.5B	2,091	962	3,209	R.S.T.5B	5,312	1,628	5,867
R.M.T.9A	3,998	1,406	4,823	R.S.T.9A	13,987	2,759	7,257
ENDOSPHERE							
NORTH CAROLINA							
E.M.N.01	5,571	37	41	E.S.N.01	7,893	95	103
E.M.N.02	6,003	170	195	E.S.N.02	6,810	68	72
E.M.N.03	8,505	150	245				
E.M.N.04	3,502	103	403	E.S.N.04	8,400	72	77
E.M.N.05	1,668	186	237	E.S.N.05	4,899	104	115
E.M.N.06	6,320	99	132	E.S.N.06	19,091	64	100
E.M.N.07	7,370	406	538	E.S.N.07	9,371	64	77
E.M.N.08	3,667	59	128	E.S.N.08	18,212	97	106
E.M.N.09	10,475	445	655	E.S.N.09	6,764	1,072	2,740
E.M.N.10	4,347	125	210	E.S.N.10	9,157	151	178
E.M.N.11	11,888	316	559	E.S.N.11	18,105	32	33
TENNESSEE							
E.M.T.03	8,514	184	225	E.S.T.03	9,426	63	66
E.M.T.04	3,431	65	71	E.S.T.04	9,765	62	90
E.M.T.05	16,707	96	98	E.S.T.05	7,480	89	100
E.M.T.06	8,020	68	73				
E.M.T.07	13,269	160	184	E.S.T.07	33,800	139	159
E.M.T.08	8,674	103	112	E.S.T.08	21,039	191	196
E.M.T.10	7,191	234	288				
E.M.T.11	8,622	237	251				
				E.S.T.2B	8,717	19	19
E.M.T.5B	14,210	101	112	E.S.T.5B	9,664	95	102
E.M.T.9A	240	63	90				

Table 3.5: List of fungal reads, OTUs, and chao index for all samples. The empty boxes represent samples that failed to amplify.

RHIZOSPHERE							
MAY				SEPTEMBER			
NORTH CAROLINA							
SAMPLE	READS	OTUs	Chao1	SAMPLE	READS	OTUs	Chao1
R.M.N.01	2,878	301	387	R.S.N.01	4,894	248	329
R.M.N.02	6,686	345	467	R.S.N.02	913	92	146
R.M.N.03	3,019	100	107	R.S.N.03	1,459	39	42
R.M.N.04	3,626	298	355	R.S.N.04	797	147	222
R.M.N.05	2,723	135	143	R.S.N.05	2,901	182	208
R.M.N.06	1,369	108	114	R.S.N.06	2,243	151	175
R.M.N.07	3,066	297	530	R.S.N.07	880	147	222
R.M.N.08	2,419	195	248	R.S.N.08	1,398	109	141
R.M.N.09	2,942	247	319	R.S.N.09	2,734	200	235
R.M.N.10	3,278	206	216	R.S.N.10	5,877	134	149
R.M.N.11	2,039	243	395	R.S.N.11	1,271	77	96
TENNESSEE							
R.M.T.03	2,110	43	45	R.S.T.03	1,223	166	259
R.M.T.04	2,557	51	54	R.S.T.04	1,805	22	24
R.M.T.05	3,244	80	89	R.S.T.05	723	110	136
R.M.T.06	2,263	68	76	R.S.T.06	2,015	142	166
R.M.T.07	1,305	193	399	R.S.T.07	2,765	119	136
R.M.T.08	4,101	346	588	R.S.T.08	1,548	126	153
R.M.T.09	9,749	543	905	R.S.T.09	2,692	95	99
R.M.T.10	3,196	228	575	R.S.T.10	4,888	126	137
R.M.T.11	3,762	245	430	R.S.T.11	5,655	132	139
R.M.T.2B	3,050	90	104	R.S.T.2B	8,660	336	437
R.M.T.5B	2,920	103	156	R.S.T.5B	2,916	97	103
R.M.T.9A	1,799	243	414	R.S.T.9A	4,621	237	271
ENDOSPHERE							
NORTH CAROLINA							
E.M.N.01	3,749	21	22	E.S.N.01	5,180	43	44
E.M.N.02	3,561	23	26	E.S.N.02	3,013	21	21
E.M.N.03	1,635	15	15				
E.M.N.04	9,200	25	46	E.S.N.04	1,888	8	8
E.M.N.05	1,395	25	28	E.S.N.05	1,912	169	184
E.M.N.06	1,010	99	149	E.S.N.06	819	22	23
E.M.N.07	960	31	53	E.S.N.07	1,394	58	97
E.M.N.08	4,815	62	72	E.S.N.08	4,294	45	62
E.M.N.09	1,129	63	76	E.S.N.09	1,627	28	30
E.M.N.10	1,918	20	29	E.S.N.10	1,902	48	56
E.M.N.11	408	27	45	E.S.N.11	2,753	96	125
TENNESSEE							
E.M.T.03	1,828	42	45	E.S.T.03	1,806	36	36
E.M.T.04	1,398	14	45	E.S.T.04	2,860	19	20
E.M.T.05	2,390	55	15	E.S.T.05	2,791	63	81
E.M.T.06	1,547	36	66	E.S.T.06	2,733	9	12
E.M.T.07	1,037	15	41	E.S.T.07	2,297	32	32
E.M.T.08	2,241	35	15	E.S.T.08	1,951	38	45
E.M.T.09	2,286	34	40	E.S.T.09	2,320	57	60
E.M.T.10	4,043	46	35	E.S.T.10	3,680	35	37
E.M.T.11	3,128	28	47	E.S.T.11	1,683	11	14
E.M.T.2B	66	24	28	E.S.T.2B	2,110	13	13
E.M.T.5B	1,599	28	42	E.S.T.5B	2,695	56	64
E.M.T.9A	2,929	29	30	E.S.T.9A	15,011	37	46

Table 3.7: A list of core fungal OTUs and their closest sequenced relatives.

OTU #	Phylum	Order	Closest sequenced relative	% ID
294	*Ascomycota*	*Chaetothyriales*	Exophiala tremulae	100%
413	*Ascomycota*	*Hypocreales*	*Metarhizium anisopliae*	100%
1467[§]	*Ascomycota*	*Hypocreales*	*Neonectria* sp.	100%
2332	*Fungi incertae sedis*	*Mortierellales*	*Mortierella* sp.	100%

Chapter 4

Characterizing archaeal communities in rhizosphere of mature trees and surrounding bulk soils from a riparian zone

Disclosure: This chapter is a manuscript under preparation. Migun Shakya was responsible for most experimental work and data analysis.

4.1 Abstract

Archaea are common members of rhizosphere microbial communities, but most of our understanding of below ground archaeal communities are derived from soil, not rhizosphere. These resident archaea that includes *Thaumarchaeota* have been implicated in nitrogen cycling and potentially could play role in plant nutrition. Based on environmental and economical importance of mature trees, it is important to characterize their resident archaea. Here, we used barcoded pyrosequencing to characterize archaeal community structure in roots of mature trees with focus on *P. deltoides*. A total of 42 samples - 6 *P. deltoides* rhizosphere and 3 bulk soils and 3 non *Populus* trees surrounding each of the 6 *Populus* tree - were included in the study. We used two genes: V1-V3 of 16S rRNA and *amoA* (ammonia monooxygenase subunit A) to survey archaeal and ammonia oxidizing archaeal (AOA) communities. Our results revealed relatively low diversity of archaea in both soils and rhizosphere compared to corresponding bacterial and fungal communities. Pairwise comparison of 16S based phylotype of rhizosphere and soil from each site revealed slightly greater diversity of archaea in rhizosphere. However, the community structure of rhizosphere and soil communities did not differentiate based on their niches. Additionally, 95% of 16S and 66% of *amoA* based phylotypes were common between two niches and only 1 16S phylotype was significantly enriched in rhizosphere. The study, however, showed a wide diversity of archaea - either affiliated to previously known lineages of ammonia oxidizing archaea or novel - are associated with mature trees and soils from riparian zones.

4.2 Introduction

Archaea are common in the rhizosphere of *Zea mays* (maize) (Chelius et al., 2001), *Oryza sativa* (rice) (Großkopf et al., 1998), *Sullius bovines* (pine) (Bomberg et al., 2003) etc. *Euryarchaeota* and *Thaumarchaeota* - a recently coined phylum that consist of non-thermophilic *Crenarchaeota* - are two phyla usually found in rhizosphere and most bulk soils. These phyla comprise of organisms that could potentially carry out oxidation of ammonia (NH_3) to nitrite (NO_2^-), a critical step in nitrogen cycle (Tourna et al., 2011)

and methanogenesis (Conrad, 2007). Thus, studying archaea-plant interface is an important step towards understanding plant microbe interface and its impact on the environment.

Since the time archaea was first discovered from a non-thermophilic soil there have been many studies that characterized archaeal communities from soil (Bates et al., 2011; Gubry-Rangin et al., 2011; Pester et al., 2012; Onodera et al., 2009). However, few studies have focused on rhizosphere, most of which were conducted on agricultural sites (Chen et al., 2008; Di et al., 2009). Beside its importance in global economy, climate change, and environment, only few studies have attempted to characterize archaeal community from mature perennial plants (Bomberg et al., 2011; Bomberg and Timonen, 2009). Therefore, many properties of archaeal communities associated with mature perennial plants remain unknown. For instance, the rhizosphere effect - change in diversity or abundance of microorganisms in the rhizosphere compared to surrounding bulk soils (Hiltner, 1904) - is an important component for understanding the plant microbe interface remains unclear for archaea. Although rhizosphere effect on archaeal communities has been reported in macrophytes like *Littorella uniflora* (American shoreweed) (Herrmann et al., 2008) and limited studies of oak and pine trees (Sliwinski and Goodman, 2004), it has not been reported in *Populus*.

Populus is a perennial plant that is used for production of lumber, pulp, paper, and biofuel (Tuskan et al., 2006). It has an extensive root system that harbors diverse groups of bacteria and fungi (Gottel et al., 2011) that have direct effect on its growth and development (van der Lelie et al., 2009). Based on the presence of archaea in diverse plants, it is likely that archaea are present in rhizosphere of *P. deltoides* and may play important roles in their hosts' health and development. Molecular techniques like PCR, NGS, and advanced bioinformatic tools have made studies of microbial community possible without the need of isolating and culturing its members. These tools allow us to study microbial communities at unprecedented scales. Our previous studies have used similar approach to characterize the community structure and the factors that are structuring bacterial and fungal communities of *P. deltoides* but did not examine archaea.

The objective of present study is to characterize archaeal communities in bulk soils and rhizosphere of naturally occurring mature trees including *P. deltoides*. We seek to test for the rhizosphere effect by comparing the rhizosphere communities to adjacent bulk

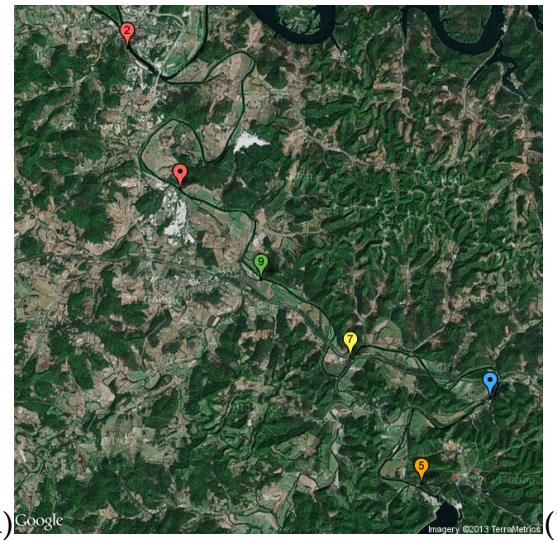
soils. We used two marker genes (V1-V3 region of 16S rRNA gene and partial *amoA* gene) for in-depth characterization. Results of our study suggest a low diversity of archaea including AOAs in both soils and rhizosphere compared to other microbes like bacteria and fungi. Furthermore, relatively higher diversity of archaea were recovered from rhizosphere compared to corresponding bulk soils. However, the community structures were not distinct based on their niches, not even for *P. deltoides* rhizosphere. Regardless, the study showed a wide diversity of archaea - either affiliated to known lineage of ammonia oxidizing archaea or not - are present in the soils and mature trees from a riparian zone.

4.3 Methods

Study site and sampling.

We collected soil and root samples along Caney Fork river in Cookeville, Tennessee on September of 2011. The study includes a subset of *P. deltoides* from our previous study of 2010 (See Chapter 3) along with additional bulk soils and rhizosphere samples from non-*Populus* trees. All samples were collected from 6 *P. deltoides* and 3 bulk soils and 3 non *Populus* tree surrounding each *P. deltoides* in september 2011. Root samples were collected by carefully excavating and tracing the roots back to the tree to ensure identity. The collected root samples were stored in ice and processed next day in lab. Furthermore, core soil samples were collected from areas between the *P. deltoides* and each non *Populus* trees. Similarly, soil samples were refrigerated to be processed next day. A total of 42 samples with 18 bulk soil and 24 root samples were collected. A list of all the samples and their source are listed in Table 4.1. From the root samples, tertiary fine roots were removed, and loosely adhered soils from these roots were removed by shaking and then washed with 100 ml of 10mM NaCl solution to remove the adhering rhizosphere soil.

The resultant wash was collected in 50mL tubes, which made up the rhizosphere samples. Part of the soil sample was used for extracting DNA and part was sent to Agricultural and Environmental Services Laboratory (AESL) of University of Georgia (`http://aesl.ces.uga.edu/`) for physical and chemical characterization. The soil characteristics of each soil are presented in Table 4.2.



(a)

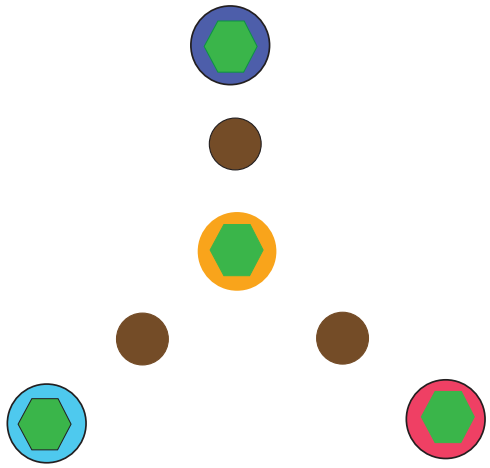

(b)

Figure 4.1: Location of trees. (a) Location of sample units (*P. deltoides*) along Caney Fork river in Tennessee. (b) An example of experimental design at one sampling unit. The symbol in the middle represents the position of *P. deltoides*, surrounded by bulk soil samples and corresponding mature trees that are not *Populus*. Common name of non *Populus* trees are listed in Table 4.1

DNA extraction

For rhizosphere samples, 2.0 ml of samples that were pelleted using a low-speed centrifugation was used for extractions using PowerSoil DNA extraction kit (MoBio, Carlsbad, CA). For soil samples,a 250mg of soil was weighed and used for extraction using PowerSoil DNA extraction kit (MoBio, Carlsbad, CA).

DNA amplification and sequencing for 16S rRNA gene.

PCR amplifications of archaeal 16S rRNA gene from the genomic DNA of 42 samples (18 bulk soil and 24 rhizosphere) were conducted using a pair of primer that targets V1-V3 region of 16S rRNA gene. The primers used are F2A-(5'-TCYSGTTGATCCYGCSRG-3') and 571R-(5'-GCTACRGVYSCTTTARRC-3') (Shakya et al., 2013) that were tagged with 8 bp barcode which was preceded by the 454 B sequence (5'-CTATGCGCCT TGCCAGCCCGCTCAG-3') in the forward primer and just the 454 A sequence in reverse primer (5'-CGTATCGCCTCC CTCGCG CCATCAG-3') (454 Life Sciences, Branford, CT, USA). A list of all the primers with their barcode are listed in Table 4.3. For each sample, a 50 μl PCR reaction was conducted using 1X High Fidelity PCR buffer, Platinum Taq High

Fidelity polymerase (Invitrogen, Carlsbad, CA), 0.2mM of deoxynucleoside triphosphates (dNTPs), 2mM $MgSO_4$, and 300nM of each primer. The samples were amplified using following thermal condition: 94°C for 2 minutes, then 30-35 cycles of 94°C for 30s, 55°C for 45s, and 72°C for 1 min followed by 72°C for 7 min before cooling at 4°C . Amplicons were sequenced using a 454 Life Sciences Genome Sequencer FLX (Roche Diagnostics, Indianapolis, IN, USA) at Oak Ridge National Laboratory, USA.

DNA amplification and sequencing for amoA gene.

PCR amplification of archaeal *amoA* gene from the subset of 42 samples were conducted using a pair of primer that targets part of the gene. The primers used are Camo19F (5'-ATGGTCTGGYTWAGACG-3') and Camo616R (5'-GCCATCCABCK RTANGTCCA-3') (Pester et al., 2012). The forward primer was tagged with 6-7 bp barcode which was preceded by the 454 B sequence and the reverse primer was preceded by the 454 A sequence. A list of all the primers with their barcode are listed in Table 4.4. For each sample, a 50 μl PCR reaction was conducted using 1X High Fidelity PCR buffer, Platinum Taq High Fidelity polymerase (Invitrogen, Carlsbad, CA), 0.2mM of deoxynucleoside triphosphates (dNTPs), 2mM $MgSO_4$, and 300nM of each primer. The samples were amplified using following thermal condition: 95 °C for 2 minutes, then 30-33 cycles of 95°C for 30s, 50°C for 1 min, and 72°C for 1 min followed by 72°C for 5 min before cooling at 4°C . Amplicons were sequenced using a 454 Life Sciences Genome Sequencer FLX (Roche Diagnostics, Indianapolis, IN, USA) at Oak Ridge National Laboratory, USA.

Sequence analysis for 16S.

We denoised pyrosequencing data using Ampliconnoise v1.27 (Quince et al., 2011), which corrects for both PCR and sequencing errors using default settings. Ampliconnoise was implemented through QIIME 1.5.0 (Caporaso et al., 2010a). The denoised sequences were then checked and removed off chimeras using uchime (Edgar et al., 2011) implementation through mothur (Schloss et al., 2009) with most abundant sequence in a sample as the template. Additionally, sequences that were shorter than 350 bp were also removed from further analysis. The high quality sequences were then used to cluster (99% sequence

similarity) into phylotypes using uclust (Edgar, 2010) implementation in QIIME. The representative sequences from each phylotypes were assigned a taxonomic classification using RDP classifier 2.2 (Cole et al., 2009) against green genes taxonomy (DeSantis et al., 2006). Furthermore the representative sequences were checked for chimeras using ChimeraSlayer (Haas et al., 2011) against the database provided with the package, although none were found. The resultant phylotypes were then filtered based on there sequence composition to have only those phylotypes that were present in at least 10% (4 samples) of all samples and had at least 20 sequences total. Furthermore, 3 phylotypes that were either unclassified or classified as bacteria were removed for further analysis. Each samples was then rarified to have equal number of sequences to remove any sampling biases $\sim$ 2999.

Reference amoA sequences.

A manually curated database with previously reported archaeal *amoA* gene was created with method based on Gubry-Rangin et al. (2011). First, National Center for Biotechnology Information (NCBI) GenBank database was searched with Entrez search terms "*amoA* and archaea" on March 18 2013. These sequences were then filtered for sequences that were too long (>750) or too short (<500) or have any ambiguous bases. Furthermore, duplicate sequences were removed and then location of the primer that was used in this study was trimmed off from both ends. The sequences that begin after the primer or shorter than the reverse primer used in the study were discarded off as well. Furthermore, sequences were then checked for stop codon in all three frames using a custom python code. Sequences that passed these thresholds were retained along with *amoA* sequences from some of the characterized ammonia oxidizing archaea. The sequences were then clustered at 85% sequence similarity based on Gubry-Rangin et al. (2011) and Pester et al. (2012) using uclust implementation in QIIME. The representative sequences from each cluster were then used as reference to cluster the sequenced *amoA* sequences.

Sequence analysis for amoA.

Similar to 16S, *amoA* sequences were also denoised using Ampliconnoise (Quince et al., 2011) followed by removal of chimeras using uchime (Edgar et al., 2011) (without the reference

sequence) implementation of mothur (Schloss et al., 2009). Furthermore, the sequences were checked for in frame for translation using a custom python script. Any sequences that have stop codon in any of the three frames were removed from downstream analysis. Also, sequences that were less than 350 nucleotide long were also removed from further analysis. The resultant sequences were the clustered against reference sequence at 85% sequence similarity using uclust (Edgar, 2010) implementation of QIIME.

Phylogenetic analysis of sequences representing 16S and amoA based phylotypes.

Phylogenetic analysis that includes sequence processing and tree construction was done using a suit of tools that includes Geneious 5.3 (Drummond et al., 2011), MEGA v5 (Tamura et al., 2011), PyNAST (Caporaso et al., 2010b), QIIME v1.5 (Caporaso et al., 2010a), and greengenes database (DeSantis et al., 2006). Representative sequences of 16S based phylotypes were aligned against greengenes reference database using PyNAST implementation in QIIME v1.5. The alignment was manually edited in Geneious v5.3, which was then uploaded into MEGA v5 for model selection and constructing phylogenetic tree. Based on the AIC score Kimura 2 parameter using discrete gamma distribution was chosen as a substitution model. Similarly, representative sequences of *amoA* based phylotypes were aligned based on codon in geneious, and was further edited as in the geneious as well. Likewise, the alignment was uploaded in MEGA v5 for model selection and construction of phylogenetic tree. Phylogenetic tree was then constructed using Kimura 2 parameter with discrete gamma distribution as the substitution model.

4.4 Results and discussion

Site characteristics.

Our study included 42 samples (18 bulk soils, 18 non *Populus*, and 6 *P. deltoides* rhizosphere) collected from riparian zones of Caney Fork river in Cookeville, TN (Figure 4.1). The six sites are named 10, 11, 2b, 5b, 7, and 9a. For every *P. deltoides* rhizosphere, we sampled 3 bulk soils and 3 non *P. deltoides* rhizosphere from its surroundings. A list of all samples and names of non *Populus* trees are shown in Table 4.1. We extracted DNA from all samples

and characterized physical and chemical features of all bulk soils (Table 4.2). Based on hierarchical clustering of measured factors, one sample (10.2) adjacent to *P. deltoides* 10 was different from all other soil samples (Figure 2), but the remaining samples clustered based on there location. Only Mn, P, and Zn showed high variation of up to 1 fold and rest of the samples variation were under 1 fold. In summary, the bulk soils along riparian zone did not vary much within our sample sites.

Archaea in rhizosphere and soil.

We used V1-V3 region of 16S rRNA gene to describe archaeal diversity. A total of 455,660 high quality sequences - denoised, dechimerized, and trimmed to 350 bp - were recovered from 41 soil and rhizosphere samples (one soil sample failed to amplify) with an average of 11,115 sequences per sample and range of 3,502 to 20,601 sequence per sample. These high quality reads clustered into 1,595 phylotypes (defined at 99% sequence similarity), most of which were singletons and only present in few samples. After removing rare phylotypes - $<$ 10% (4 samples) of the samples and $<$20 sequences total - and ones that classified as bacteria, the number of phylotypes dramatically decreased to 63. However, removal of those phylotypes accounted only 1.4% of total sequence. Based on high number of reads, low number of phylotypes per sample, and rarefaction curves from this study (Figure 1) it is possible that we might have captured most of the archaea detectable using the primer pairs used in the study. However, biases due to primer mismatches could undervalue actual diversity (Shakya et al., 2013). The observed lower diversity of overall archaea in below ground environment is on par with average archaeal diversity in soils from around the world (Auguet et al., 2009; Bates et al., 2011).

Out of 63 phylotypes, 6 belonged to phylum *Euryarchaeota*, 4 to Marine Benthic Group A, and rest (53 phylotypes) were similar to genus *Nitrososphaera* of *Thaumarchaeota*. Among all taxonomic groups, *Nitrososphaera* was the most dominant. This group, which consist of potential ammonia oxidizing archaeon (AOA) represented $\sim$ 99% of overall sequences, 95%- 99% of sequences per sample, and all phylotypes with relative abundance greater than 5% per sample. It is clear that regardless of habitat, riparian zone below ground archaeal communities are dominated by small number of taxa. Similar pattern of

dominance by few archaeal phylotypes have been observed in rhizosphere of crop plants like maize and soybean (Nelson et al., 2010).

The phylogenetic placement of sequences representing 13 most abundant phylotypes, all of which were classified as *Thaumarchaeota* are shown in Figure 4.2. All the sequences belong to group 1.1b crenarchaeota clade, the dominant member of soil archaeal community (Auguet et al., 2009; Bates et al., 2011). 7 of 13 abundant phylotypes had close affiliation to *Nitrososphaera* from soil and rest formed separate lineages. Out of 13 dominant phylotypes, four ((phylotype # 188 (Avg.:28.4%), 1029 (21.8%), 1125 (16%), and 1243 (12.5%)) accounted for ∼80% of total sequences. These phylotypes, however, showed great variation across all samples and only one phylotype (# 1243) was present in all 41 samples and varying from <1% to 52%. Phylogenetic placement of these four abundant phylotypes suggest two (1125 and 1243) are closely related to *Nitrososphaera* isolated from garden (Tourna et al., 2011) and agricultural soil (Kim et al., 2012), and two formed clades with uncultured representatives (Figure 4.2). A BLAST search of two phylotypes in the uncultured clade against cultured or isolated archaea revealed that the closest one (∼90% percent identity) is *Candidatus* "Nitrososphaera gargensis" (Hatzenpichler et al., 2008). Clearly, these phylotypes' represent novel archaea that have only been identified through SSU rRNA. However, given the affiliation of these phylotypes to *Thaumarchaeota* and their high abundance, it is likely that they are important part of nitrogen cycle functions of their host and environment (Leininger et al., 2006).

Putative ammonia oxidizing archaea in rhizosphere and soil.

With *amoA*, a total of 193,657 high-quality sequences - denoised, dechimerized, checked for translation frame, and trimmed to 350 bp - were recovered from 18 soil and rhizosphere samples with an average of 10,759 sequences per sample and range of 986 to 18,985 reads per sample. The 18 samples comprised rhizosphere of all 6 *P. deltoides* and a bulk soil and a non *Populus* tree's rhizosphere surrounding it. These reads clustered into 171 *amoA* phylotypes at 85% sequence similarity against reference clusters derived from high quality *amoA* sequences deposited in NCBI (Gubry-Rangin et al., 2011; Pester et al., 2012). Similar to 16S phylotypes, most (54%) *amoA* phylotypes were singletons. Thus, after removing

the rare ones (present in < 2 samples and <20 sequences total.), we recorded 22 *amoA* phylotypes.

Similar to distribution of 16S based phylotypes, all samples were dominated by six phylotypes (90%). Furthermore, phylotype # 15 accounted for 72% of total sequences and was present in all, but one sample. Another phylotype # 80 was the second most dominant phylotype comprising 14% of all sequences. As these phylotypes were clustered against a reference database, # 15 also represents most abundant cluster of reference sequences from NCBI as well. These results suggest that a group of most abundant potential AOAs in this study and possibly globally, still lack genomic information. We also found one phylotype that did not cluster with any representative reference sequences (phylotype labeled None56), thus representing a novel phylotype that have not been detected before. Additionally, the blast search of 'None56' against all NCBI sequences did not reveal any sequence with 100% match.

Phylogenetic analysis of representative sequences from *amoA* phylotypes with cultured representatives of *Thaumarchaeota* revealed diverse groups of putative AOAs in our samples (Figure 4.2(b)). Five phylotypes were affiliated to lineage with known archaea and rest formed a separate lineage with no affiliations to cultured/isolated/enriched/sequenced AOAs. In contrast with 16S based phylotypes, *amoA* primers in the study are able to detect a phylotype (phylotype # 101) that was closely related to AOAs from marine environment. Additionally, the only phylotype that clustered with reference sequence representing a cultured archaea was phylotype # 69. It clustered with *Ca.* "N. gargensis Ga9.2". Both abundant phylotypes # 15 and # 80 belonged to a separate clade that did not have sequenced or cultured AOAs. In agreement with the archaeal diversity based on 16S, the study suggests that most abundant, and likely important phylotypes from soil and rhizosphere environments are still uncharacterized.

The 16S primer set used in the study was also tested against archaeal 'synthetic communities' in a previous study (Shakya et al., 2013). The primer set, although missed certain *Thermoproteales*, it was able to reconstruct most of the original diversity. Therefore, 16S primers used in the study should be able to capture wide diversity of archaea, if present. *amoA* primer used in the study has been used to study communities of putative AOAs Pester et al. (2012). Comparison of archaeal diversity revealed from 16S and *amoA* based primer

sets revealed that former captured higher number of potential AOAs (*Thaumarchaeota*) than the latter. However, by using *amoA* gene, we were able to detect novel phylotypes that were potentially missed by the 16S gene. Nevertheless, the general community structure revealed by both marker genes were identical, with many rare phylotypes and few dominant ones.

Comparative analysis of archaea in rhizosphere and soil.

We discovered an average of 21 phylotypes per sample from rhizosphere with a range of 11 to 30 per sample. The average phylotypes per bulk soil sample was 18, with a range of 12 to 25 per sample. The higher average in rhizosphere richness may indicate its capability to support higher diversity of archaea than corresponding bulk soils. Additionally, pairwise comparisons of average rhizosphere phylotypes in each of six sites with surrounding bulk soil also show similar trend. 4 sites had higher rhizosphere diversity and 2 sites had equal number of phylotypes in both rhizosphere and soil. Similar richness of phylotypes have been recorded for archaea in a wide variety of plants (Sliwinski and Goodman, 2004), but contrasting results were observed for bacterial communities in oak (Uroz et al., 2010). These observations suggest factors that select bacteria and archaea in rhizosphere may be different. This is in agreement to Valentine (2007), who suggested adaptation to energy stress controls the ecology and evolution of archaea while bacteria become adapted to maximize energy availability. However, distribution of 16S based phylotypes across rhizosphere and soil revealed 95% were shared between two niches, 1 was unique to soil, and 2 to rhizosphere (Figure 4.3(a)). Phylotype unique to soil was in relatively low abundance ($< 1\%$) and specific to single site (site 11). Similarly, the rhizosphere specific phylotypes were also ($< 1\%$) low abundance. These observations suggest only few archaeal phylotypes were mutually exclusive to either of the niches and although we detected higher diversity of archaea in rhizosphere compared to soil, the archaea residing in rhizosphere are capable of proliferating in soil and vice versa.

Principal coordinate analysis of Bray-Curtis dissimilarity measure (based on phylotype abundance) and Sorensen Dice coefficient (based on presence absence of phylotypes) did not reveal clustering of samples based on rhizosphere and soil (data not shown). These results suggest lack of niche partitioning of archaeal communities into bulk soils and

rhizosphere. However, at the phylotype level, out of 13 abundant phylotypes, one (# 1125) was significantly enriched in rhizosphere (Figure 4.4 (c)). This phylotype was closely related to sequences found in the BLAST data base that have been recovered from trembling aspen (Lesaulnier et al., 2008).

Based on *amoA*, we detected an average of 13 phylotypes in soil with a range of 10 to 18 per sample, and an average of 11 phylotypes in rhizosphere with a range of 4 to 17. However, pairwise comparisons of *amoA* based phylotypes revealed no significant differences in rhizosphere or soil. The lack of clear richness of AOAs between rhizosphere and soil suggest that AOAs might not have preference over one site or another. However, we detected some rhizosphere and soil specific phylotypes as only 14 phylotypes were shared, 4 were unique to soil, and 3 to rhizosphere (Figure 4.3). pH is known as one of the main driver of AOAs (Gubry-Rangin et al., 2011). We did not observe wide range of pH among all our soil samples and since we don't expect much variation in pH within few meters of the site, we did not observe variation between the distribution of these AOAs either. Furthermore, distinct *amoA* communities were not observed using beta-diversity measures like Bray-Curtis and Sorenson dice and no phylotypes showed significant enrichment in either rhizosphere or soil. The lack of difference in chemically and physically similar environments are in par with studies that have only revealed differences in community structure of potential AOA in significantly diverse environment types Alves et al. (2013); Gubry-Rangin et al. (2011); Pester et al. (2012).

P. deltoides and surrounding tree species

Based on 16S, our study discovered an average of 19 phylotypes per sample from rhizosphere of *P. deltoides* with a range of 11 to 30 per sample. Similarly, an average of 22 phylotypes per sample with a range of 12 to 26 per sample were recovered from rhizosphere of non *Populus* trees. And, as mentioned previously, the average phylotypes in soil was 18. Compared to *P. deltoides* rhizosphere, non *Populus* show slightly higher richness, however, we do not have enough observations and equal number of samples to claim non *Populus* rhizosphere to be enriched with higher diversity of archaea. Similarly, we detected 44 phylotypes from rhizosphere of *P. deltoides*, 60 from non *Populus*, and 61 from bulk soils and none of the

phylotypes were unique to *P. deltoides*. Furthermore, pairwise comparison of phylotypes of each *P. deltoides* rhizosphere to its surrounding non *Populus* and bulk soil showed varying results. For example, in 4 sites, average phylotype of non *Populus* tree were higher than corresponding *P. deltoides* and in other two sites, it was lower. Ternary plots based on the average abundance of these phylotypes also show that most of the phylotypes, including the abundant ones are almost equally distributed in all three sites (Figure 4.4). Some rare ones, based on 16S, however showed to be slightly enriched in soil, but most of the phylotypes were discovered from all three niches. A similar inconsistent results were also observed in phylotypes based on *amoA*. Based on partial *amoA*, we found 15 phylotypes in *P. deltoides*, 18 in soil, and 14 in non *Populus* rhizosphere. Two phylotypes were unique to *P. deltoides* and four to soil. None of *amoA* based phylotypes' relative abundance differed between *P.deltoides* and soil or other trees. One 16S based phylotype showed significant enrichment in *P. deltoides* compared to non *Populus*, but no significant difference in abundance was observed between bulk soil and *P. deltoides* (Figure 4.4). However, none of the other abundant phylotypes were specifically enriched in one rhizosphere to another or to a soil environment. These observation suggest that the AOAs and archaea in general don't have preference over soil, rhizosphere of *P. deltoides*, and rhizosphere in general.

4.5 Conclusions

Here we present a study that characterizes the diversity and community structure of archaea and AOAs from rhizosphere of mature trees including *P. deltoides* and soil in a riparian zone. The study reveals that rhizosphere and bulk soils have similar community structure consisting of few dominant and many rare phylotypes. It also suggests that archaeal community structure don't vary much within the narrow range soil physical and chemical properties. However, the study reveals archaea from rhizosphere and soil demonstrates that there is a high variability in structure and relative abundance of dominant phylotypes regardless of their habitat. Most archaeal groups, rare or abundant, did not show preference to soil or rhizosphere of *P. deltoides*, or rhizosphere in general. Although all the samples were dominated by phylotypes belonging to genus *Nitrososphaera*, a putative ammonia oxidizer, study like ours cannot confirm their role in ammonia oxidizing. However, the use

of *amoA* gene marker also showed dominance of putative AOAs, thus it can be cautiously suggested that these archaea contributes to N availability for host and environment through nitrification. Future studies using metagenomics or single cell genomics to target dominant phylotypes have potential to confirm the role of these archaea and also reveal previously unknown functions of the organism and ecosystem.

Table 4.1: List of samples and their source. The S in the sample name represents the samples originated from soil and R represents the samples originated from Rhizosphere

Sample	Tree Source	phylotypes (16S)	phylotypes *amoA*
R10	Populus	17	12
R10.1	Oak	17	13
R10.2	Prunus	19	NA
R10.3	Tulip Poplar	20	NA
R11	Populus	30	12
R11.1	Sycamore	25	17
R11.2	Red maple	22	NA
R11.3	Box elder	26	NA
R2b	Populus	11	7
R2b.1	Hackberry	20	11
R2b.2	Box elder	29	NA
R2b.3	Box elder	20	NA
R5b	Populus	12	4
R5b.1	Beech	NA	NA
R5b.2	Dogwood	21	8
R5b.3	Chest Oak	24	NA
R7	Populus	21	10
R7.1	Hickory	20	11
R7.2	Box elder	18	NA
R7.3	Maple	16	NA
R9a	Populus	21	14
R9a.1	maple	25	13
R9a.2	Box elder	27	NA
R9a.3	Hackberry	22	NA
S10.1	Soil	16	18
S10.2	Soil	12	NA
S10.3	Soil	14	NA
S11.1	Soil	25	16
S11.2	Soil	22	NA
S11.3	Soil	18	NA
S2b.1	Soil	18	11
S2b.2	Soil	16	NA
S2b.3	Soil	17	NA
S5b.1	Soil	20	10
S5b.2	Soil	17	NA
S5b.3	Soil	18	NA
S7.1	Soil	22	12
S7.2	Soil	17	NA
S7.3	Soil	17	NA
S9a.1	Soil	19	11
S9a.2	Soil	16	NA
S9a.3	Soil	24	NA

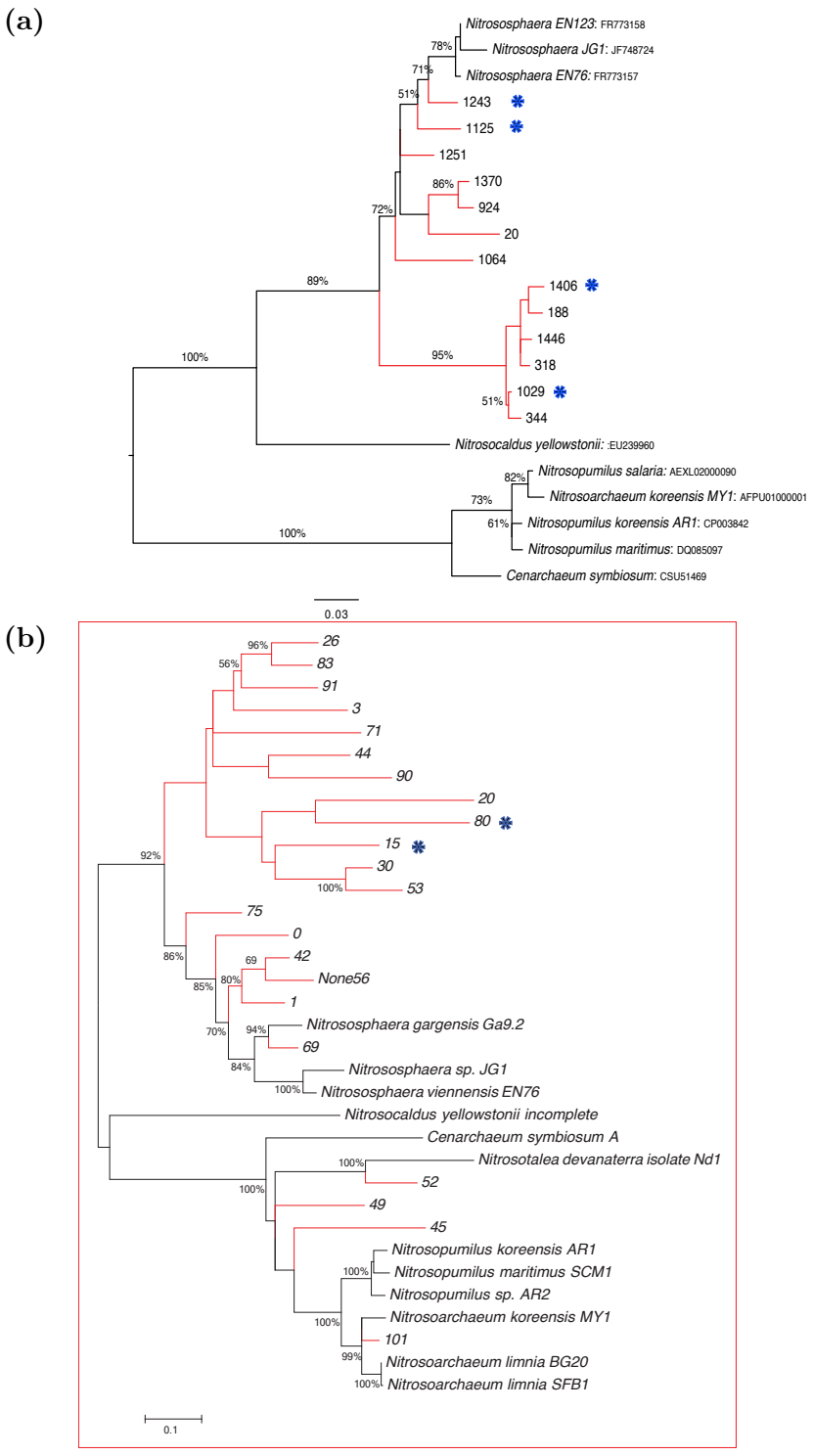


Figure 4.2: Maximum likelihood phylogenetic trees of **(a)** 13 abundant phylotypes based on 16S (V1-V3 $\sim$ 372 base position) and **(b)** phylotypes based on partial *amoA* sequences including corresponding gene sequences from known *Thaumarchaeotas*. The colored branches indicates samples from the current study and the '*' by the phylotype indicates that it is one of the most abundant phylotype in our samples. Numbers indicate bootstrap support based on 100 replicates, and the values with < 50 are not shown. The scale bar indicates the inferred number of substitutions per site.

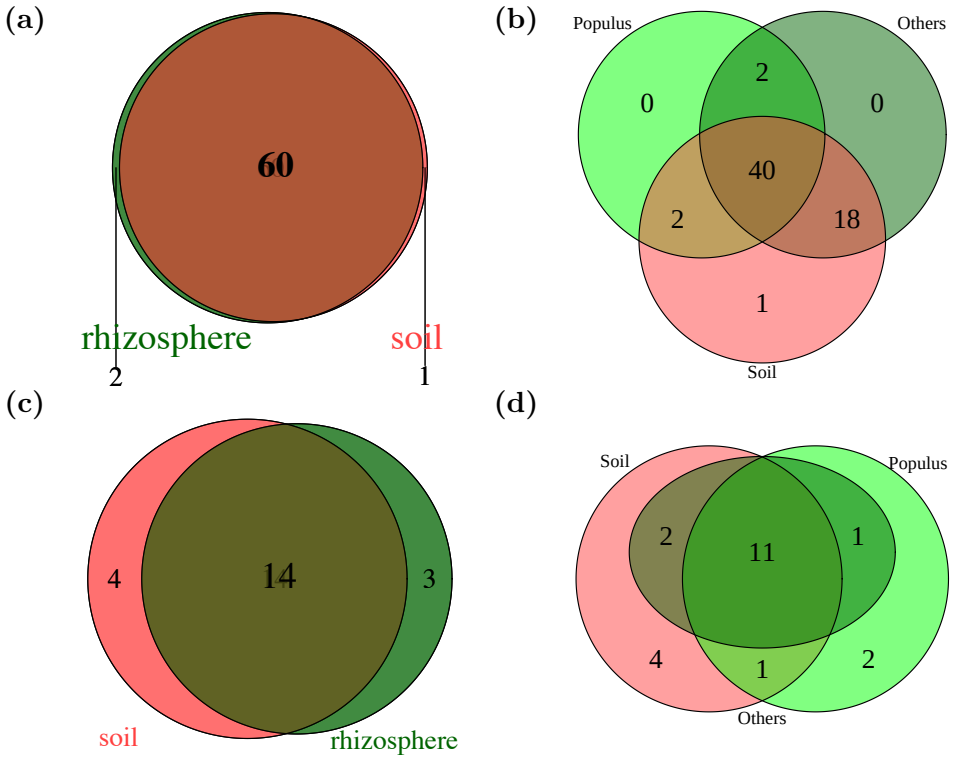


Figure 4.3: Venn diagrams representing phylotypes **(a)** shared between rhizosphere and soil, **(b)** rhizosphere of *P. deltoides* and non *Populus*, and bulk soils, **(c)** *amoA* based phylotypes shared between rhizosphere and soil, and **(d)** *amoA* based phylotypes shared between rhizosphere of *P. deltoides* and non *Populus*, and bulk soils.

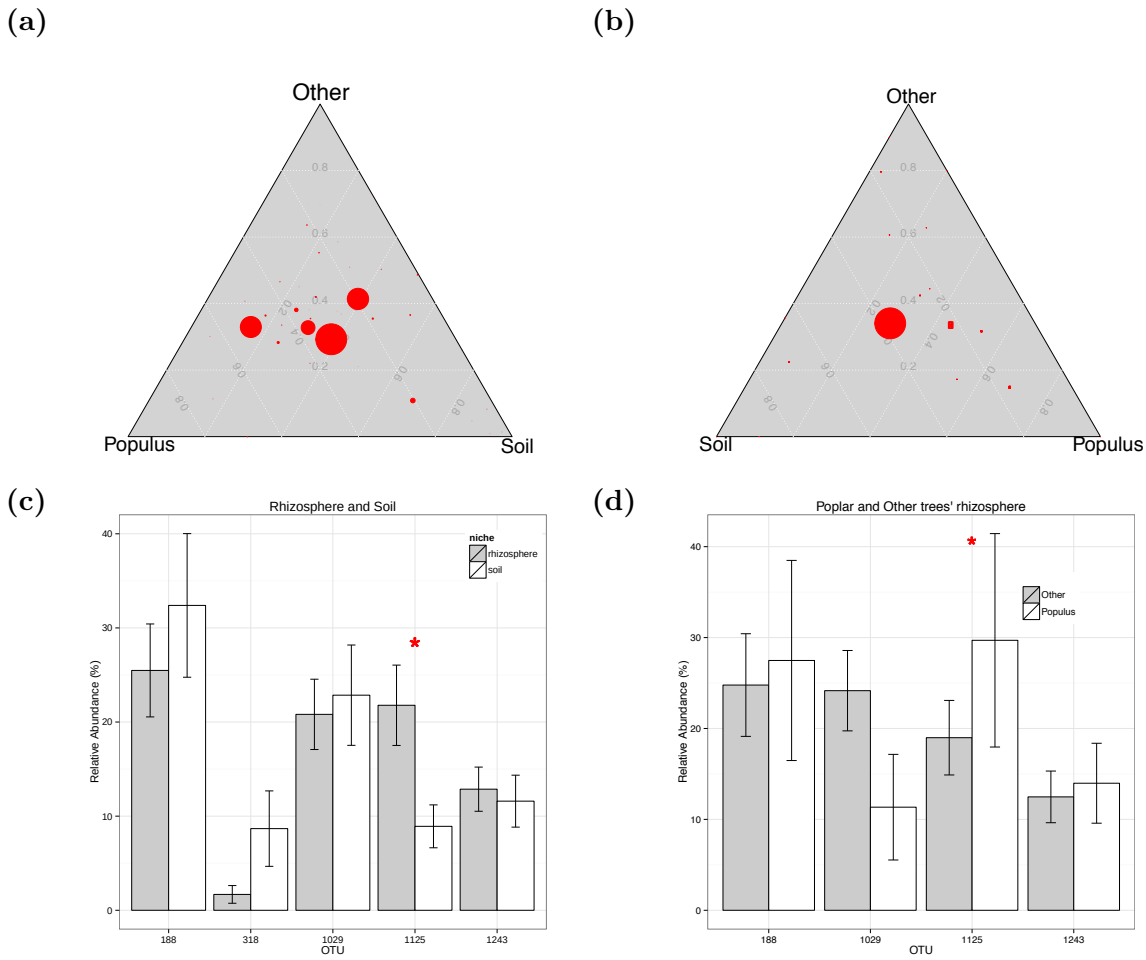


Figure 4.4: Ternary plots based on average abundance of **(a)** 16S based phylotypes and **(b)** *amoA* based phylotypes across all three niches. Size of the dots represent OTU relative abundance in the overall dataset, and position within the triangle the degree of relative enrichment between the three niches. Pairwise comparison of abundant phylotypes between **(c)** rhizosphere and soil, and **(d)** rhizosphere of *Populus* and other surrounding trees. The asterisk represents the statistically significant ($p<0.05$) difference between composition of that phylotype between two niches.

Table 4.2: Soil physical and chemical properties. ** in the sample name represents data that are average of 2010 collection from the same site as these data were not obtained in 2011 due to technical problem. * represents data point that are average of 2010 as well. LBC: Lime Buffer Capacity

Sample	Soil Type	ppm LBC CaCO3/pH	pH CaCl2	pH H_2O	Sand %	Silt %	Clay %	Total C %	Total N %	OM %	Ca ppm	K ppm	Mg ppm	Mn ppm	P ppm	Zn ppm
S2B.1	Clay Loam	443	7.31	7.91	28.0	38.0	34.0	6.66	0.33	6.46	6508	131.6	213.1	1.10	27.6	1
S2B.2	Clay Loam	511	7.42	8.02	23.9	40.2	35.9	7.35	0.36	7.31	6957	147.4	221.2	0.57	33.8	1.10
S2B.3	Clay Loam	Not app.	7.53	8.13	21.9	42.1	36.0	5.5	0.2	3.51	6491	154.7	178.5	0.63	17.5	1.11
S5B.1	Loam	324	6.07	6.67	49.8	36.2	14.0	1.11	0.14	2.04	2025	47.3	136.7	18.99	90.1	6.43
S5B.2	Sandy Loam	374	6.22	6.82	55.8	32.2	12.0	2.21	0.22	3.48	1912	65.7	140.0	27.79	59.1	9.05
S5B.3	Loam	363	5.72	6.32	47.9	38.1	14.0	2	0.22	3.25	1708	50.5	132.5	29.08	43.6	7.33
S7.1	Clay Loam	449	6.35	6.95	21.9	46.0	32.0	2.5	0.26	4.77	3870	83.0	200.7	23.92	430.5	14.59
S7.2	Clay Loam	571	6.42	7.02	21.8	46.2	32.0	4.38	0.4	7.71	4924	101.3	256.9	39.24	397.7	27.10
S7.3	Clay Loam	558	6.34	6.94	21.7	46.2	32.1	3.77	0.36	6.43	4392	95.8	238.6	40.98	358.5	21
S9A.1	Silty Clay Loam	479	6.27	6.87	17.7	52.2	30.0	3.15	0.28	5.14	3159	145.0	220.3	34.22	66.2	6.28
S9A.2	Silt Loam	454	5.76	6.36	17.7	56.2	26.0	2.30	0.25	4.07	2210	121.8	185.7	27.67	69.1	4.97
S9A.3	Silt Loam	552	6.27	6.87	19.7	54.2	26.1	4.35	0.39	7.42	3605	257.7	283.9	35.60	101.7	11.14
S10.1**	Sandy Loam	493.5	6.86	7.46	64	22.53	13.47	5.83	0.42	9.81	4826.5	111.15	389.60	62.38	174	22.67
S10.2	Sandy Loam	819	6.77	7.37	82*	9.9*	8.1*	11.4	0.83	19.5	8125	204.20	549.5	65.33	101.4	14.75
S10.3**	Sandy Loam	493.5	6.86	7.46	64	22.53	13.47	5.83	0.42	9.81	4826.5	111.15	389.60	62.38	174	22.67
S11.1**	Loam	485	7.13	7.73	31.90	45.90	22.20	2.53	0.22	4.88	5528	83.27	191.60	55.94	574.7	8.03
S11.2	Silty Clay Loam	410	6.69	7.29	19.7	48.2	32.0	1.65	0.17	3.07	3060	63.7	376.2	22.61	421.9	26.06
S11.3	Sandy Loam	293	6.30	6.90	59.7	24.2	16.0	0.96	0.13	1.62	2155	49.4	178.7	15.81	342.8	9.95

Table 4.3: List of 454 primers used in the study that targets V1-V3 region of 16S rRNA gene.

Barcode	Primer Name	Full Primer (454+barcode+primer)
AAGCCGCC	FLXB_A2FA1	CTATGCGCCTTGCCAGCCCGCTCAGAAGCCGCCTCYSGTTGATCCYGCSRG
CAAGAACC	FLXB_A2FA2	CTATGCGCCTTGCCAGCCCGCTCAGCAAGAACCTCYSGTTGATCCYGCSRG
AGTTGGCC	FLXB_A2FA3	CTATGCGCCTTGCCAGCCCGCTCAGAGTTGGCCTCYSGTTGATCCYGCSRG
TATCAACC	FLXB_A2FA4	CTATGCGCCTTGCCAGCCCGCTCAGTATCAACCTCYSGTTGATCCYGCSRG
AACCAGCC	FLXB_A2FA5	CTATGCGCCTTGCCAGCCCGCTCAGAACCAGCCTCYSGTTGATCCYGCSRG
CAAGAACC	FLXB_A2FA6	CTATGCGCCTTGCCAGCCCGCTCAGCAAGAACCTCYSGTTGATCCYGCSRG
AGTTGGCC	FLXB_A2FA7	CTATGCGCCTTGCCAGCCCGCTCAGAGTTGGCCTCYSGTTGATCCYGCSRG
TATCAACC	FLXB_A2FA8	CTATGCGCCTTGCCAGCCCGCTCAGTATCAACCTCYSGTTGATCCYGCSRG
AGGCGGCC	FLXB_A2FA9	CTATGCGCCTTGCCAGCCCGCTCAGAGGCGGCCTCYSGTTGATCCYGCSRG
CGGTATCC	FLXB_A2FA10	CTATGCGCCTTGCCAGCCCGCTCAGCGGTATCCTCYSGTTGATCCYGCSRG
TGACGACC	FLXB_A2FA11	CTATGCGCCTTGCCAGCCCGCTCAGTGACGACCTCYSGTTGATCCYGCSRG
ACAAGGCC	FLXB_A2FA12	CTATGCGCCTTGCCAGCCCGCTCAGACAAGGCCTCYSGTTGATCCYGCSRG
AGACCTCC	FLXB_A2FA13	CTATGCGCCTTGCCAGCCCGCTCAGAGACCTCCTCYSGTTGATCCYGCSRG
TAGGAATC	FLXB_A2FA14	CTATGCGCCTTGCCAGCCCGCTCAGTAGGAATCTCYSGTTGATCCYGCSRG
CCGGCCAC	FLXB_A2FA15	CTATGCGCCTTGCCAGCCCGCTCAGCCGGCCACTCYSGTTGATCCYGCSRG
AATGGTAC	FLXB_A2FA16	CTATGCGCCTTGCCAGCCCGCTCAGAATGGTACTCYSGTTGATCCYGCSRG
TCTCCGTC	FLXB_A2FA17	CTATGCGCCTTGCCAGCCCGCTCAGTCTCCGTCTCYSGTTGATCCYGCSRG
TCTCGACC	FLXB_A2FA18	CTATGCGCCTTGCCAGCCCGCTCAGTCTCGACCTCYSGTTGATCCYGCSRG
CCAGGACC	FLXB_A2FA19	CTATGCGCCTTGCCAGCCCGCTCAGCCAGGACCTCYSGTTGATCCYGCSRG
ACTCCTCC	FLXB_A2FA20	CTATGCGCCTTGCCAGCCCGCTCAGACTCCTCCTCYSGTTGATCCYGCSRG
TTCCTGCC	FLXB_A2FA21	CTATGCGCCTTGCCAGCCCGCTCAGTTCCTGCCTCYSGTTGATCCYGCSRG
TTCATACC	FLXB_A2FA22	CTATGCGCCTTGCCAGCCCGCTCAGTTCATACCTCYSGTTGATCCYGCSRG
CGTCGTCC	FLXB_A2FA23	CTATGCGCCTTGCCAGCCCGCTCAGCGTCGTCCTCYSGTTGATCCYGCSRG

Table 4.4: List of 454 primers used in the study to amplify part of archaeal amoA gene.

Barcode	Primer Name	Full Primer (454+barcode+primer)
AAGCCGC	FLXB_Camo19F_1	CTATGCGCCTTGCCAGCCCGCTCAGAAGCCGCATGGTCTGGYTWAGACG
CGCAAC	FLXB_Camo19F_2	CTATGCGCCTTGCCAGCCCGCTCAGCGCAACATGGTCTGGYTWAGACG
TGAAGC	FLXB_Camo19F_3	CTATGCGCCTTGCCAGCCCGCTCAGTGAAGCATGGTCTGGYTWAGACG
ACTTGC	FLXB_Camo19F_4	CTATGCGCCTTGCCAGCCCGCTCAGACTTGCATGGTCTGGYTWAGACG
TCACAC	FLXB_Camo19F_5	CTATGCGCCTTGCCAGCCCGCTCAGTCACACATGGTCTGGYTWAGACG
CGTGAC	FLXB_Camo19F_6	CTATGCGCCTTGCCAGCCCGCTCAGCGTGACATGGTCTGGYTWAGACG
ACGCGC	FLXB_Camo19F_7	CTATGCGCCTTGCCAGCCCGCTCAGACGCGCATGGTCTGGYTWAGACG
CCTCTC	FLXB_Camo19F_8	CTATGCGCCTTGCCAGCCCGCTCAGCCTCTCATGGTCTGGYTWAGACG
ACTCAC	FLXB_Camo19F_9	CTATGCGCCTTGCCAGCCCGCTCAGACTCACATGGTCTGGYTWAGACG
AGACAC	FLXB_Camo19F_10	CTATGCGCCTTGCCAGCCCGCTCAGAGACACATGGTCTGGYTWAGACG
CGACTC	FLXB_Camo19F_11	CTATGCGCCTTGCCAGCCCGCTCAGCGACTCATGGTCTGGYTWAGACG
AGCTTC	FLXB_Camo19F_12	CTATGCGCCTTGCCAGCCCGCTCAGAGCTTCATGGTCTGGYTWAGACG
CAAGAAC	FLXB_Camo19F_13	CTATGCGCCTTGCCAGCCCGCTCAGCAAGAACATGGTCTGGYTWAGACG
AGTTGGC	FLXB_Camo19F_14	CTATGCGCCTTGCCAGCCCGCTCAGAGTTGGCATGGTCTGGYTWAGACG
TATCAAC	FLXB_Camo19F_15	CTATGCGCCTTGCCAGCCCGCTCAGTATCAACATGGTCTGGYTWAGACG
AGGCGGC	FLXB_Camo19F_16	CTATGCGCCTTGCCAGCCCGCTCAGAGGCGGCATGGTCTGGYTWAGACG
CGGTATC	FLXB_Camo19F_17	CTATGCGCCTTGCCAGCCCGCTCAGCGGTATCATGGTCTGGYTWAGACG
TGACGAC	FLXB_Camo19F_18	CTATGCGCCTTGCCAGCCCGCTCAGTGACGACATGGTCTGGYTWAGACG

Chapter 5

Conclusion

5.1 Conclusions

This research was undertaken with the goal of characterizing and understanding microbial (archaea, bacteria, and fungi) communities in roots - both rhizosphere and endosphere - of mature *P. deltoides* trees. NGS is one of the most suitable approaches for exhaustive characterization of microbial community, if inherent biases and errors are taken into consideration during experimental design, data analysis, and interpretation. However, testing for biases is complicated by incomplete knowledge of the diversity, abundance, and genomic details of microbes. Most microbes in environments are still uncultured, unsequenced, and identified only based on marker genes.

To overcome such uncertainty, I constructed mixes of genomic DNA, comprising of known abundances of sequenced bacteria and archaea, or in other words 'synthetic communities'. It allowed direct and quantitative comparisons of two widely used approaches in microbial ecology, shotgun metagenomics and SSU rRNA gene-based diversity characterization. In terms of recapitulating the actual taxonomic diversity of 'synthetic communities', metagenomic sequencing outperformed most SSU rRNA gene primer sets used in this study. None of the primer sets were ideal for quantitatively representing the entire diversity of even our relatively simple community. Among the bacterial primer sets for rRNA gene regions, V13 recovered most accurately the composition of the synthetic community. None of the archaeal sets performed comparably to the bacterial ones. However, the results from the study revealed the use of quality-filtered data can nearly eliminate diversity artifacts from SSU rRNA amplicon data. Both metagenomic strategies - transposon based shearing of DNA followed by sequencing in 454 and physical shearing of DNA followed by sequencing in Illumina HiSeq - recovered the quantitative distribution of the various archaeal and bacterial taxa remarkably well even though organisms spanned two orders of magnitude in abundance. Between two metagenomics method tested, a certain degree of bias was observed for genomes with extreme genomic GC content in transposon-based library sequencing. In summary, 'synthetic communities' like the one used in this study could serve as important analytical controls for validating novel community studies.

The dissertation research represents one of the most in-depth and complete characterization of microbes - archaea, bacteria, and fungi - from roots of *P. deltoides*. The

study revealed overall diversity of archaeal communities (average phylotype at 99% ~19) in their roots are significantly lower compared to corresponding bacterial (average phylotype at 97% ~1500) and fungal (average at 99% ~172) communities. Additionally, the use of functional gene *amoA* as marker to characterize potential ammonia oxidizing archaeaon revealed even lower number of phylotypes (average~10). The archaeal communities of *P. deltoides*, studied using both 16S and *amoA* based phylotype revealed similar community structure dominated by just couple of phylotypes that belonged to phylum *Thaumarchaeota*. Moreover, 99% of all 16S archaeal sequences were classified as *Thaumarchaeota* and phylogenetic analysis of most abundant ones revealed presence of diverse groups, some of which are closely affiliated to clades with known AOAs and some formed a separate clade with no characterized AOAs. Similar results were revealed with phylogenetic analysis of *amoA*. Comparisons of *P. deltoides* archaeal communities based on both 16S and *amoA* with surrounding non *Populus* trees and soil did not reveal selection of unique phylotypes in their roots. However, pairwise comparison of average diversity of 16S based archaea in soil and rhizosphere of all trees (both *Populus* and non *Populus*) from each site revealed a slightly higher diversity of archaea in rhizosphere, but the same was not true for *P. deltoides*. In summary, archaea are common inhabitants of *P. deltoides* root and bulk soils of riparian zones. At extreme conditions of reduced oxygen, flooding, or high CO_2 pressure archaea may become abundant and contribute to rhizosphere processes (Buée et al., 2009; Chen et al., 2008), but as of now much remains to be discovered about their role in roots. It is clear that the archaeal communities in soil and rhizosphere of *P. deltoides* are dominated by putative ammonia oxidizers that could potentially contributing to N availability of host plants and environment.

NGS approaches for exhaustive characterization of bacterial communities have been widely implemented to study associated bacteria from plants like *Arabidopsis spp.*, maize, and others (Bulgarelli et al., 2012; Lundberg et al., 2012; Peiffer et al., 2013), but only few studies included fungal communities from the same plant. Here, I characterized both fungal and bacterial community from same tree in *P. deltoides*. I discovered that roots of *P. deltoides* housed diverse group of bacterial phyla, but were dominated only by nine phyla (*Proteobacteria, Actinobacteria, Acidobacteria, Firmicutes, Planctomycetes, Verrucomicrobia, TM7, Chloroflexi, and Gemmatimonadetes*). Likewise, I detected a total

of eight fungal phyla in the roots of *P. deltoides* with *Ascomycota* dominating the overall fungal communities followed by *Basidiomycota*,*Chytridiomycota* and others of the largely unresolved basal lineages in the former *Zygomycota* that are commonly reported as *Fungi incertae sedis*. Additionally, the primer pair used for bacterial community was able to minimize amplification of mitochondrial and chloroplast sequences of *P. deltoides*, while recapitulating important bacterial groups (namely *Actinobacteria*) that were missed in the previous study (Gottel et al., 2011). Regardless, the study reiterated the difference in communities between rhizosphere and endosphere which was maintained across *P. deltoides* from two geographical settings (Gottel et al., 2011). The differences were evident in higher taxonomic level with seven phyla having different abundance in two niches and at lower taxonomic level (OTUs) with rhizosphere housing 9-10 times more phylotypes than corresponding endosphere. I was also able to partially explain the variation in bacterial and fungal community structures from rhizosphere and endosphere of trees in native environments. The most important factors that were affecting these communities were soil properties (pH), seasonal changes, and geographical settings. In summary, the study revealed new details about the microbes and microbial community structure in the roots of *P. deltoides* that could be utilized in future studies to optimize their use in biofuel production and other environmentally relevant uses.

5.2 Future Directions.

In my study to understand errors and biases in microbial diversity characterization methods, I focused on two sequencing platforms (454 and Illumina) and two common microbial diversity characterization approaches: amplicon based sequencing and metagenomics, but only metagenomic sequencing was performed in both platforms. Illumina is now also used for amplicon sequencing (Caporaso et al., 2011), but its error profile has not been characterized yet. So, the 'synthetic communities' assembled in this study can be used to test the efficacy of amplicon sequencing in Illumina and other new platforms like Ion-Torrent as well. The limited diversity of 'synthetic communities' compared to natural communities is always a disadvantage, but with more microorganisms now available in culture, it would be beneficial to increase the number of organisms in the community. Additionally, DNA extraction biases

were not tested in our study, and based on available cultures of microorganisms, approaches similar to this study can be used to test for biases due to DNA extraction method.

I investigated bacterial and fungal communities in roots of naturally occurring riparian population of *P. deltoides* as these two taxa consist of members that are major players in their growth and health. In this study, I was able to account partial (14%-24%) variation in bacterial and fungal communities of rhizosphere and endosphere to soil chemistry, seasonal change, and geographical settings. However, the attributed community variation can be increased by continuing the study in a more controlled environment like common gardens. Additionally, my study failed to attribute community variation to genotype differences, but the role of genotype can't be discounted yet. Different set of SSR markers that are able to tease out differences in putative clonal population or the ones that corroborate with root exudation properties could be used to test the hypothesis again. The study also deciphered OTUs that could reveal mechanism of colonization. For instance, in both fungal and bacterial communities, there were OTUs residing exclusively in rhizosphere or endosphere and both. Isolation, sequencing, and comparative genomic studies of the organisms represented by these OTUs could reveal genetic basis of colonization and proliferation.

The archaeal study was able to capture most of the archaeal diversity for the primer set used and revealed that regardless of habitat, rhizosphere or soil, all samples were dominated by 6-13 phylotypes. Most of these phylotypes do not have culture representatives or genome sequences, so much remains to be learned. Given their potential role in ammonia oxidation, the resident *Thaumarchaeotas* could be important players that contribute to N availability in these environments. Thus, next step could be obtaining genomic information for these phylotypes using single cell genomics. To specifically target potential AOAs, we have developed antibodies against a cultured representative *N. viennensis* that could be used to isolate cells for single cell genomics and culture studies. Additionally, the contribution of AOAs in Nitrogen cycle compared to ammonia oxidizing bacteria (AOB) in these environments are also not known. To understand the importance of archaea in these environments a comparative quantification of resident AOAs and AOBs using qPCR could reveal the dominant ammonia oxidizers from these environment.

Bibliography

A, M. (2001). *Studies on the importance of endophytic bacteria for the biological control of the root-knot nematode Meloidogyne incognita on tomato*. PhD thesis, University of Bonn, Germany. 8

Adey, A., Morrison, H. G., Asan, Xun, X., Kitzman, J. O., Turner, E. H., Stackhouse, B., MacKenzie, A. P., Caruccio, N. C., Zhang, X., and Shendure, J. (2010). Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology*, 11(12):R119. 30

Aira, M., Gómez-Brandón, M., Lazcano, C., Bååth, E., and Domínguez, J. (2010). Plant genotype strongly modifies the structure and growth of maize rhizosphere microbial communities. *Soil Biology and Biochemistry*, 42(12):2276–2281. 7, 9, 66, 82

Alves, R. J. E., Wanek, W., Zappe, A., Richter, A., Svenning, M. M., Schleper, C., and Urich, T. (2013). Nitrification rates in arctic soils are associated with functionally distinct populations of ammonia-oxidizing archaea. *The ISME journal*. 110

Auguet, J.-C., Barberan, A., and Casamayor, E. O. (2009). Global ecological patterns in uncultured archaea. *The ISME journal*, 4(2):182–190. 106, 107

Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S., and Vivanco, J. M. (2006). The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu. Rev. Plant Biol.*, 57:233–266. 4

Baker, G., Smith, J., and Cowan, D. (2003). Review and re-analysis of domain-specific 16s primers. *Journal of Microbiological Methods*, 55(3):541–555. 24, 61

Bates, S. T., Berg-Lyons, D., Caporaso, J. G., Walters, W. A., Knight, R., and Fierer, N. (2011). Examining the global distribution of dominant archaeal populations in soil. *ISME J*, 5(5):908–17. 24, 37, 61, 100, 106, 107

Behie, S., Zelisko, P., and Bidochka, M. (2012). Endophytic insect-parasitic fungi translocate nitrogen directly from insects to plants. *Science*, 336(6088):1576–1577. 84

Benhamou, N., Gagné, S., Le Quéré, D., and Dehbi, L. (2000). Bacterial-mediated induced resistance in cucumber: beneficial effect of the endophytic bacterium serratia plymuthica on the protection against infection by pythium ultimum. *Phytopathology*, 90(1):45–56. 3

Berendsen, R. L., Pieterse, C. M., and Bakker, P. A. (2012). The rhizosphere microbiome and plant health. *Trends in plant science*, 17(8):478–486. 65

Berg, G., Zachow, C., Lottmann, J., Götz, M., Costa, R., and Smalla, K. (2005). Impact of plant species and site on rhizosphere-associated fungi antagonistic to verticillium dahliae kleb. *Applied and environmental microbiology*, 71(8):4203–4213. 7, 8

Bever, J. D. (2003). Soil community feedback and the coexistence of competitors: conceptual frameworks and empirical tests. *New Phytologist*, 157(3):465–473. 66

Boer, W., Folman, L., Summerbell, R., and Boddy, L. (2006). Living in a fungal world: impact of fungi on soil bacterial niche development. *FEMS microbiology reviews*, 29(4):795–811. 82

Bomberg, M., Jurgens, G., Saano, A., Sen, R., and Timonen, S. (2003). Nested pcr detection of archaea in defined compartments of pine mycorrhizospheres developed in boreal forest humus microcosms. *FEMS microbiology ecology*, 43(2):163–171. 4, 99

Bomberg, M., Münster, U., Pumpanen, J., Ilvesniemi, H., and Heinonsalo, J. (2011). Archaeal communities in boreal forest tree rhizospheres respond to changing soil temperatures. *Microbial ecology*, 62(1):205–217. 4, 6, 100

Bomberg, M. and Timonen, S. (2009). Effect of tree species and mycorrhizal colonization on the archaeal population of boreal forest rhizospheres. *Applied and environmental microbiology*, 75(2):308–315. 100

Bonfante, P. and Anca, I.-A. (2009). Plants, mycorrhizal fungi, and bacteria: a network of interactions. *Annual review of microbiology*, 63:363–383. 82

Borcard, D. and Legendre, P. (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, 153(1):51–68. 75

Bradshaw, H., Ceulemans, R., Davis, J., and Stettler, R. (2000). Emerging model systems in plant biology: poplar (populus) as a model forest tree. *Journal of Plant Growth Regulation*, 19(3):306–313. 10

Bragg, L., Stone, G., Imelfort, M., Hugenholtz, P., and Tyson, G. W. (2012). Fast, accurate error-correction of amplicon pyrosequences using acacia. *Nature Methods*, 9(5):425–426. 14

Broeckling, C. D., Broz, A. K., Bergelson, J., Manter, D. K., and Vivanco, J. M. (2008). Root exudates regulate soil fungal community composition and diversity. *Applied and Environmental Microbiology*, 74(3):738–744. 83

Buée, M., De Boer, W., Martin, F., Van Overbeek, L., and Jurkevitch, E. (2009). The rhizosphere zoo: an overview of plant-associated communities of microorganisms, including phages, bacteria, archaea, and fungi, and of some of their structuring factors. *Plant and Soil*, 321(1-2):189–212. 3, 4, 8, 15, 83, 121

Bulgarelli, D., Rott, M., Schlaeppi, K., van Themaat, E. V. L., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., et al. (2012). Revealing structure and assembly cues for arabidopsis root-inhabiting bacterial microbiota. *Nature*, 488(7409):91–95. 6, 8, 66, 121

Campbell, J., Foster, C., Vishnivetskaya, T., Campbell, A., Yang, Z., Wymore, A., Palumbo, A., Chesler, E., and Podar, M. (2012). Host genetic and environmental effects on mouse intestinal microbiota. *The ISME Journal*. 52

Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., Fierer, N., Pena, A., Goodrich, J., Gordon, J., et al. (2010a). Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336. 25, 69, 103, 105

Caporaso, J. G., Bittinger, K., Bushman, F. D., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2010b). Pynast: a flexible tool for aligning sequences to a template alignment. *Bioinformatics*, 26(2):266–267. 69, 105

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N., and Knight, R. (2011). Global patterns of 16s rrna diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4516–4522. 11, 14, 18, 51, 82, 122

Cardon, Z. G. (2007). *The rhizosphere: an ecological perspective*. Academic press. 4, 5, 7

Carroll, G. (1986). The biology of endophytism in plants with particular reference to woody perennials. In Fokkema, N. J. and Heuvel, J. v. d., editors, *Microbiology of the Phyllosphere*, volume 1, page 2t. Cambridge University Press, Cambridge, UK. 5

Caruccio, N. (2011). Preparation of next-generation sequencing librapreparation of next-generation sequencing libraries using nextera^{TM} technology: Simultaneous dna fragmentation and adaptor tagging by in vitro transposition. *Methods in Molecular Biology*, 733:241–255. 30

Chelius, M., Triplett, E., et al. (2001). The diversity of archaea and bacteria in association with the roots of zea mays l. *Microbial Ecology*, 41(3):252–263. 99

Chen, X.-P., Zhu, Y.-G., Xia, Y., Shen, J.-P., and He, J.-Z. (2008). Ammonia-oxidizing archaea: important players in paddy rhizosphere soil? *Environmental Microbiology*, 10(8):1978–1987. 100, 121

Clarke, K. and Warwick, R. (1994). *Change in marine communities: an approach to statistical analysis and interpretation*. Plymouth marine laboratory, Natural environment research council. 22

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., and Tiedje, J. M. (2009). The ribosomal database project: improved alignments and new tools for rrna analysis. *Nucleic Acids Res*, 37(Database issue):D141–5. 25, 36, 104

Conrad, R. (2007). Microbial ecology of methanogens and methanotrophs. *Advances in agronomy*, 96:1–63. 100

Costa, R., Götz, M., Mrotzek, N., Lottmann, J., Berg, G., and Smalla, K. (2006). Effects of site and plant species on rhizosphere community structure as revealed by molecular analysis of microbial guilds. *FEMS Microbiology Ecology*, 56(2):236–249. 7, 8, 9

Danielsen, L., Thürmer, A., Meinicke, P., Buée, M., Morin, E., Martin, F., Pilate, G., Daniel, R., Polle, A., and Reich, M. (2012). Fungal soil communities in a young transgenic poplar plantation form a rich reservoir for fungal root communities. *Ecology and evolution*, 2(8):1935–1948. 65

DeLong, E. F. and Pace, N. R. (2001). Environmental diversity of bacteria and archaea. *Syst Biol*, 50(4):470–8. 10, 17

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and environmental microbiology*, 72(7):5069–5072. 69, 104, 105

Di, H., Cameron, K., Shen, J. P., Winefield, C., O'Callaghan, M., Bowatte, S., and He, J. (2009). Nitrification driven by bacteria and not archaea in nitrogen-rich grassland soils. *Nature Geoscience*, 2(9):621–624. 100

Doornbos, R. F., Geraats, B. P., Kuramae, E. E., Van Loon, L., and Bakker, P. A. (2011). Effects of jasmonic acid, ethylene, and salicylic acid signaling on the rhizosphere bacterial community of arabidopsis thaliana. *Molecular Plant-Microbe Interactions*, 24(4):395–407. 3

Doty, S. L., Oakley, B., Xin, G., Kang, J. W., Singleton, G., Khan, Z., Vajzovic, A., and Staley, J. T. (2009). Diazotrophic endophytes of native black cottonwood and willow. *Symbiosis*, 47(1):23–33. 65, 66

Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., et al. (2011). Geneious v5. 4. 105

Eckburg, P. B. and Relman, D. A. (2007). The role of microbes in crohn's disease. *Clinical Infectious Diseases*, 44(2):256–262. 7

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than blast. *Bioinformatics*, 26(19):2460–2461. 69, 104, 105

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., and Knight, R. (2011). Uchime improves sensitivity and speed of chimera detection. *Bioinformatics*, 27(16):2194–2200. 69, 103, 104

Engelbrektson, A., Kunin, V., Wrighton, K. C., Zvenigorodsky, N., Chen, F., Ochman, H., and Hugenholtz, P. (2010). Experimental factors affecting pcr-based estimates of microbial species richness and evenness. *ISME J*, 4(5):642–7. 10, 18, 37

Fierer, N., Jackson, J., Vilgalys, R., and Jackson, R. (2005). Assessment of soil microbial community structure by use of taxon-specific quantitative pcr assays. *Applied and Environmental Microbiology*, 71(7):4117–4120. 19

Frank, J., Reich, C., Sharma, S., Weisbaum, J., Wilson, B., and Olsen, G. (2008). Critical evaluation of two primers commonly used for amplification of bacterial 16s rrna genes. *Applied and Environmental Microbiology*, 74(8):2461–2470. 24, 37, 61

Gihring, T., Green, S., and Schadt, C. (2011). Massively parallel rrna gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environmental microbiology*, 14(2):285–290. 11, 18

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc*, 2010(1):pdb prot5368. 32

Gomez-Alvarez, V., Teal, T. K., and Schmidt, T. M. (2009). Systematic artifacts in metagenomes from complex microbial communities. *ISME J*, 3(11):1314–7. Gomez-Alvarez, Vicente Teal, Tracy K Schmidt, Thomas M England ISME J. 2009 Nov;3(11):1314-7. doi: 10.1038/ismej.2009.72. Epub 2009 Jul 9. 18, 52

Goslee, S. C. and Urban, D. L. (2007). The ecodist package for dissimilarity-based analysis of ecological data. *Journal of Statistical Software*, 22(7):1–19. 70, 72, 78

Gottel, N. R., Castro, H. F., Kerley, M., Yang, Z., Pelletier, D. A., Podar, M., Karpinets, T., Uberbacher, E., Tuskan, G. A., Vilgalys, R., et al. (2011). Distinct microbial communities within the endosphere and rhizosphere of populus deltoides roots across contrasting soil types. *Applied and environmental microbiology*, 77(17):5934–5944. 5, 6, 8, 66, 69, 72, 80, 100, 122

Graff, A. and Conrad, R. (2005). Impact of flooding on soil bacterial communities associated with poplar (populus sp.) trees. *FEMS microbiology ecology*, 53(3):401–415. 65, 66

Großkopf, R., Stubner, S., and Liesack, W. (1998). Novel euryarchaeotal lineages detected on rice roots and in the anoxic bulk soil of flooded rice microcosms. *Applied and environmental microbiology*, 64(12):4983–4989. 99

Gubry-Rangin, C., Hai, B., Quince, C., Engel, M., Thomson, B. C., James, P., Schloter, M., Griffiths, R. I., Prosser, J. I., and Nicol, G. W. (2011). Niche specialization of terrestrial archaeal ammonia oxidizers. *Proceedings of the National Academy of Sciences*, 108(52):21206–21211. 7, 100, 104, 107, 110

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methe, B., DeSantis, T. Z., Petrosino, J. F., Knight, R., and Birren, B. W. (2011). Chimeric 16s rrna sequence formation and detection in sanger and 454-pyrosequenced pcr amplicons. *Genome Res*, 21(3):494–504. 10, 11, 18, 37, 72, 104

Hannula, S. E., de Boer, W., and van Veen, J. (2012). A 3-year study reveals that plant growth stage, season and field site affect soil fungal communities while cultivar and gm-trait have minor effects. *PloS One*, 7(4):e33819. 8, 9, 66, 82

Hatzenpichler, R., Lebedeva, E. V., Spieck, E., Stoecker, K., Richter, A., Daims, H., and Wagner, M. (2008). A moderately thermophilic ammonia-oxidizing crenarchaeote from a hot spring. *Proceedings of the National Academy of Sciences*, 105(6):2134–2139. 107

Hernesmaa, A., Björklöf, K., Kiikkilä, O., Fritze, H., Haahtela, K., and Romantschuk, M. (2005). Structure and function of microbial communities in the rhizosphere of scots pine after tree-felling. *Soil Biology and Biochemistry*, 37(4):777–785. 8

Herrmann, M., Saunders, A. M., and Schramm, A. (2008). Archaea dominate the ammonia-oxidizing community in the rhizosphere of the freshwater macrophyte littorella uniflora. *Applied and environmental microbiology*, 74(10):3279–3283. 100

Hess, M., Sczyrba, A., Egan, R., Kim, T. W., Chokhawala, H., Schroth, G., Luo, S., Clark, D. S., Chen, F., Zhang, T., Mackie, R. I., Pennacchio, L. A., Tringe, S. G., Visel, A., Woyke, T., Wang, Z., and Rubin, E. M. (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science*, 331(6016):463–7. 30, 51

Hiltner, L. (1904). Über neuere erfahrungen und probleme auf dem gebiete der bodenbakteriologie unter besonderer berücksichtigung der gründüngung und brache. *Arbeiten der Deutschen Landwirtschaftlichen Gesellschaft*, 98:59–78. 3, 4, 100

Holland, M. A. (1997). Occam's razor applied to hormonology (are cytokinins produced by plants?). *Plant Physiology*, 115(3):865. 3

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S. S. (2009). Polymerase chain reaction primers miss half of rrna microbial diversity. *ISME J*, 3(12):1365–73. 10, 18

Hur, M., Kim, Y., Song, H.-R., Kim, J. M., Im Choi, Y., and Yi, H. (2011). Effect of genetically modified poplars on soil microbial communities during the phytoremediation of waste mine tailings. *Applied and environmental microbiology*, 77(21):7611–7619. 10

Huse, S. M., Welch, D. M., Morrison, H. G., and Sogin, M. L. (2010). Ironing out the wrinkles in the rare biosphere through improved otu clustering. *Environ Microbiol*, 12(7):1889–98. 10, 14, 18, 27, 38, 39, 52

Huson, D., Auch, A., Qi, J., and Schuster, S. (2007). Megan analysis of metagenomic data. *Genome research*, 17(3):377–386. 23, 35

Iverson, V., Morris, R., Frazar, C., Berthiaume, C., Morales, R., and Armbrust, E. (2012). Untangling genomes from metagenomes: revealing an uncultured class of marine euryarchaeota. *Science*, 335(6068):587–590. 53

Izumi, H., Anderson, I. C., Killham, K., and Moore, E. R. (2008). Diversity of predominant endophytic bacteria in european deciduous and coniferous trees. *Canadian journal of microbiology*, 54(3):173–179. 5

Jallow, M. F., Dugassa-Gobena, D., and Vidal, S. (2004). Indirect interaction between an unspecialized endophytic fungus and a polyphagous moth. *Basic and Applied Ecology*, 5(2):183–191. 3

Jansson, S. and Douglas, C. J. (2007). Populus: a model system for plant biology. *Annu. Rev. Plant Biol.*, 58:435–458. 10

Jumpponen, A., JONES, K. L., David Mattox, J., and YAEGE, C. (2010). Massively parallel 454-sequencing of fungal communities in quercus spp. ectomycorrhizas indicates seasonal dynamics in urban and rural sites. *Molecular Ecology*, 19(s1):41–53. 6

Kai, M., Vespermann, A., and Piechulla, B. (2008). The growth of fungi and arabidopsis thaliana is influenced by bacterial volatiles. *Plant signaling & behavior*, 3(7):482–484. 82

Kan, J., Clingenpeel, S., Macur, R. E., Inskeep, W. P., Lovalvo, D., Varley, J., Gorby, Y., McDermott, T. R., and Nealson, K. (2011). Archaea in yellowstone lake. *ISME J*, 5(11):1784–95. 37

Khan, A. L., Hamayun, M., Khan, S. A., Kang, S.-M., Shinwari, Z. K., Kamran, M., ur Rehman, S., Kim, J.-G., and Lee, I.-J. (2012a). Pure culture of metarhizium anisopliae lhl07 reprograms soybean to higher growth and mitigates salt stress. *World Journal of Microbiology and Biotechnology*, 28(4):1483–1494. 84

Khan, A. L., Hamayun, M., Waqas, M., Kang, S.-M., Kim, Y.-H., Kim, D.-H., and Lee, I.-J. (2012b). Exophiala sp. lhl08 association gives heat stress tolerance by avoiding oxidative damage to cucumber plants. *Biology and Fertility of Soils*, 48(5):519–529. 84

Kim, B. K., Jung, M.-Y., Yu, D. S., Park, S.-J., Oh, T. K., Rhee, S.-K., and Kim, J. F. (2011). Genome sequence of an ammonia-oxidizing soil archaeon,"candidatus nitrosoarchaeum koreensis" my1. *Journal of bacteriology*, 193(19):5539–5540. 4

Kim, J.-G., Jung, M.-Y., Park, S.-J., Rijpstra, W. I. C., Sinninghe Damsté, J. S., Madsen, E. L., Min, D., Kim, J.-S., Kim, G.-J., and Rhee, S.-K. (2012). Cultivation of a highly enriched ammonia-oxidizing archaeon of thaumarchaeotal group i. 1b from an agricultural soil. *Environmental Microbiology*, 14(6):1528–1543. 107

Koren, S., Treangen, T. J., and Pop, M. (2011). Bambus 2: scaffolding metagenomes. *Bioinformatics*, 27(21):2964–71. 51

Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2009). Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology*, 12(1):118–123. 11, 13, 14, 38, 52

Kuske, C. R., Ticknor, L. O., Miller, M. E., Dunbar, J. M., Davis, J. A., Barns, S. M., and Belnap, J. (2002). Comparison of soil bacterial communities in rhizospheres of three plant species and the interspaces in an arid grassland. *Applied and Environmental Microbiology*, 68(4):1854–1863. 7, 9

Lane, D. (1991). 16s/23s rrna sequencing. *Nucleic acid techniques in bacterial systematics*. 24

Lauber, C. L., Hamady, M., Knight, R., and Fierer, N. (2009). Pyrosequencing-based assessment of soil ph as a predictor of soil bacterial community structure at the continental scale. *Applied and environmental microbiology*, 75(15):5111–5120. 8, 82

LeBlanc, P. M., Hamelin, R. C., and Filion, M. (2007). Alteration of soil rhizosphere communities following genetic transformation of white spruce. *Applied and environmental microbiology*, 73(13):4128–4134. 7, 9

Legendre, P. and Anderson, M. J. (1999). Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69(1):1–24. 75

Leininger, S., Urich, T., Schloter, M., Schwark, L., Qi, J., Nicol, G., Prosser, J., Schuster, S., and Schleper, C. (2006). Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, 442(7104):806–809. 107

Lesaulnier, C., Papamichail, D., McCorkle, S., Ollivier, B., Skiena, S., Taghavi, S., Zak, D., and Van Der Lelie, D. (2008). Elevated atmospheric co2 affects soil microbial diversity associated with trembling aspen. *Environmental Microbiology*, 10(4):926–941. 110

Ley, R. E., Hamady, M., Lozupone, C., Turnbaugh, P. J., Ramey, R. R., Bircher, J. S., Schlegel, M. L., Tucker, T. A., Schrenzel, M. D., Knight, R., and Gordon, J. I. (2008). Evolution of mammals and their gut microbes. *Science*, 320(5883):1647–51. 19

Lindow, S. E. and Leveau, J. H. (2002). Phyllosphere microbiology. *Current Opinion in Biotechnology*, 13(3):238–243. 3

Liu, B., Gibbons, T., Ghodsi, M., Treangen, T., and Pop, M. (2011). Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC genomics*, 12(Suppl 2):S4. 51

Liu, K.-L., Porras-Alfaro, A., Kuske, C. R., Eichorst, S. A., and Xie, G. (2012). Accurate, rapid taxonomic classification of fungal large-subunit rrna genes. *Applied and environmental microbiology*, 78(5):1523–1533. 69, 70

Long, H. H., Sonntag, D. G., Schmidt, D. D., and Baldwin, I. T. (2010). The structure of the culturable root bacterial endophyte community of nicotiana attenuata is organized by soil composition and host plant ethylene production and perception. *New Phytologist*, 185(2):554–567. 8

Lottmann, J., O'Callaghan, M., Baird, D., and Walter, C. (2010). Bacterial and fungal communities in the rhizosphere of field-grown genetically modified pine trees (pinus radiata d.). *Environmental biosafety research*, 9(1):25. 6, 8, 66, 82

Lozupone, C. and Knight, R. (2005). Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and environmental microbiology*, 71(12):8228–8235. 75

Lu, X. and Koide, R. T. (1994). The effects of mycorrhizal infection on components of plant growth and reproduction. *New Phytologist*, 128(2):211–218. 66

Lu, Y. and Conrad, R. (2005). In situ stable isotope probing of methanogenic archaea in the rice rhizosphere. *Science*, 309(5737):1088–1090. 4

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., et al. (2012). Defining the core arabidopsis thaliana root microbiome. *Nature*, 488(7409):86–90. 4, 8, 9, 66, 80, 81, 121

MacLean, D., Jones, J. D., and Studholme, D. J. (2009). Application of'next-generation'sequencing technologies to microbial genetics. *Nature Reviews Microbiology*, 7(4):287–296. 2

Mano, H., Tanaka, F., Nakamura, C., Kaga, H., and Morisaki, H. (2007). Culturable endophytic bacterial flora of the maturing leaves and roots of rice plants (oryza sativa) cultivated in a paddy field. *Microbes and Environments*, 22(2):175–185. 8

Markowitz, V. M., Chen, I. M., Chu, K., Szeto, E., Palaniappan, K., Grechkin, Y., Ratner, A., Jacob, B., Pati, A., Huntemann, M., Liolios, K., Pagani, I., Anderson, I., Mavromatis, K., Ivanova, N. N., and Kyrpides, N. C. (2012). Img/m: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res*, 40(Database issue):D123–9. 32

Marschner, P., Crowley, D., and Yang, C. H. (2004). Development of specific rhizosphere bacterial communities in relation to plant species, nutrition and soil type. *Plant and Soil*, 261(1-2):199–208. 7, 8, 9, 82

Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., et al. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nature methods*, 4(6):495–500. 36

McCarren, J., Becker, J. W., Repeta, D. J., Shi, Y., Young, C. R., Malmstrom, R. R., Chisholm, S. W., and DeLong, E. F. (2010). Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *Proc Natl Acad Sci U S A*, 107(38):16420–7. 18, 30

McCutcheon, J. and Moran, N. (2007). Parallel genomic evolution and metabolic interdependence in an ancient symbiosis. *Proceedings of the National Academy of Sciences*, 104(49):19392–19397. 24

Mendes, R., Garbeva, P., and Raaijmakers, J. M. (2013). The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS microbiology reviews*. 65

Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J. H., Piceno, Y. M., DeSantis, T. Z., Andersen, G. L., Bakker, P. A., et al. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, 332(6033):1097–1100. 2, 7

Moore, F. P., Barac, T., Borremans, B., Oeyen, L., Vangronsveld, J., Van der Lelie, D., Campbell, C. D., and Moore, E. R. (2006). Endophytic bacterial diversity in poplar trees growing on a btex-contaminated site: the characterisation of isolates with potential to enhance phytoremediation. *Systematic and applied microbiology*, 29(7):539–556. 66

Morales, S. E. and Holben, W. E. (2009). Empirical testing of 16s rrna gene pcr primer pairs reveals variance in target specificity and efficacy not suggested by in silico analysis. *Appl Environ Microbiol*, 75(9):2677–83. 11, 18, 40

Mougel, C., Offre, P., Ranjard, L., Corberand, T., Gamalero, E., Robin, C., and Lemanceau, P. (2006). Dynamic of the genetic structure of bacterial and fungal communities at different developmental stages of medicago truncatula gaertn. cv. jemalong line j5. *New Phytologist*, 170(1):165–175. 9

Muyzer, G., Hottenträger, S., Teske, A., and Wawer, C. (1996). Denaturing gradient gel electrophoresis of pcr-amplified 16s rdna—a new molecular approach to analyse the genetic diversity of mixed microbial communities. *Molecular microbial ecology manual*, 3(4):1–23. 24, 61

Nelson, D. M., Cann, I. K., and Mackie, R. I. (2010). Response of archaeal communities in the rhizosphere of maize and soybean to elevated atmospheric co2 concentrations. *PloS one*, 5(12):e15897. 107

Nübel, U., Engelen, B., Felske, A., Snaidr, J., Wieshuber, A., Amann, R., Ludwig, W., and Backhaus, H. (1996). Sequence heterogeneities of genes encoding 16s rrnas in paenibacillus polymyxa detected by temperature gradient gel electrophoresis. *Journal of Bacteriology*, 178(19):5636–5643. 24

Oksanen, J., Kindt, R., Legendre, P., OHara, B., Stevens, M. H. H., Oksanen, M. J., and Suggests, M. (2007). The vegan package. *Community ecology package. Disponível em: http://www. R-project. org. Acesso em*, 10(01):2008. 70

Onodera, Y., Nakagawa, T., Takahashi, R., and Tokuyama, T. (2009). Seasonal change in vertical distribution of ammonia-oxidizing archaea and bacteria and their nitrification in temperate forest soil. *Microbes and environments*, (0):1001260154. 100

Ovreås, L., Forney, L., Daae, F., and Torsvik, V. (1997). Distribution of bacterioplankton in meromictic lake saelenvannet, as determined by denaturing gradient gel electrophoresis of pcr-amplified gene fragments coding for 16s rrna. *Applied and Environmental Microbiology*, 63(9):3367–3373. 24, 61

Pace, N. R. (2009). Mapping the tree of life: progress and prospects. *Microbiol Mol Biol Rev*, 73(4):565–76. 10, 17

Patil, K., Haider, P., Pope, P., Turnbaugh, P., Morrison, M., Scheffer, T., and McHardy, A. (2011). Taxonomic metagenome sequence assignment with structured output models. *Nature methods*, 8(3):191–192. 51

Peay, K. G., Bruns, T. D., Kennedy, P. G., Bergemann, S. E., and Garbelotto, M. (2007). A strong species–area relationship for eukaryotic soil microbes: island size matters for ectomycorrhizal fungi. *Ecology Letters*, 10(6):470–480. 78, 82

Peiffer, J. A., Spor, A., Koren, O., Jin, Z., Tringe, S. G., Dangl, J. L., Buckler, E. S., and Ley, R. E. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proceedings of the National Academy of Sciences*, 110(16):6548–6553. 4, 8, 9, 121

Pérez-Pantoja, D., Donoso, R., Agulló, L., Córdova, M., Seeger, M., Pieper, D. H., and González, B. (2012). Genomic analysis of the potential for aromatic compounds biodegradation in burkholderiales. *Environmental microbiology*, 14(5):1091–1117. 84

Pester, M., Rattei, T., Flechl, S., Gröngröft, A., Richter, A., Overmann, J., Reinhold-Hurek, B., Loy, A., and Wagner, M. (2012). amoa-based consensus phylogeny of ammonia-oxidizing archaea and deep sequencing of amoa genes from soils of four different geographic regions. *Environmental microbiology*, 14(2):525–539. 100, 103, 104, 107, 108, 110

Pignatelli, M. and Moya, A. (2011). Evaluating the fidelity of de novo short read metagenomic assembly using simulated data. *PloS one*, 6(5):e19984. 35, 36

Podila, G. K., Sreedasyam, A., and Muratet, M. A. (2009). Populus rhizosphere and the ectomycorrhizal interactome. *Critical Reviews in Plant Science*, 28(5):359–367. 10

Polz, M. F. and Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate pcr. *Appl Environ Microbiol*, 64(10):3724–30. 10, 11, 13, 18

Porat, I., Vishnivetskaya, T., Mosher, J., Brandt, C., Yang, Z., Brooks, S., Liang, L., Drake, M., Podar, M., Brown, S., et al. (2010). Characterization of archaeal community in contaminated and uncontaminated surface stream sediments. *Microbial ecology*, 60(4):784–795. 37

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, 5(3):e9490. 69

Prosser, J. (2010). Replicate or lie. *Environmental microbiology*, 12(7):1806–1810. 51

Quince, C., Lanzen, A., Curtis, T., Davenport, R., Hall, N., Head, I., Read, L., and Sloan, W. (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature methods*, 6(9):639–641. 52

Quince, C., Lanzen, A., Davenport, R. J., and Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC Bioinformatics*, 12:38. 11, 14, 18, 25, 26, 28, 37, 52, 69, 72, 103, 104

Reeder, J. and Knight, R. (2010). Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. *Nature methods*, 7(9):668. 14, 52

Reysenbach, A., Liu, Y., Banta, A., Beveridge, T., Kirshtein, J., Schouten, S., Tivey, M., Von Damm, K., and Voytek, M. (2006). A ubiquitous thermoacidophilic archaeon from deep-sea hydrothermal vents. *Nature*, 442(7101):444–447. 19

Rodriguez, R., White Jr, J., Arnold, A., and Redman, R. (2009). Fungal endophytes: diversity and functional roles. *New Phytologist*, 182(2):314–330. 65

Roesti, D., Ineichen, K., Braissant, O., Redecker, D., Wiemken, A., and Aragno, M. (2005). Bacteria associated with spores of the arbuscular mycorrhizal fungi glomus geosporum and glomus constrictum. *Applied and environmental microbiology*, 71(11):6673–6679. 82

Saunders, M. and Kohn, L. M. (2009). Evidence for alteration of fungal endophyte community assembly by host defense compounds. *New Phytologist*, 182(1):229–238. 8

Schimel, J. P. and Gulledge, J. (1998). Microbial community structure and global trace gases. *Global Change Biology*, 4(7):745–758. 2

Schloss, P. D., Gevers, D., and Westcott, S. L. (2011). Reducing the effects of pcr amplification and sequencing artifacts on 16s rrna-based studies. *PLoS One*, 6(12):e27310. 11, 18, 38, 39, 52

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*, 75(23):7537–41. 25, 69, 103, 105

Schulz, B. and Boyle, C. (2005). The endophytic continuum. *Mycological Research*, 109(6):661–686. 3

Seghers, D., Wittebolle, L., Top, E. M., Verstraete, W., and Siciliano, S. D. (2004). Impact of agricultural practices on the zea mays l. endophytic community. *Applied and Environmental Microbiology*, 70(3):1475–1482. 8

Shade, A. and Handelsman, J. (2012). Beyond the venn diagram: the hunt for a core microbiome. *Environmental Microbiology*, 14(1):4–12. 83

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., and Podar, M. (2013). Comparative metagenomic and rrna microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental microbiology*. 81, 91, 102, 106, 108

Shi, S., Richardson, A. E., O'Callaghan, M., DeAngelis, K. M., Jones, E. E., Stewart, A., Firestone, M. K., and Condron, L. M. (2011). Effects of selected root exudate components on soil bacterial communities. *FEMS microbiology ecology*, 77(3):600–610. 83

Sieber, T. N., Waisel, Y., Eshel, A., Kafkafi, U., et al. (2002). *Fungal root endophytes*. Number Ed. 3. Marcel Dekker Inc. 5

Simon, H. M., Dodsworth, J. A., and Goodman, R. M. (2000). Crenarchaeota colonize terrestrial plant roots. *Environmental microbiology*, 2(5):495–505. 4

Singh, B. K., Millard, P., Whiteley, A. S., and Murrell, J. C. (2004). Unravelling rhizosphere–microbial interactions: opportunities and limitations. *Trends in microbiology*, 12(8):386–393. 2

Singh, B. K., Nunan, N., Ridgway, K. P., McNicol, J., Young, J. P. W., Daniell, T. J., Prosser, J. I., and Millard, P. (2008). Relationship between assemblages of mycorrhizal fungi and bacteria on grass roots. *Environmental microbiology*, 10(2):534–541. 82

Sliwinski, M. K. and Goodman, R. M. (2004). Comparison of crenarchaeal consortia inhabiting the rhizosphere of diverse terrestrial plants with those in bulk soil in native environments. *Applied and environmental microbiology*, 70(3):1821–1826. 4, 100, 109

Smalla, K., Wieland, G., Buchner, A., Zock, A., Parzy, J., Kaiser, S., Roskot, N., Heuer, H., and Berg, G. (2001). Bulk and rhizosphere soil bacterial communities studied by denaturing gradient gel electrophoresis: plant-dependent enrichment and seasonal shifts revealed. *Applied and Environmental Microbiology*, 67(10):4742–4751. 7, 8, 9, 82

Smith, S. E. and Read, D. J. (2008). *Mycorrhizal symbiosis*. Academic Press. 5

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, 103(32):12115–12120. 14

Spencer, M. D., Hamp, T. J., Reid, R. W., Fischer, L. M., Zeisel, S. H., and Fodor, A. A. (2011). Association between composition of the human gastrointestinal microbiome and development of fatty liver with choline deficiency. *Gastroenterology*, 140(3):976–986. 7

Stackebrandt, E. and Ebers, J. (2006). Taxonomic parameters revisited: tarnished gold standards. *Microbiology today*, 33(4):152. 38

Sturz, A., Christie, B., Matheson, B., Arsenault, W., and Buchanan, N. (1999). Endophytic bacterial communities in the periderm of potato tubers and their potential to improve resistance to soil-borne plant pathogens. *Plant pathology*, 48(3):360–369. 8

Sun, L., Qiu, F., Zhang, X., Dai, X., Dong, X., and Song, W. (2008). Endophytic bacterial diversity in rice (oryza sativa l.) roots estimated by 16s rdna sequence analysis. *Microbial ecology*, 55(3):415–424. 5

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., and Mai, V. (2012). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Brief Bioinform*, 13(1):107–21. 18, 38

Suzuki, M. and Giovannoni, S. (1996). Bias caused by template annealing in the amplification of mixtures of 16s rrna genes by pcr. *Applied and environmental microbiology*, 62(2):625–630. 24, 37, 61

Taghavi, S., Garafola, C., Monchy, S., Newman, L., Hoffman, A., Weyens, N., Barac, T., Vangronsveld, J., and van der Lelie, D. (2009). Genome survey and characterization of endophytic bacteria exhibiting a beneficial effect on growth and development of poplar trees. *Applied and environmental microbiology*, 75(3):748–757. 3, 65, 66

Takai, K. and Horikoshi, K. (2000). Rapid detection and quantification of members of the archaeal community by quantitative pcr using fluorogenic probes. *Applied and Environmental Microbiology*, 66(11):5066–5072. 24

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, 28(10):2731–2739. 105

Team, R. C. et al. (2008). R: A language and environment for statistical computing. *Vienna, Austria: R Foundation for Statistical Computing*, pages 1–1731. 70, 72

Tourna, M., Stieglmeier, M., Spang, A., Könneke, M., Schintlmeister, A., Urich, T., Engel, M., Schloter, M., Wagner, M., Richter, A., et al. (2011). Nitrososphaera viennensis, an ammonia oxidizing archaeon from soil. *Proceedings of the National Academy of Sciences*, 108(20):8420–8425. 99, 107

Tringe, S. G. and Hugenholtz, P. (2008). A renaissance for the pioneering 16s rrna gene. *Curr Opin Microbiol*, 11(5):442–6. 10, 17

Tringe, S. G. and Rubin, E. M. (2005). Metagenomics: Dna sequencing of environmental samples. *Nat Rev Genet*, 6(11):805–14. 17

Turnbaugh, P. J. and Gordon, J. I. (2009). The core gut microbiome, energy balance and obesity. *The Journal of physiology*, 587(17):4153–4158. 7, 83

Turner, T. R., James, E. K., and Poole, P. S. (2013). The plant microbiome. *Genome biology*, 2013(14):209. 65

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006). The genome of black cottonwood, populus trichocarpa (torr. & gray). *Science*, 313(5793):1596–1604. 10, 65, 100

Tuskan, G. A., Gunter, L. E., Yang, Z. K., Yin, T., Sewell, M. M., and DiFazio, S. P. (2004). Characterization of microsatellites revealed by genomic sequencing of populus trichocarpa. *Canadian Journal of Forest Research*, 34(1):85–93. 67, 68

Uroz, S., Buée, M., Murat, C., Frey-Klett, P., and Martin, F. (2010). Pyrosequencing reveals a contrasted bacterial diversity between oak rhizosphere and surrounding soil. *Environmental Microbiology Reports*, 2(2):281–288. 4, 5, 6, 8, 109

Valentine, D. L. (2007). Adaptations to energy stress dictate the ecology and evolution of the archaea. *Nature Reviews Microbiology*, 5(4):316–323. 109

van Berloo, R. (2008). Ggt 2.0: versatile software for visualization and analysis of genetic data. *Journal of Heredity*, 99(2):232–236. 70

van der Heijden, M. G., Klironomos, J. N., Ursic, M., Moutoglis, P., Streitwolf-Engel, R., Boller, T., Wiemken, A., and Sanders, I. R. (1998). Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature*, 396(6706):69–72. 66

van der Lelie, D., Taghavi, S., Monchy, S., Schwender, J., Miller, L., Ferrieri, R., Rogers, A., Wu, X., Zhu, W., Weyens, N., et al. (2009). Poplar and its bacterial endophytes: coexistence and harmony. *Critical Reviews in Plant Science*, 28(5):346–358. 3, 65, 100

Van Overbeek, L. and Van Elsas, J. D. (2008). Effects of plant genotype and growth stage on the structure of bacterial communities associated with potato (solanum tuberosum l.). *FEMS microbiology ecology*, 64(2):283–296. 9

VerBerkmoes, N. C., Denef, V. J., Hettich, R. L., and Banfield, J. F. (2009). Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol*, 7(3):196–205. 18

Vestergård, M., Henry, F., Rangel-Castro, J. I., Michelsen, A., Prosser, J. I., and Christensen, S. (2008). Rhizosphere bacterial community composition responds to arbuscular mycorrhiza, but not to reductions in microbial activity induced by foliar cutting. *FEMS microbiology ecology*, 64(1):78–89. 82

Viebahn, M., Veenman, C., Wernars, K., Loon, L. C., Smit, E., and Bakker, P. A. (2005). Assessment of differences in ascomycete communities in the rhizosphere of field-grown wheat and potato. *FEMS microbiology ecology*, 53(2):245–253. 9

Watanabe, K., Kodama, Y., and Harayama, S. (2001). Design and evaluation of pcr primers to amplify bacterial 16s ribosomal dna fragments used for community fingerprinting. *Journal of Microbiological Methods*, 44(3):253–262. 24

Weisburg, W., Barns, S., Pelletier, D., and Lane, D. (1991). 16s ribosomal dna amplification for phylogenetic study. *Journal of bacteriology*, 173(2):697–703. 24, 61

Werner, J., Koren, O., Hugenholtz, P., DeSantis, T., Walters, W., Caporaso, J., Angenent, L., Knight, R., and Ley, R. (2011). Impact of training sets on classification of high-throughput bacterial 16s rrna gene surveys. *The ISME journal*, 6(1):94–103. 11, 18

Weston, D. J., Pelletier, D. A., Morrell-Falvey, J. L., Tschaplinski, T. J., Jawdy, S. S., Lu, T.-Y., Allen, S. M., Melton, S. J., Martin, M. Z., Schadt, C. W., et al. (2012). Pseudomonas fluorescens induces strain-dependent and strain-independent host plant responses in defense networks, primary metabolism, photosynthesis, and fitness. *Molecular Plant-Microbe Interactions*, 25(6):765–778. 65

Wilson, D. (1995). Endophyte: the evolution of a term, and clarification of its use and definition. *Oikos*, 73(2):274–276. 3, 5

Wrighton, K., Thomas, B., Sharon, I., Miller, C., Castelle, C., VerBerkmoes, N., Wilkins, M., Hettich, R., Lipton, M., Williams, K., et al. (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science*, 337(6102):1661–1665. 53

Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B. J., Hooper, S. D., Pati, A., Lykidis, A., Spring, S., Anderson, I. J., D'Haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C., Chen, F., Cheng, J. F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D., Rubin, E. M., Kyrpides, N. C., Klenk, H. P., and Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*, 462(7276):1056–60. 37, 51

Yang, C.-H., Crowley, D., and Menge, J. (2001). 16s rdna fingerprinting of rhizosphere bacterial communities associated with healthy and phytophthora infected avocado roots. *FEMS Microbiology Ecology*, 35(2):129–136. 7

Yang, C.-H. and Crowley, D. E. (2000). Rhizosphere microbial community structure in relation to root location and plant iron nutritional status. *Applied and Environmental Microbiology*, 66(1):345–351. 7

Yang, J., Kloepper, J. W., and Ryu, C.-M. (2009). Rhizosphere bacteria help plants tolerate abiotic stress. *Trends in plant science*, 14(1):1–4. 3

Yilmaz, S., Allgaier, M., and Hugenholtz, P. (2010). Multiple displacement amplification compromises quantitative analysis of metagenomes. *Nat Methods*, 7(12):943–4. 30

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y., Tu, Q., Xie, J., Van Nostrand, J., He, Z., and Yang, Y. (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME journal*, 5(8):1303–1313. 40, 52

Appendices

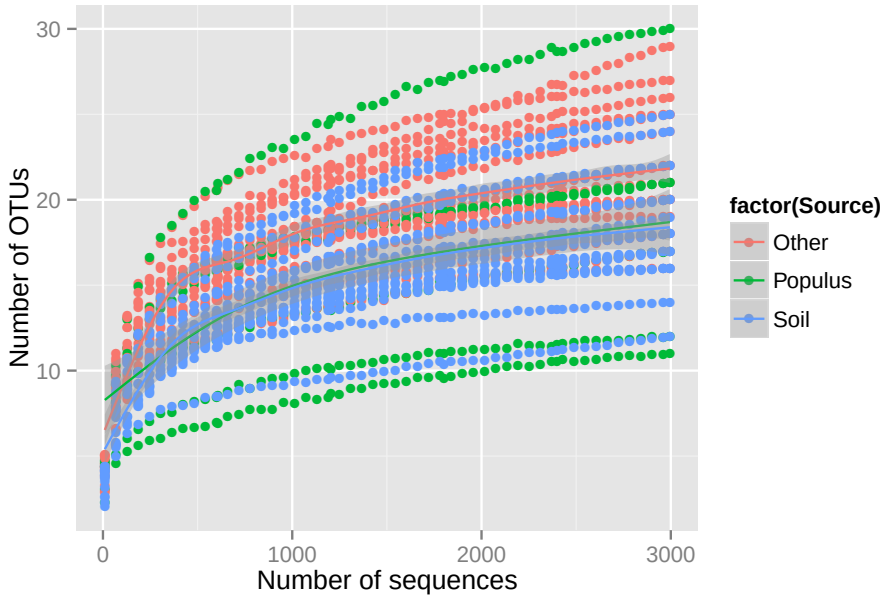


Figure 1: The rarefaction curve generated for the archaeal populations in all 41 samples. The points are color coded to represent the source of samples that could be soil, *P. deltoides* or surrounding trees' rhizosphere. The shaded line represents the average of corresponding points.

```
#!/usr/bin/env python
from __future__ import division

__author__ = "Migun Shakya"
__copyright__ = "Copyright 2007-2012, The Cogent Project"
__credits__ = ["Migun Shakya"]
__license__ = "GPL"
__version__ = "1.5.3-dev"
__maintainer__ = "Migun SHakya"
__email__ = "microbeatic@gmail.com"
__status__ = "Production"

from cogent import DNA, LoadSeqs, Sequence
from cogent.core.genetic_code import DEFAULT as standard_code
seqs = LoadSeqs(filename='amo_processed_1.fasta',moltype=DNA,aligned=False)
outfile=open("in-frame.fasta","w")

have_seen = {}
for label,seq in seqs.items():
        for i in range(3):
                frame = standard_code.getStopIndices(seq, start=i)
                if not frame:
                        if label in have_seen:
                                print "Multiple reading frames with out stop-codon found with: %s!" % label
                        else:
                                have_seen[label] = 1
                                outfile.write('>frame_'+str(i)+'_'+label+'\n'+str(seq)+'\n')
```

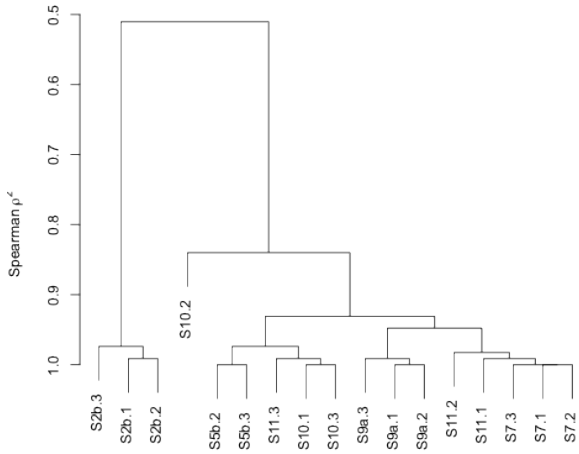


Figure 2: Cluster analysis of bulk soil samples based on measured variables.

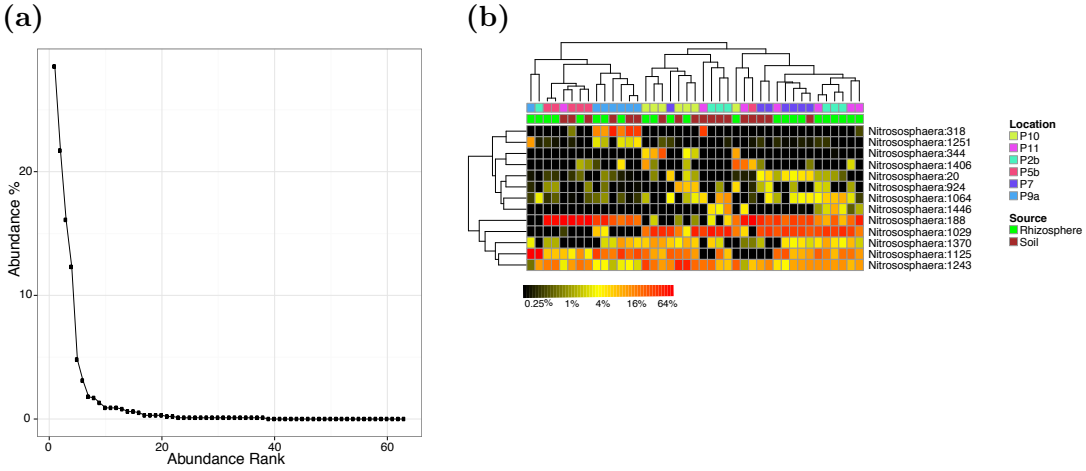


Figure 3: **(a)** Rank abundance plot of 16S rRNA gene based OTUs. **(b)** A heat map of OTUs that had a relative abundant of greater than 5% in any of the samples. The heatmap is scaled at logarithmic scale of 2.

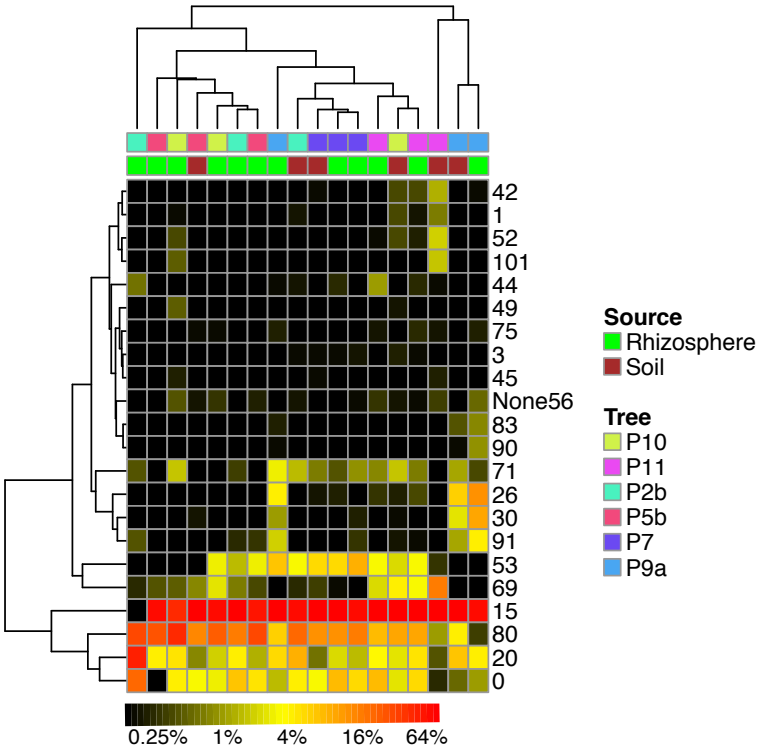


Figure 4: A heat map of OTUs based on partial *amoA* sequences that were clustered at 85% against a reference database. The heatmap is scaled at logarithmic scale of 2.

Vita

Migun Shakya was born to Mina Shakya and Buddha Ratna Shakya in Patan, Nepal. After completing high school at St. Xavier's in Kathmandu, Nepal he came to United States for his further studies. He completed his Bachelors of Science degree from Southwestern Oklahoma State University and later attended Genome Science and Technology Program at University of Tennessee, Knoxville for his PhD.