THE MAIZE RHIZOSPHERE MICROBIOME

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Zhao Jin

August 2014

THE MAIZE RHIZOSPHERE MICROBIOME

Zhao Jin, Ph. D.

Cornell University 2014

The rhizosphere microbiome, which is the microbial community living in close proximity to plant roots, is important for plant growth and development. Besides environmental factors, plant genetic control is key in cultivating the rhizosphere microbiome. Whether the genetic variation influences the taxonomy or function of the rhizosphere microbiome remains equivocal. I approached this question by culturing and sequencing 48 *Pseudomonas* isolates from two maize genotypes grown at two different fields, and analyzing the *Pseudomonas* genomes to identify components under maize genetic control. I observed a small but significant association of maize genotypes with the variation in the metabolic genes of the *Pseudomonas* isolates, while I did not see an association of maize genotypes with the abundance of the isolates.

Plant age is another important factor in shaping the rhizosphere microbiome, as plant age reflects changes in plant genetic control. Treating the rhizosphere microbiome as a quantitative trait, the proportion of this phenotypic variation attributable to plant genetic control can be measured as the heritability of the rhizosphere microbiome. To address how much variation in the maize rhizosphere microbiome is accounted for by maize genotypic variation, and to monitor how the heritability of the maize rhizosphere microbiome changes as maize grows and

develops, we sampled the maize rhizosphere microbiome from 27 diverse maize lines grown in three different fields over the entire maize growing season. I followed the temporal dynamics of the microbiome, and estimated the proportion of variation in the beta diversity of the rhizosphere microbiome samples at each time point explained by maize genotypes, fields, and genotype by field interactions. I found that the maize genotype effect starts to increase at week 2 after planting, suggesting that the maize genetic control is taking effect. I observed the strongest maize genotype effect around flowering time. I also identified some potential heritable taxa as well as OTUs whose abundances vary over maize developmental stages. In addition, I observed increased species loss starting at week 2, which corresponds to the time point when maize genetic control starts to take effect, whereas species loss peaks at flowering time when maize imposes the strongest genetic control on the rhizosphere microbiome.

Metagenomes are full of microbial “dark matters” that may harbor vast functional capacities. To optimize the function and decipher a functional region in a plant-growth promoting bacterial protein from the maize rhizosphere, I retrieved the rhizosphere bacterial protein regions and swapped them into *E. coli* to construct variant libraries, and selected the variant libraries for several rounds using nitrogen source limitation. I observed the fixation of known essential active site residues before the selection, and the fixation of several residues after selection, suggesting they are important for protein function. My results showed successful optimization and functional characterization of a region in this maize rhizosphere enzyme.

BIOGRAPHICAL SKETCH

Zhao was born in Wuhan, one of the Three Furnaces of China, a few days before the Chinese New Year in 1983. As a child, she enjoyed playing on the horizontal and parallel bars, chasing grasshoppers, and raising her pets, including chickens, parrots, a hedgehog, and turtles. She had wanted to become a zookeeper, or a diplomat as she was passionate about languages. After learning that China was only contributing to 1% of the Human Genome Project, she joined Wuhan University to study biological sciences as an undergrad in 2000. In 2004, she went to graduate school at Penn State University, where she studied iron-sulfur cluster assembly in Cyanobacteria, and got her master's degree in Biochemistry, Microbiology, and Molecular Biology in 2007. She then worked on lung diseases as a technician at the University of Pittsburgh. She enjoyed her stay in Pennsylvania, and especially loved Pittsburgh, a city in many ways resembles her hometown. With her interest in bacteria, she began her PhD studies in the Field of Microbiology at Cornell in 2009, and joined the lab of Dr. Ruth Ley's in 2010, when she began her research on plants' second genome, the rhizosphere microbiome, and explored the influence of maize genetic variation on the microbiome. She enjoyed every moment at the Ley lab and Cornell. With the generous approval and trust from her advisor and committee members, she moved to California for the last year of her PhD studies to work on her dissertation and stay with her husband. In Stanford University, she was warmly received and kindly hosted by Drs. Julie Parsonnet and Catherine Ley, and is now having a great time there wrapping up her PhD studies.

DEDICATION

献给我亲爱的爸爸妈妈

ACKNOWLEDGMENTS

I would like to express my greatest gratitude to my advisor, Dr. Ruth Ley, for her consistent support and guidance through my PhD years, and her trust and thoughtfulness in kindly granting me the privilege of working remotely in California for the last year of my PhD studies. Ruth has always been very encouraging and really helpful. Her passion for research is infectious, and her guidance for my study is insightful. I am truly fortunate to have Ruth as my PhD advisor, and I enjoyed every moment I had in the Ley lab.

I would like to thank my committee members, Drs. James Booth and Michael Stanhope, for their valuable and insightful suggestions for my research. Both professors have been very approachable and patient whenever I went to them with research questions. I have benefited tremendously from their help and guidance on my research work, and I am grateful to their assistance and critique during the final stages of my studies.

I appreciate very much the hospitality of Drs. Julie Parsonnet and Catherine Ley for hosting me and providing me an office space at Stanford University. I enjoyed going to the Parsonnet lab meetings to learn about interesting translational human microbiome research, and was delighted to have the lovely and chatty company of Cat, Lauren, and Mu in the office.

I thank my collaborators and former Ley lab members for making my research projects possible: Dr. Ed Buckler at Cornell for providing us the maize inbred lines, Drs. Susannah Tringe, Tanja Woyke, and Tijana Glavina del Rio as well as Mrs. Stephanie Malfatti, Sirisha Sunkara, and Lynne Goodwin at the Joint Genome Institute

for sequencing my 48 *Pseudomonas* genomes and 53 rhizosphere microbiome 16S amplicon plates, Dr. Jeff Dangl at University of North Carolina at Chapel Hill and Drs. Devin Coleman-Derr and Scott Clingenpeel at JGI for helpful discussions, and Drs. Rob Knight and Antonio Gonzalez Peña at the University of Colorado at Boulder for assisting my microbiome time-series analysis. In addition, my thanks go to Drs. Aymé Spor, Jason Peiffer, and Omry Koren for helping plant the maize inbred lines and collecting the rhizosphere soil samples, as well as Sarah Asman, Elle Glenny, Stephen Wu, and Eric Morris for their assistance in sample collection and DNA extraction.

I would like to thank everyone in the Ley lab for their friendship. Especially, in alphabetic order, Anders for his inspiring discussions and interesting anecdotes of his visits to China, Beth for her timely orders for my experimental supplies, Jessica for her constant availability whenever I asked for a hand, Julia for her awesome assistance on Qiime and R, Sara for her yummy, exquisite, and unique desserts and her immense help for editing the ACC deaminase manuscript, Sha for being my swimming buddy, and Wei for her huge help in my experiments.

My four years in Ithaca has been cheerful and memorable with the company of many friends: Xiaoxiao, Zihui, Qi, Xiaonan, Yibei, Zhen, Xuan, Ziying, Yi, Mizue, Yingying, and Yueting. I loved dining out and playing board games with them.

Last but not least, I thank my sweet husband, Ying Liu, for loving me and for accompanying me all these years with trust and support. Finally, I wish to thank my dear parents Mr. Dehua Jin and Ms. Junyi Liu, for showering me with their unconditional and unfailing love, support, and confidence all the time.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC – 1-aminocyclopropane 1-carboxylic acid

ACC-DR – ACC deaminase region

OTU – Operational Taxonomic Unit

UniFrac – unique fraction metric

QIIME – Quantitative Insights into Microbial Ecology

GLM – Generalized linear model

GLMM – Generalized linear mixed model

ANOVA – Analysis of variance

CHAPTER 1

An Overview of the Maize Rhizosphere Microbiome

Plant rhizosphere microbiome

Plants conduct photosynthesis and convert light energy to carbohydrates. Up to 40% of the plant photosynthates are released in the form of root exudates into the rhizosphere (Singh et al 2004), which is the area in close proximity to plant roots (Hartmann et al 2008). Root exudates are carbon-rich, and contain organic compounds such as sugars, organic acids, amino acids, fatty acids, proteins, and a number of plant secondary metabolites (Badri and Vivanco 2009). These compounds create unique ecological niches for microbes surrounding the roots, attracting to the vicinity a huge number of microorganisms that are collectively named as the rhizosphere microbiota, with the sum of microbial genomes being regarded as the microbiome (Hooper and Gordon 2001) (Figure 1.1).

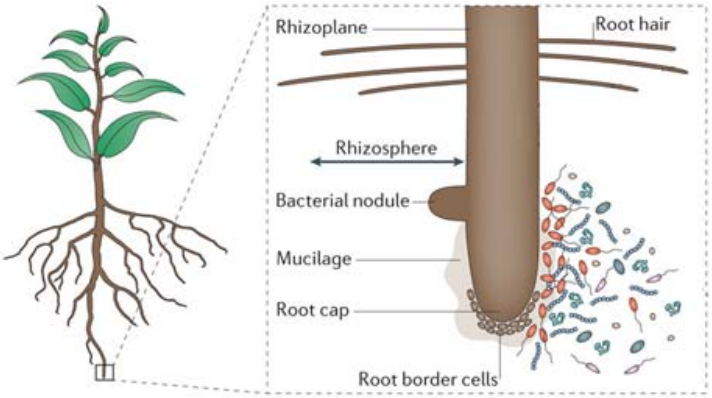

Figure 1.1 The plant rhizosphere and the rhizosphere microbiome. Adapted from Philippot et al 2013.

Plants and their rhizosphere microbiome are considered as “superorganisms” (Mendes et al 2011), in which plants interact closely with the microbes. Mediated through root exudates, plants directly or indirectly influence their rhizosphere microbiome. Besides providing energy source to the microbiome, root exudates and other rhizodeposits act as signals to recruit microbes to the rhizosphere. One

example is the legumes-rhizobia symbiosis: flavonoids secreted at the legume root surface attract rhizobia to colonize and infect root hairs and regulate bacterial nodulation factor gene expression (Abdel-Lateif et al 2012). Root exudates also shield plants from pathogenic microbes. For example, maize roots secrete the anti-fungal secondary metabolites 2,4-dihydroxy-7-methoxy-1,4-benzoxazin-3-one (DIMBOA) and 2,4-dihydroxy-1,4-benzoxazin-3-one (DIBOA) (Frey et al 2009). Other aspects of the influence from plants on their rhizosphere microbiome include adjusting the soil pH (Hinsinger et al 2003), facilitating growth of beneficial microbes (Cai et al 2009), interference with bacterial cell-cell communicating (Gao et al 2003, Proust et al 2011), and so on. Although the influence from plants is not the only factor, it has been proposed that these positive and negative influences from plants were key in shaping the rhizosphere microbiome (Dennis et al 2010).

Rhizosphere microbiome also interacts with and influences plants in a number of ways (Figure 1.2). One beneficial effect from roots-associated microbes includes decomposing soil minerals that are inaccessible to plants, thus providing plants with essential nutrients (Van Der Heijden et al 2008). For example, bacteria and fungi produce phytase that immobilizes inorganic phosphate, making it available for plants (Richardson and Simpson 2011). Roots-associated microbes also benefit plants in many other aspects, such as protecting plants from infection by soil-borne pathogens (Garbeva et al 2004, Mendes et al 2011), fixing nitrogen (Hsu and Buckley 2009), promoting root growth by producing phytohormones (Mavrodi et al 2006), and relieving plant abiotic stresses such as heavy metal contamination (Gamalero and Glick 2012), high salinity (Egamberdieva and Lugtenberg 2014), and drought (Kim et al 2012). These influences from the rhizosphere microbiome on plants are critical to plant growth and development.

Figure 1.2 The influence of the rhizosphere microbiome on plants. Adapted from Mendes et al 2013.

The rhizosphere microbiome has been compared to the gut microbiome (Berendsen et al 2012) as they share many similarities. These include mediation of host nutrient uptake, suppression of pathogen invasion to host, regulation of host immunity, and so on (Berendsen et al 2012). In addition, similar to mammals and their gut microbiome (Ley et al 2008), it has been proposed that plants and their rhizosphere microbiome co-evolve (Bakker et al 2012, Rosenberg and Zilber-Rosenberg 2013). Genetic variation is the building block for evolution. Therefore, to further understand the interactions between plants and their rhizosphere microbiome, it is important to learn more about how plant genotypes control the rhizosphere microbiome, and what compositions of the microbiome are related to plant genetic variation.

Plant genetic variation controls rhizosphere microbiome

Soil-borne microbial communities are generally affected by a number of abiotic factors. These mostly consist of factors resulting in heterogeneous environment and different geographic patterns, such as a number of soil physiochemical parameters (Chaparro et al 2012), including soil pH (Lauber et al 2009), temperature and moisture content (Bell et al 2009), carbon content (Cruz-Martínez et al 2012), carbon/nitrogen ratio (Nuccio et al 2013), mineral composition (Carson et al 2009), and so on. These factors play an important role in shaping the soil microbiome.

The influence from host plants, and the interactions between host plants and abiotic factors contribute further to the differences between the rhizosphere and bulk soil microbial communities, which has been termed the rhizosphere effect (Berendsen et al 2012). In addition, it has been shown that the influence from plants along with environmental heterogeneity resulted in variation in the rhizosphere microbiome. For example, our recent survey on the rhizosphere microbiome collected from multiple maize inbred lines grown at five different fields in the Northeast and Midwest demonstrated that biogeography, i.e., field heterogeneity, as well as maize genetic differences, both contributed to the variation in the diversity of the maize rhizosphere microbiome (Peiffer et al 2013). Numerous studies have further pointed the importance of plant genetic variation, which controls the types and timing of root exudates, as a critical regulator for rhizosphere microbiome. It is well known that different plant species harbor distinct rhizosphere microbiomes (Berendsen et al 2012). Within a plant species, studies on *Arabidopsis thaliana* rhizosphere bacterial communities have revealed that different Arabidopsis genotypes produced unique root exudates that closely regulated recruitment of different rhizosphere bacteria

(Micallef et al 2009b). In addition, it was shown that two Arabidopsis genotypes secreted different root exudates over time, which led to different assembly of rhizosphere bacterial communities (Micallef et al 2009a). These experimental evidences have all supported the relationship between plant genetic control and the variation in the rhizosphere microbiome, and that plant genetic variation is crucial in the differential assembly of rhizosphere microbiome.

Plant genetic control may select on the taxonomy of the rhizosphere microbiota. Different plant genotypes may recruit unique microbial taxa to their rhizosphere. For example, several potato cultivars shared many bacterial taxa in their rhizosphere, but also attracted some cultivar-dependent bacterial taxa to their roots (Weinert et al 2011). Another study showed that two different bacterial genera, *Pseudomonas* and *Serratia*, responded differently to volatile organic compounds (VOC) from two plants, *Lotus corniculatus* and *Saponaria officinalis*, suggesting that *Pseudomonas* preferred a more narrow selection of VOC, whereas *Serratia* could be recruited by a broader spectrum of VOC (Junker and Tholl 2013). It is possible that different plant genotypes that secret distinct VOC profiles in their root exudates are likely to attract different bacterial taxa to their rhizosphere. On the other hand, the relative abundance of the same bacterial taxa may also vary in the rhizosphere of different plants. For example, different relative abundances of bacteria from the orders Pseudomonadales, Actinomycetales, and Enterobacteriales were discovered in the study on the rhizosphere microbiome of three potato cultivars (Weinert et al 2011). In another study that compared the disease-suppressing rhizosphere microbiome to the disease-prone rhizosphere microbiome of sugar beet plants, researchers showed that the relative abundances of several bacterial classes, including Gammaproteobacteria, Betaproteobacteria, and Firmicutes, were

associated with whether the sugar beet rhizosphere microbiome was resistant or susceptible to *Rhizocbonia solani* infection (Mendes et al 2011). Thus, plant genetic variation is related to the differences in microbial taxa in the rhizosphere microbiome.

Plant genotypes may also select on the functional capacity of the rhizosphere microbiome. Bacteria have been divided to R-strategists, which grow fast on available nutrients, and K-strategists, which grow slowly and are more ubiquitous (Fierer et al 2007). The distinct blends of root exudates secreted by different plant genotypes may therefore attract different R-strategist bacteria to the rhizosphere based on the functional capacity of the bacteria to utilize the nutrients. For example, previous studies have shown the selection from different maize genotypes on the 2,4-diacethylphloroglucinol-producing *Pseudomonas* strains (Picard and Bosco 2006), the selection from rice cultivars on ammonia-oxidizing bacteria (Briones Jr et al 2003), and the preferences for the type-I methanotrophs over type-II methanotrophs (Wu et al 2009). Thus, plant genetic variation is also related to the differences in the functional capacity of the rhizosphere microbiome.

Although the impact of plant genetics on their rhizosphere microbiome has been widely studied, relatively less is known about the influence of plant genetic control on the rhizosphere microbiome within a plant species. Previous studies on the rhizosphere bacterial communities from different potato cultivars have uncovered that potato cultivars with more similar genotypes showed smaller differences in their rhizosphere bacterial communities compared to those of more different potato cultivars (Weinert et al 2009), and that potato cultivars recruited several cultivar-specific bacterial taxa and differed in the relative abundances of their shared bacterial taxa (Weinert et al 2011). One study on the Arabidopsis root microbiome from eight Arabidopsis accessions grown in a greenhouse revealed a small but significant

difference in the relative abundance of a few bacterial taxa in the endophytic compartment that could be used to differentiate the Arabidopsis accessions (Lundberg et al 2012). Another study on the Arabidopsis root microbiome from two Arabidopsis ecotypes grown in controlled field conditions identified one bacterial taxa with different abundance in the root microbiome from the two Arabidopsis genotypes (Bulgarelli et al 2012). Our recent survey on the maize rhizosphere microbiome from 27 maize inbred lines grown at five different fields at flowering time also discovered a small but truly significant maize genotype effect on the variation in the rhizosphere microbiome diversity (Peiffer et al 2013). Other studies on the rhizosphere microbiome from different genotypes within Medicago (Offre et al 2007, Zancarini et al 2012), Arabidopsis (Micallef et al 2009b), and soybean (Wang et al 2009) have also found that different plant genotypes within a plant species showed variation in their rhizosphere microbial populations. These studies established that plant genotypic variation, even within a plant species, is related to different rhizosphere microbiome.

One important aspect that received even less attention on the influence from plant genotypes on their rhizosphere microbiome is the heritability of the rhizosphere microbiome. Heritability refers to the proportion of phenotypic variation in a population accounted for by genetic variation of individuals. If the plant rhizosphere microbiome is treated as a quantitative trait, it is probably affected by plant genotypes and/or other abiotic factors, as well as the interaction between plant genetics and those factors. The heritability of the rhizosphere microbiome therefore answers how much of the variation in the rhizosphere microbiome is attributable to plant genetic control and/or other factors. An earlier investigation on the intraspecific heritability of root microbial communities from *Populus angustifolia* found that intraspecific plant

genotypic variation explained over 60% of the variation in microbial biomass nitrogen levels, and nearly 70% of the variation in the microbial community composition (Schweitzer et al 2008). In our recent study on the maize rhizosphere microbiome, the measurement of heritability was applied to the α- or β- diversity indexes of the microbiome by calculating how much of the variation in the microbiome diversity was accounted for by the variation in maize genotypes, field conditions, and maize genotype by field interactions using analysis of variance (Peiffer and Ley 2013, Peiffer et al 2013). This study revealed that within a field, maize genotypes explained nearly half and more than 20% of the α- and β-diversity of the rhizosphere microbiome, respectively, suggesting that the maize rhizosphere microbiome diversity is heritable (Peiffer et al 2013). Studies on mammalian gut microbiome provided additional insights into measuring the heritability and identifying the heritable components of microbiomes. One study employed quantitative trait locus (QTL) analysis to examine whether certain mouse gut microbiome bacterial taxa could be treated as quantitative traits that were associated with over five hundred single nucleotide polymorphisms (SNPs) in the animals. It was suggested that several mouse genomic regions and QTLs are associated with the variation in the relative abundance of a few lower-order bacterial taxa (Benson et al 2010). Human gut microbiome heritability studies using twins were equivocal, concluding that a strong host genotype effect may or may not contribute to the variation in the human gut microbiome (Spor et al 2011). Other studies have identified several bacterial families as heritable components of the chicken gut microbiome (Meng et al 2014, Zhao et al 2013). These results indicate that the heritability of microbiomes may be small, and may require better experimental design and analyses to be discovered.

Plant developmental stages influence rhizosphere microbiome

Besides plant genetic control and abiotic factors related to spatial heterogeneity, the taxonomy and/or functional capacity of the rhizosphere microbiome may also be under the influence of plant developmental stages. In general, time may be an important factor behind the changes in many soil physiochemical properties, such as moisture (Baskan et al 2013), nitrogen availability (Cain et al 1999), and C/N ratio (Zhang et al 2011), which may influence rhizosphere microbiome. More specifically, the composition of root exudates changes as plants age (Baudoin et al 2002, Chaparro et al 2013), which may reflect the variation of plant genetic control over time. Many studies probing rhizosphere microbiome in relation to plant development have been conducted in Arabidopsis. In a recent survey that examined the composition of the Arabidopsis rhizosphere microbiome at four Arabidopsis developmental stages, researchers discovered that young Arabidopsis seedlings cultivated a more different rhizosphere microbiome from older Arabidopsis plants, and that several bacterial taxa displayed temporal patterns in response to plant developmental stages. In addition, a number of genes in the Arabidopsis rhizosphere microbiome showed differential expression over time (Chaparro et al 2014). A chromatographic analysis on the root exudates collected from Arabidopsis at different developmental stages showed that Arabidopsis secreted varying percentages of sugars, sugar alcohols, amino acids, and phenolics over time. The variation in the root exudates were correlated with the functional genes involved in metabolizing the root exudates in the rhizosphere microbiome (Chaparro et al 2013). Studies of rhizosphere microbiomes of other plants also suggested a temporal pattern. For examples, rhizosphere bacterial taxa at the family and genus levels have been shown to vary significantly over the maize developmental stages of one maize

cultivar (Li et al 2014), and the structures of bacterial and fungal populations varied as Medicago transited from vegetation to reproduction (Mougel et al 2006), whereas the structure of bacterial communities changed significantly following potato developmental stages in all potato cultivars examined (van Overbeek and van Elsas 2008). Therefore, plant developmental stages also contribute to the variation in the rhizosphere microbiome.

The maize rhizosphere microbiome

To study within a plant species, the heritability of the rhizosphere microbiome, the change in the heritability over plant developmental stages, and whether plant genotypes select taxonomical or functional bacterial populations, I focused my research on the maize rhizosphere microbiome. Maize is one of the staple food crops in the world, and harbors extensive natural diversity and tractable genotypic and phenotypic information (McMullen et al 2009). The maize nested association mapping (NAM) population is a suite of maize strains developed by Ed Buckler and colleagues (Yu et al 2006) including 5,000 recombinant inbred lines (RILs) with identified genotypes and defined QTLs (McMullen et al 2009). Maize also has well-described growth stages (Meier 2001), with an important one, flowering, being mapped to numerous QTLs (Buckler et al 2009). These previous efforts allowed me to associate rhizosphere microbial phenotypic variation with maize genotypic variation and temporal factors. For this dissertation research, I used the maize rhizosphere soil samples from 27 NAM founder lines (Table 1.1) grown in a randomized complete block design in three fields in Ithaca, Lansing, and Aurora at New York state, one field in Urbana, Illinois, and one field in Columbia, Missouri, as described previously

(Peiffer et al 2013). The maize rhizosphere microbiome samples were collected every week from week one after planting to week 15 after planting. The week 20 maize rhizosphere microbiome samples were also collected.

Table 1.1 The 27 maize inbred lines and subgroups used in this dissertation.

Maize inbred lines	Subgroups
B73	Stiff stalk
B97	Non-stiff stalk
CML103	Tropical-subtropical
CML228	Tropical-subtropical
CML247	Tropical-subtropical
CML277	Tropical-subtropical
CML322	Tropical-subtropical
CML333	Tropical-subtropical
CML52	Tropical-subtropical
CML69	Tropical-subtropical
Hp301	Popcorn
Il14h	Sweet corn
Ki11	Tropical-subtropical
Ki3	Tropical-subtropical
Ky21	Non-stiff stalk
M162w	Non-stiff stalk
M37w	Mixed
Mo17	Non-stiff stalk
Mo18w	mixed
MS71	Non-stiff stalk
NC350	Tropical-subtropical
NC358	Tropical-subtropical
Oh43	Non-stiff stalk
Oh7B	Non-stiff stalk
P39	Sweet corn
Tx303	Mixed
Tzi8	Non-stiff stalk

Aims of study

Metagenomics has opened a window into the functional capacities of microbial communities in the environment, revealing a vast array of uncharacterized proteins that may be useful in many fields. While the structure and function of a small percentage of proteins in metagenomes are known, the remaining uncharacterized fraction remains a “dark matter” (Rinke et al 2013), ignored and omitted from most analyses. The first aim of this dissertation research is to optimize the function and decipher a functional region in the plant-growth promoting bacterial protein, 1-aminocyclopropane-1-carboxylic acid (ACC) deaminase, from the maize rhizosphere microbiome.

As described above, plant genotypes may select on the taxonomy or functional capacity of the rhizosphere microbiome. The second aim of this dissertation research is to investigate the influence of maize genetic variation on its rhizosphere *Pseudomonas* populations, and to find out whether maize genotypes were significantly associated with the variation in the *Pseudomonas* isolate genomes.

Relatively fewer studies were focused on the heritability of rhizosphere microbiome, and currently, no longitudinal study has been conducted to investigate the heritability of rhizosphere microbiome over time. Thus, the third aim is to measure the heritability of maize rhizosphere microbiome over the entire maize growth season, and to investigate whether the maize genetic control on the rhizosphere microbiome changes over time.

REFERENCES

Abdel-Lateif K, Bogusz D, Hocher V (2012). The role of flavonoids in the establishment of plant roots endosymbioses with arbuscular mycorrhiza fungi, rhizobia and Frankia bacteria. *Plant Signal Behav* **7:** 636-641.

Badri DV, Vivanco JM (2009). Regulation and function of root exudates. *Plant Cell Environ* **32:** 666-681.

Bakker MG, Manter DK, Sheflin AM, Weir TL, Vivanco JM (2012). Harnessing the rhizosphere microbiome through plant breeding and agricultural management. *Plant and soil* **360:** 1-13.

Baskan O, Kosker Y, Erpul G (2013). Spatial and temporal variation of moisture content in the soil profiles of two different agricultural fields of semi-arid region. *Environmental monitoring and assessment* **185:** 10441-10458.

Baudoin E, Benizri E, Guckert A (2002). Impact of growth stage on the bacterial community structure along maize roots, as determined by metabolic and genetic fingerprinting. *Applied Soil Ecology* **19:** 135-145.

Bell CW, Acosta-Martinez V, McIntyre NE, Cox S, Tissue DT, Zak JC (2009). Linking microbial community structure and function to seasonal differences in soil moisture and temperature in a Chihuahuan desert grassland. *Microbial ecology* **58:** 827-842.

Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J *et al* (2010). Individuality in gut

microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A* **107:** 18933-18938.

Berendsen RL, Pieterse CMJ, Bakker PAHM (2012). The rhizosphere microbiome and plant health. *Trends in Plant Science* **17:** 478-486.

Briones Jr AM, Okabe S, Umemiya Y, Ramsing N-B, Reichardt W, Okuyama H (2003). Ammonia-oxidizing bacteria on root biofilms and their possible contribution to N use efficiency of different rice cultivars. *Plant and soil* **250:** 335-348.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C *et al* (2009). The genetic architecture of maize flowering time. *Science* **325:** 714-718.

Bulgarelli D, Rott M, Schlaeppi K, Ver Loren van Themaat E, Ahmadinejad N, Assenza F *et al* (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* **488:** 91-95.

Cai T, Cai W, Zhang J, Zheng H, Tsou AM, Xiao L *et al* (2009). Host legume-exuded antimetabolites optimize the symbiotic rhizosphere. *Mol Microbiol* **73:** 507-517.

Cain ML, Subler S, Evans JP, Fortin MJ (1999). Sampling spatial and temporal variation in soil nitrogen availability. *Oecologia* **118:** 397-404.

Carson JK, Campbell L, Rooney D, Clipson N, Gleeson DB (2009). Minerals in soil select distinct bacterial communities in their microhabitats. *FEMS microbiology*

ecology **67:** 381-388.

Chaparro JM, Sheflin AM, Manter DK, Vivanco JM (2012). Manipulating the soil microbiome to increase soil health and plant fertility. *Biology and Fertility of Soils* **48:** 489-499.

Chaparro JM, Badri DV, Bakker MG, Sugiyama A, Manter DK, Vivanco JM (2013). Root exudation of phytochemicals in Arabidopsis follows specific patterns that are developmentally programmed and correlate with soil microbial functions. *PLoS One* **8:** e55731.

Chaparro JM, Badri DV, Vivanco JM (2014). Rhizosphere microbiome assemblage is affected by plant development. *ISME J* **8:** 790-803.

Cruz-Martínez K, Rosling A, Zhang Y, Song M, Andersen GL, Banfield JF (2012). Effect of rainfall-induced soil geochemistry dynamics on grassland soil microbial communities. *Applied and environmental microbiology* **78:** 7587-7595.

Dennis PG, Miller AJ, Hirsch PR (2010). Are root exudates more important than other sources of rhizodeposits in structuring rhizosphere bacterial communities? *FEMS microbiology ecology* **72:** 313-327.

Egamberdieva D, Lugtenberg B (2014). Use of Plant Growth-Promoting Rhizobacteria to Alleviate Salinity Stress in Plants. *Use of Microbes for the Alleviation of Soil Stresses, Volume 1*. Springer. pp 73-96.

Fierer N, Bradford MA, Jackson RB (2007). Toward an ecological classification of soil bacteria. *Ecology* **88:** 1354-1364.

Frey M, Schullehner K, Dick R, Fiesselmann A, Gierl A (2009). Benzoxazinoid biosynthesis, a model for evolution of secondary metabolic pathways in plants. *Phytochemistry* **70:** 1645-1651.

Gamalero E, Glick BR (2012). 19 Plant Growth-Promoting Bacteria and Metals Phytoremediation. *Phytotechnologies: Remediation of Environmental Contaminants***:** 361.

Gao MS, Teplitski M, Robinson JB, Bauer WD (2003). Production of substances by Medicago truncatula that affect bacterial quorum sensing. *Molecular Plant-Microbe Interactions* **16:** 827-834.

Garbeva P, van Veen JA, van Elsas JD (2004). Microbial diversity in soil: selection microbial populations by plant and soil type and implications for disease suppressiveness. *Annu Rev Phytopathol* **42:** 243-270.

Hartmann A, Rothballer M, Schmid M (2008). Lorenz Hiltner, a pioneer in rhizosphere microbial ecology and soil bacteriology research. *Plant and soil* **312:** 7-14.

Hinsinger P, Plassard C, Tang CX, Jaillard B (2003). Origins of root-mediated pH changes in the rhizosphere and their responses to environmental constraints: A review. *Plant and soil* **248:** 43-59.

Hooper LV, Gordon JI (2001). Commensal host-bacterial relationships in the gut. *Science* **292:** 1115-1118.

Hsu SF, Buckley DH (2009). Evidence for the functional significance of diazotroph community structure in soil. *ISME J* **3:** 124-136.

Junker RR, Tholl D (2013). Volatile organic compound mediated interactions at the plant-microbe interface. *J Chem Ecol* **39:** 810-825.

Kim Y-C, Glick BR, Bashan Y, Ryu C-M (2012). Enhancement of plant drought tolerance by microbes. *Plant Responses to Drought Stress*. Springer. pp 383-413.

Lauber CL, Hamady M, Knight R, 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:** 5111-5120.

Ley RE, Lozupone CA, Hamady M, Knight R, Gordon JI (2008). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology* **6:** 776-788.

Li X, Rui J, Mao Y, Yannarell A, Mackie R (2014). Dynamics of the bacterial community structure in the rhizosphere of a maize cultivar. *Soil Biology and Biochemistry* **68:** 392-401.

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S *et al*

(2012). Defining the core Arabidopsis thaliana root microbiome. *Nature* **488:** 86-90.

Mavrodi DV, Blankenfeldt W, Thomashow LS (2006). Phenazine compounds in fluorescent *Pseudomonas* spp. biosynthesis and regulation. *Annu Rev Phytopathol* **44:** 417-445.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q *et al* (2009). Genetic properties of the maize nested association mapping population. *Science* **325:** 737-740.

Meier U (2001). Growth stages of mono-and dicotyledonous plants. BBCH monograph. *German federal biological research centre for agriculture and forestry, Berlin*.

Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JH *et al* (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332:** 1097-1100.

Mendes R, Garbeva P, Raaijmakers JM (2013). The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS microbiology reviews* **37:** 634-663.

Meng H, Zhang Y, Zhao L, Zhao W, He C, Honaker CF *et al* (2014). Body weight selection affects quantitative genetic correlated responses in gut microbiota. *PLoS One* **9:** e89862.

Micallef SA, Channer S, Shiaris MP, Colon-Carmona A (2009a). Plant age and genotype impact the progression of bacterial community succession in the Arabidopsis rhizosphere. *Plant Signal Behav* **4:** 777-780.

Micallef SA, Shiaris MP, Colon-Carmona A (2009b). Influence of Arabidopsis thaliana accessions on rhizobacterial communities and natural variation in root exudates. *J Exp Bot* **60:** 1729-1742.

Mougel C, Offre P, Ranjard L, Corberand T, Gamalero E, Robin C *et al* (2006). Dynamic of the genetic structure of bacterial and fungal communities at different developmental stages of Medicago truncatula Gaertn. cv. Jemalong line J5. *New Phytol* **170:** 165-175.

Nuccio EE, Hodge A, Pett-Ridge J, Herman DJ, Weber PK, Firestone MK (2013). An arbuscular mycorrhizal fungus significantly modifies the soil bacterial community and nitrogen cycling during litter decomposition. *Environmental microbiology* **15:** 1870-1881.

Offre P, Pivato B, Siblot S, Gamalero E, Corberand T, Lemanceau P *et al* (2007). Identification of bacterial groups preferentially associated with mycorrhizal roots of Medicago truncatula. *Appl Environ Microbiol* **73:** 913-921.

Peiffer JA, Ley RE (2013). Exploring the maize rhizosphere microbiome in the field: A glimpse into a highly complex system. *Commun Integr Biol* **6:** e25177.

Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL *et al* (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc Natl Acad Sci U S A* **110:** 6548-6553.

Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH (2013). Going back to the roots: the microbial ecology of the rhizosphere. *Nature reviews Microbiology* **11:** 789-799.

Picard C, Bosco M (2006). Heterozygosis drives maize hybrids to select elite 2,4-diacethylphloroglucinol-producing *Pseudomonas* strains among resident soil populations. *FEMS Microbiol Ecol* **58:** 193-204.

Proust H, Hoffmann B, Xie XN, Yoneyama K, Schaefer DG, Yoneyama K *et al* (2011). Strigolactones regulate protonema branching and act as a quorum sensing-like signal in the moss Physcomitrella patens. *Development* **138:** 1531-1539.

Richardson AE, Simpson RJ (2011). Soil Microorganisms Mediating Phosphorus Availability. *Plant Physiology* **156:** 989-996.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF *et al* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431-437.

Rosenberg E, Zilber-Rosenberg I (2013). Role of Microorganisms in Adaptation,

Development, and Evolution of Animals and Plants: The Hologenome Concept. *The Prokaryotes: Prokaryotic Biology and Symbiotic Associations***:** 347-358.

Schweitzer JA, Bailey JK, Fischer DG, LeRoy CJ, Lonsdorf EV, Whitham TG *et al* (2008). Plant-soil microorganism interactions: heritable relationship between plant genotype and associated soil microorganisms. *Ecology* **89:** 773-781.

Singh BK, Millard P, Whiteley AS, Murrell JC (2004). Unravelling rhizosphere-microbial interactions: opportunities and limitations. *Trends Microbiol* **12:** 386-393.

Spor A, Koren O, Ley R (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology* **9:** 279-290.

Van Der Heijden MG, Bardgett RD, Van Straalen NM (2008). The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecology letters* **11:** 296-310.

van Overbeek L, van Elsas JD (2008). Effects of plant genotype and growth stage on the structure of bacterial communities associated with potato (Solanum tuberosum L.). *FEMS Microbiol Ecol* **64:** 283-296.

Wang G, Xu Y, Jin J, Liu J, Zhang Q, Liu X (2009). Effect of soil type and soybean genotype on fungal community in soybean rhizosphere during reproductive growth stages. *Plant and soil* **317:** 135-144.

Weinert N, Meincke R, Gottwald C, Heuer H, Gomes NC, Schloter M *et al* (2009). Rhizosphere communities of genetically modified zeaxanthin-accumulating potato plants and their parent cultivar differ less than those of different potato cultivars. *Appl Environ Microbiol* **75:** 3859-3865.

Weinert N, Piceno Y, Ding GC, Meincke R, Heuer H, Berg G *et al* (2011). PhyloChip hybridization uncovered an enormous bacterial diversity in the rhizosphere of different potato cultivars: many common and few cultivar-dependent taxa. *FEMS Microbiol Ecol* **75:** 497-506.

Wu L, Ma K, Lu Y (2009). Rice roots select for type I methanotrophs in rice field soil. *Syst Appl Microbiol* **32:** 421-428.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF *et al* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38:** 203-208.

Zancarini A, Mougel C, Voisin AS, Prudent M, Salon C, Munier-Jolain N (2012). Soil nitrogen availability and plant genotype modify the nutrition strategies of M. truncatula and the associated rhizosphere microbial communities. *PLoS One* **7:** e47096.

Zhang CH, Wang ZM, Ju WM, Ren CY (2011). Spatial and temporal variability of soil C/N ratio in Songnen Plain maize belt. *Huan jing ke xue* **32:** 1407-1414.

Zhao L, Wang G, Siegel P, He C, Wang H, Zhao W *et al* (2013). Quantitative genetic

background of the host influences gut microbiomes in chickens. *Sci Rep* **3:** 1163.

Chapter 2

Mutational Analysis and Optimization of a Maize Rhizosphere Metagenome Enzyme

Zhao Jin, Sara C. Di Rienzi, Anders Janzon, Jeff J. Werner, Douglas M. Fowler, Jeffrey L. Dangl, Ruth E. Ley

Abstract

Metagenomics has opened a window into the functional capacities of microbial communities in the environment, revealing a vast array of uncharacterized proteins that may have use in medicine, industry, and agriculture. While protein crystal structures and traditional mutational analyses are proven methods to determine the functional regions of a protein and to optimize its enzymatic activity, these methods are time consuming and difficult. Here I describe the use of a metagenomic library to optimize the function and decipher a functional region in the plant-growth promoting bacterial protein, 1-aminocyclopropane-1-carboxylate (ACC) deaminase region (DR), encoded by a rhizosphere microbial metagenome. I competed these ACC-DR variants in a selection assay based on ACC deaminase's capacity to provide nitrogen for the growth of *E. coli in vitro*. The most successful ACC deaminase region (ACC-DR) variants were identified after multiple rounds of selection using 454 pyrosequencing. I observed that the previously studied essential active site residues were already fixed in the metagenomic library and that residues within the previously structurally identified ACC deaminase small domain and helix 3 went to fixation after selection. In addition, I identified a divergent essential residue that hints at alternate substrates or other constraints in nature, and a cluster of neutral residues that did not influence the performance of ACC-DR variants in the selection assay. I observed the same fixation of one important and one divergent residue after selection in an artificial ACC-DR variant library generated by DNA oligomer synthesis. Therefore, by use of a simple competition assay and a metagenomic library, I was able to optimize and functionally characterize a region of a metagenomic enzyme.

Introduction

Environmental metagenomes are a rich and mostly uncharacterized reservoir of protein diversity encoded by a vast diversity of microorganisms. Metagenomes are mined for novel enzymes and products, such as the discovery of antibiotic resistant proteins (McGarvey et al 2012) and cellulose-degrading enzymes (Nacke et al 2012) from soil metagenomes. Metagenome sequences are also frequently generated to describe the functional attributes of microbial systems (Dantas et al 2013). While the structure and function of a small percent of proteins in metagenomes are known, the remaining uncharacterized fraction remains a “dark matter” (Rinke et al 2013), ignored and omitted from most analyses. Hence, facile and high throughput methods to understand the relationship between protein sequence and function of novel metagenomic proteins are needed.

Recently, deep-mutational scanning was developed as a method to elucidate the sequence-function relationships and optimal sequence of proteins (Fowler et al 2010). Using a doped DNA oligomer library and Illumina sequencing, Fowler et al. were able to map the mutational preferences of hundreds of thousands of protein variants for an important human protein domain and to show the fitness effects of all possible point mutations in the protein domain. Given the diversity of protein variants in the metagenomes, I hypothesized that the metagenome itself could be used as the pool of variants. Furthermore by using the metagenome as a source of enzyme variants, non-functional protein variants would already have been excluded, thereby reducing the total sequence search space.

The enzyme I targeted to construct a metagenomic library for optimization and mutational analysis is 1-aminocyclopropane-1-carboxylate (ACC) deaminase, an important plant-growth promoting protein. ACC deaminase is encoded by a wide variety of soil bacteria from the Proteobacteria, Firmicutes, and Actinobacteria phyla

(Glick et al 2007, Onofre-Lemus et al 2009). In the soil environment directly proximal to plant roots, the rhizosphere, bacteria convert ACC, the precursor of the plant stress hormone ethylene, to alpha-ketobutyrate and ammonia. This activity has been associated with relief from a number of plant stresses (Sheehy et al 1991) and promotion of root elongation (Glick 2004, Glick and Stearns 2011).

Previous structural and mutational studies on the *Pseudomonas* and yeast ACC deaminase proteins demonstrated that the ACC deaminase protein structures are very similar with several highly conserved amino acid residues involved in binding the substrate and cofactor. Using degenerate primers based on the alignment of bacterial ACC deaminase proteins, I amplified by PCR a 37 amino acid region from the full-length ACC deaminase. This region contains several previously identified conserved residues as well as variable residues, and is hereafter referred to as the ACC deaminase region (ACC-DR). Based on the previous structural analyses on bacterial and yeast ACC deaminase proteins (Fujino et al 2004, Karthikeyan et al 2004, Ose et al 2003, Yao et al 2000) and the alignment of characterized ACC deaminase sequences (Figure 2.1), the ACC-DR gene sequence displays high levels of conservation at regions coding for the active sites of the enzyme, whereas regions encoding non-active sites exhibit far higher levels of variation.

In order to further elucidate the function of the remaining residues in the ACC deaminase region and to test the use of a metagenomic library for mutational analysis and optimization, I cloned a maize rhizosphere metagenomic library of over 1000 ACC-DR variants into *E. coli* and conducted a growth selection assay based on the ability of ACC-DR gene variants to break down ACC and make nitrogen available for cell survival. Abundant (or fixed) ACC-DR gene variants after multiple rounds of selection were deemed the most efficient in this context. I first tested competing

rhizosphere bacterial ACC-DR variants. I then conducted the same selection assay using an artificial ACC-DR variant library generated by doped DNA oligomer synthesis using a winning ACC-DR variant from the soil libraries as the template for the oligo synthesis.

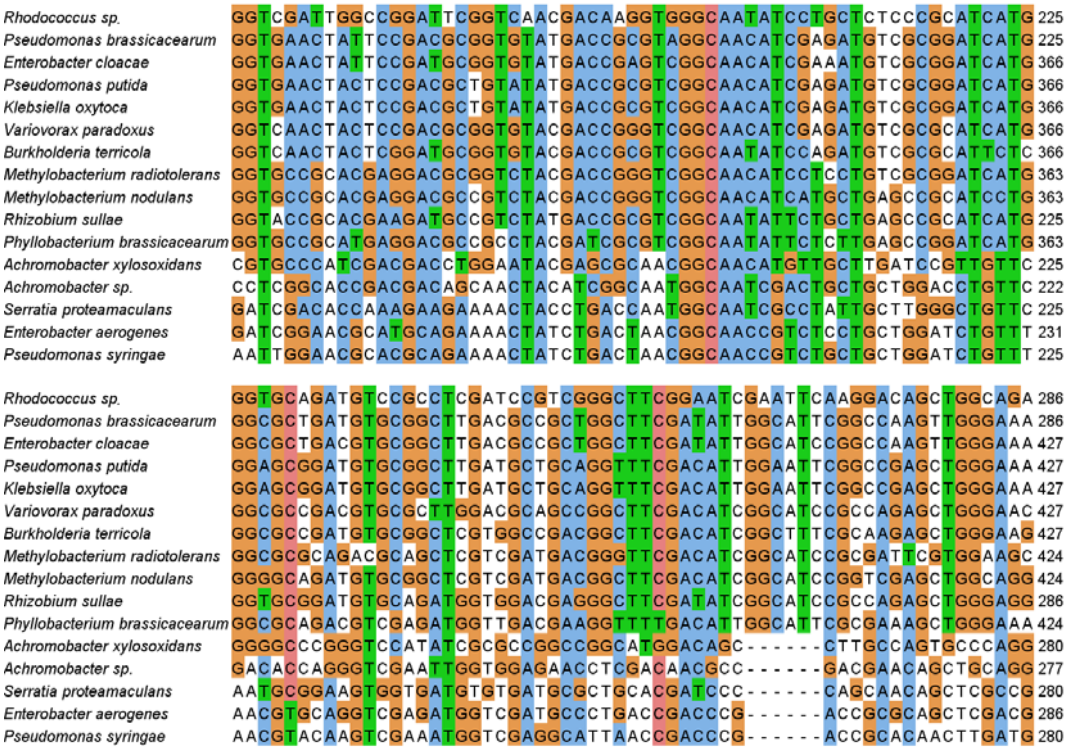

Figure 2.1 Alignment of soil bacterial ACC deaminase gene sequences. Shown in this figure is the Clustalx-colored alignment of 16 soil bacterial ACC deaminase gene sequences. The full-length alignment was truncated to the ACC-DR sequences due to the limit in figure size.

Multiple rounds of selection in replicate assays on the rhizosphere bacterial ACC-DR variant libraries demonstrated that the growth competition-based selection assay selected for the most beneficial residues in and around the active site of the ACC deaminase enzyme, and was able to reveal the importance of residues not previously known to be critical for ACC function. In the artificial ACC-DR variant

library, I observed the same fixation of one important residue and one divergent residue as those in the soil bacterial ACC-DR variant libraries after selection. Therefore, I demonstrated that a metagenome can be used as a starting source of variation in protein structure mutational analysis and optimization assays.

Methods

Rhizosphere samples and ACC deaminase plasmids

Missouri rhizosphere samples from maize inbred lines Oh43, MS71, M37W, and NC358 were collected in 2010 and DNA was extracted as previously described (Peiffer et al 2013) by a team of people in my lab. A Lansing rhizosphere sample from one maize plant of the week two B73 maize inbred line was used for the library construction.

A plasmid containing the *Pseudomonas cloacae* ACC deaminase and its flanking region, p4U2, was a generous gift from Dr. Bernard Glick at the University of Waterloo (Li and Glick 2001). The ACC deaminase region was deleted from p4U2 using the primer acdSdelF 5'-AATAGCGGCCTGGCCTTCGGCGCAGGAAAACTGGGTGAACTACT-3' and the Agilent QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies, Santa Clara, CA). The PCR reaction was as follows: 51 µl containing 5 µl 10X QuikChange reaction buffer, 1 µl p4U2 plasmid DNA (~50 ng), 1 µl acdSdelF primer, 1 µl QuikChange dNTP mix, 1.5 µl QuikSolution reagent, and 1 µl QuikChange Lightening Enzyme. Thermal cycling consisted of an initial denaturation at 95°C for 2min, 18 cycles of denaturation at 95°C for 20 s, annealing at 60°C for 10 s, and elongation at 68°C for 5 min, followed by a final extension at 68°C for 5 min. The plasmid without the *P. cloacae* ACC deaminase region is referred to as *p4U2ΔACC-*

DR.

Cloning of ACC deaminase variants into *E. coli*

The rhizosphere bacterial ACC deaminase regions were first amplified by PCR from the Lansing week2 A1 maize rhizosphere DNA using the primers acdSinserF 5'-AATAGCGGCCTGGCCTTCGGCGGSAACAAGACGCGCAAG-3' and acdSinserR 5'-CGGAGTAGTTCACCCAGTTTTCCTGCACSAGCACGCACTTCATG-3'. To capture the maximum diversity of ACC deaminase from the rhizosphere sample, three separate groups of PCR reactions were conducted, and each separate group consisted of five replicate PCR reactions. Each PCR replicate was 20 µl containing 2 µl rhizosphere DNA (~10 ng), 0.4 µl acdSinserF, 0.4 µl acdSinserR, and 10 µl 2X Phusion HF Master Mix (New England BioLabs, Ipswich, MA). Thermal cycling consisted of initial denaturation at 98°C for 30s, 30 cycles of denaturation at 98°C for 10 s, annealing at 57°C for 30 s, and elongation at 72°C for 1 min, followed by a final extension at 72°C for 7 min. For each group of PCR reactions, amplicons from the five replicate PCRs were combined, and purified using the QIAquick PCR Purification Kit (Qiagen, Valencia, CA). The three groups of amplicons were inserted into p4U2-del in three separate mutagenesis reactions similar to those described above. Each mutagenesis reaction generated a pool of rhizosphere bacterial ACC deaminase variants thereafter referred to as lineages 1, 2, and 3.

Growth-based selection assay

To assess the ACC deaminase function of the variants, each lineage of the plasmid library containing rhizosphere bacterial ACC deaminase variants was transformed into *E. coli* XL10-Gold chemical ultracompetent cells (Agilent Technologies, Santa Clara, CA) following the manufacturer's protocol. To maximize the diversity of ACC deaminase variants transformed into *E. coli*, three replicate

transformations were conducted for each lineage of ACC deaminase variants. The three transformations for each lineage of ACC deaminase variants were combined (total volume 1.65 ml), Lysogeny broth (LB) supplemented with 50 mg/ml Ampicillin was added to 5 ml, and grown at 37 °C overnight. The overnight culture was spun down, and washed twice in 0.1 M Tris-HCl buffer (pH 7.5). For each lineage, the washed cell pellets were resuspended in 1.65 ml DF minimal media (Dworkin and Foster 1958) minus $(NH_4)_2SO_4$, and supplemented with 0.2% dextrose, 50 mM $MgSO_4$, 1 mM $CaCl_2$, 50 mg/ml Ampicillin and 10 µg/ml thiamine. 300 µl of the washed overnight culture were frozen as the before selection ACC deaminase variant samples. 300 µl resuspended cell pellets normalized by the OD_{600} values of the previous round of cultures were added to 30 ml supplemented DF minimal media minus $(NH_4)_2SO_4$ in three replicates, and 1 ml 0.5 M 1-aminocyclopropane-1-carboxylate (ACC) was added lastly to the medium (hereafter this growth medium is called the DF/ACC medium) as the sole nitrogen source. The *E. coli* cells with ACC deaminase variants were grown at 30 °C for five days in the first round of selection. At the end of the first round of selection, the cultures were harvested, spun down, washed, and resuspended. 300 µl resuspended cells normalized by the OD600 values of the previous round of cultures were transferred into 30 ml fresh DF/ACC medium to start the second round of selection. The second round of selection consisted of three days of growth at 30 °C, and were passaged into fresh DF/ACC medium to start the third round of selection in a similar way. A total of six rounds of selection were conducted on the ACC deaminase variant library. Importantly, 300 µl of cultures were collected as ACC deaminase variant pool samples after each round of selection. Note that *E. coli* cells containing different rhizosphere ACC-DR variants had heterogeneous growth rates within each variant library and between libraries, so I

did not sync the *E. coli* cells to the same growth stage but rather ensured that I provided the same amount of *E. coli* cells to each round of selection across all three variant libraries based on normalization of OD_{600} values, and gave each library of variants the same length of growth time. Also note that because the time zero culture for lineage2 in Library B was lost due to the broken flask where the culture was grown, Library B ACC-DR variants contained only two lineages.

The test for cheaters, which tested whether *E. coli* cells without a functional ACC deaminase could grow on the nitrogen produced by *E. coli* cells with a functional ACC deaminase, was performed similarly to the selection assay except that two rounds of selection were conducted.

Illumina sequencing of the soil bacterial ACC deaminase region

The soil bacterial ACC deaminase region was amplified by PCR from the extracted DNA of the four Missouri maize soil samples using the degenerate primers a2F 5'-AATAGCGGCCTGGCCTTCGGCGCAGGAAAACTGGGTGAACTACT-3' and a2R 5'-CACSAGCACGCACTTCATG-3'. The amplicons were purified with the Agencourt AMPure XP PCR purification beads (Beckman Coulter, Indianapolis, IN). Addition of Illumina linker and adaptor sequences, and sequencing of the ACC deaminase regions on the Illumina Genome Analyzer IIx (Illumina Inc., San Diego, CA) were conducted by the Cornell University Life Sciences Core Laboratories Center.

Illumina sequences were analyzed by Dr. Jeff Werner. Illumina sequences were processed using in-house Perl scripts as follows. Paired-end sequences were joined based on aligning the overlapping region of 23 base pairs, with no internal gaps allowed. Reads were filtered by trimming at sites of low-quality bases (Q20 cutoff) from single-direction reads and discarding reads that lost more than six bases. Joined read pairs in which the overlapping sequence region between the forward and

reverse read disagreed internally. Up to three tailing bases (for each direction) that disagreed with the complimentary sequence were allowed to be trimmed, and it was confirmed that trimmed tailing bases had comparatively lower quality scores. The ACC-DR DNA sequences were translated to their corresponding amino acid sequences. The correct reading frame was determined by comparison to a known amino acid template sequence. Amino acid sequences were then clustered by absolute identity using UCLUST (Edgar 2010), to tabulate the protein-level diversity available in the metagenome pool of variants.

454 Sequencing of ACC Deaminase Variant Pools

Plasmid DNA was extracted from the ACC deaminase variant pool samples collected before selection and after each round of selection using the QIAprep Spin Miniprep Kit (Qiagen, Valencia, CA). The ACC deaminase regions were amplified by PCR from the plasmid DNA using the following composite primer pair: forward primer = 454 Titanium Lib-I Primer A/5-base barcode/a2F primer, and reverse primer = 454 Titanium Lib-I Primer B/a2R primer. Each sample was amplified in quadruplicate 20 µl-PCR reaction containing 1 µl plasmid DNA (~10 ng), 0.4 µl forward primer, 0.4 µl reverse primer, and 10 µl 2X Phusion HF Master Mix (New England BioLabs, Ipswich, MA). Thermal cycling consisted of initial denaturation at 98°C for 30s, 30 cycles of denaturation at 98°C for 10 s, annealing at 51.2°C for 30 s, and elongation at 72°C for 1 min, followed by a final extension at 72°C for 7 min. Following PCR, DNA amplicons were purified with the Agencourt AMPure XP PCR purification beads (Beckman Coulter, Indianapolis, IN), quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, Grand Island, NY), and pooled in equimolar ratios into a single sample with a final concentration of 30 ng/µl. Pyrosequencing was performed using the Roche GS FLX Titanium chemistry (454 Life Sciences, Branford, CT) at the

engencore facility in the University of South Carolina.

Analysis of ACC Deaminase Variant Pools

454 reads were analyzed using the QIIME software package (Quantitative Insights into Microbial Ecology) using default parameters for each step (Caporaso et al 2010). Sequences were chimera-checked and clustered into ACC deaminase variant clusters using Otupipe (Edgar et al 2011) at a sequence similarity threshold of 0.99. Each ACC deaminase variant cluster was represented by its most abundant sequence. A total of 33625 quality-filtered reads were obtained for 51 samples, an average of 659 reads per sample (min = 97, max = 10056). The forward and reverse primers were removed using a customized script. Due to the many indels in the sequences, a custom script was employed to maintain the correct length of the sequences. The P. cloacae ACC deaminase region was selected as the 'backbone' sequence. Using the EMBOSS water program (Rice et al 2000), each 454 read trimmed of both primers was aligned to the backbone. If an insertion was found relative to the backbone, the insertion was deleted in the 454 reads. If a deletion was found relative to the backbone, a gap was inserted into the 454 reads at the corresponding position. The insertion in the 454 sequences was easy to identify; however, the content of the gaps (i.e. what base to fill in the gaps) was impossible to determine within the limited context. Therefore, inevitably, a number of the resulting sequences still contained gaps. However, after this process, all sequences were of the same length, and the correct reading frame was maintained. The DNA sequences were translated into amino acid sequences, and the sequences that contained more than one unknown residue were excluded from the analysis. After this quality-filtering, 26764 reads remained for 51 samples, an average of 524 reads per sample (min = 83, max = 6710).

To calculate the β-diversity between the ACC deaminase variant pools before and after each round of selection, the OTU table was rarefied once at the depth of 80 reads per sample. A phylogenetic tree was built for the representative sequences of the ACC deaminase variant clusters using ClustalW (Larkin et al 2007), and the tree was used for calculating β-diversity using the UniFrac distance metrics (Lozupone and Knight 2005). To calculate the frequency of each DNA base or amino acid residue at every DNA/protein position, the OTU table was normalized by frequency. The 'plyr' (Wickham 2011) and 'reshape2' packages (Wickham 2007) were applied to numerate the DNA base/amino acid residue frequency in R v.2.15.0 (R Development Core Team 2010). The amino acid and DNA waffle plots were generated based on the frequency of the residues and bases using the R package ‘ggplot2’ (Wickham 2009). The structure of the H26 ACC deaminase variant was computed using homology modeling on the SWISS-MODEL server (Arnold et al 2006) with the P. sp. ACP ACC deaminase (the Q26 ACC deaminase variant, PDB ID 1TYZ) as the template (Karthikeyan et al 2004). The structures were visualized and aligned in PyMol (DeLano 2002). The characterized ACC deaminase protein sequences from bacteria, fungi, and plant were aligned using MUSCLE (Edgar 2004), and the soil bacterial ACC deaminase gene sequences were aligned based on the alignment of the corresponding protein sequences using PAL2NAL (Suyama et al 2006). The alignments were visualized in JalView 2 (Waterhouse et al 2009).

To identify the amino acid residues that were fixed or neutral in the selection assay, a linear regression was fitted to the frequencies of ACC-DR variants in each library at time zero and after each round of selection for each residue at each position. A residue is defined as being neutral if it has both positive and negative slopes in the three libraries, and is fixed if it has a starting frequency of over 0.1 and positive

slopes in all three libraries.

To test whether the eight important residues identified by the selection assay hitchhiked to fixation, independence tests were used to identify whether the residues were associated. For the ACC-DR variant sequences at time zero before the selection assay, one was assigned to the sequences at a certain position if the sequences contained the fixed residue at that positions, and a zero otherwise. The loglinear model-based independence tests was applied to the 8-way contingency tables generated from the binary data for ACC-DR variants in the three soil libraries. (Add SCA methods here later).

Molecular Evolution Analyses of ACC Deaminase Variant Pools

The codon-based Z-test (Nei and Gojobori 1986) and the Tajima's neutrality test (Tajima 1989) were performed in the Molecular Evolutionary Genetics Analysis (MEGA 6.0) program (Tamura et al 2013) on the ACC-DR DNA variants from all three soil libraries at time zero before the selection assay. The ACC-DR variants DNA variants were filtered by length and aligned based on their encoding protein sequences. A total of 16,432 sequences were obtained for the molecular evolution analyses. The codon-based Z-test calculates the test statistic d_S-d_N, with d_S and d_N representing the synonymous and nonsynonymous substitutions per site, respectively. The dataset was bootstrapped 500 times to estimate the variance, and the modified Nei-Gojobori method with Jukes-Cantor correction (assumed transition/transversion bias=15) (Zhang et al 1998) was selected as the substitution model. Any position that contained alignment gaps or missing data was eliminated for pairwise sequence comparisons. The probability of rejecting the null hypothesis of strict neutral (d_N=d_S) in favor of the alternative hypothesis (purifying selection with d_N<d_S) was measured and tabulated, and the level of significance was set at 5%. Tajima's neutrality test was

conducted using all codon (1st, 2nd, and 3rd) positions, and all positions containing gaps and missing data were eliminated. The Illumina-sequenced soil ACC-DR variant pool was analyzed in the same way, except that a random subset of 5000 sequences were used for the analyses.

DNA oligomer synthesis of for the artificial ACC deaminase variant pool

The DNA oligo was synthesized as described previously (Fowler et al 2010) (Gene Link, Hawthorne, NY). One of the DNA variants of the ACC deaminase 'LA' variant was chosen as the 'wildtype' backbone of the oligo. The DNA sequence of the winning 'LA' variant 'CTCGAATACCTGATCCCCGAGGCGCTGGCGCAGGGCTGCGACACGCTGGTGT CGATCGGCGGCATCCAGTCGAACCAGACACGCCAGGTTGCGGCCGTGGCTGC CCACCTGGG', which encoded 'LEYLIPEALAQGCDTLVSIGGIQSNQTRQVAAVAAHL'), and each base was doped with 2.1% non-wildtype nucleotides. The cloning of ACC deaminase variants into *E. coli*, the construction of the *E. coli* ACC deaminase variant library, and the growth-based selection assay on the artificial ACC deaminase variant pools were the same as described above for the soil rhizosphere bacterial ACC deaminase variant pools.

Results

Diversity of ACC deaminase genes in the Rhizosphere

To assess whether the rhizosphere sample was suitable as a source of ACC deaminase protein variants, I first probed the genetic diversity of bacterial ACC deaminase genes in four Missouri rhizosphere soil samples collected in 2010. Rhizosphere bacterial ACC deaminase genes are GC rich and highly polymorphic (Blaha et al 2006) with a few widely conserved regions (Figure 2.1). Thus, I designed a degenerate primer pair to amplify a 113-bp region from the ACC deaminase genes by PCR. This region, hereafter referred to as the ACC deaminase region (ACC-DR), contains several amino acid residues previously shown to be conserved in the active site as well as some variable residues (Karthikeyan et al 2004, Ose et al 2003, Yao et al 2000) (Figure 2.2).

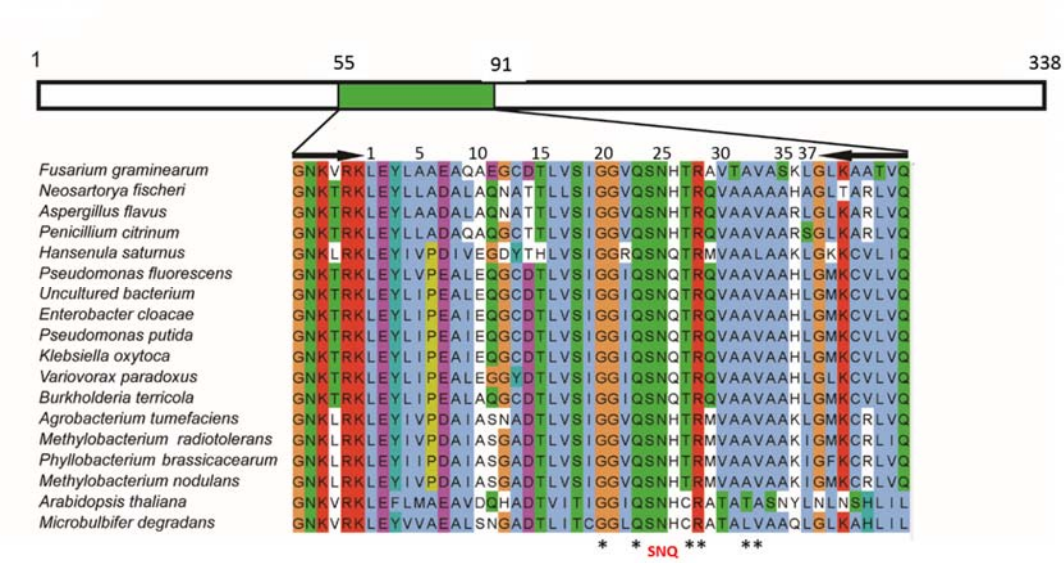

Figure 2.2 Alignment of ACC-DR from various organisms. Shown in this figure is the Clustalx-colored alignment of characterized ACC deaminase proteins in 18 organisms from bacteria, fungi, and plant. The full-length ACC deaminase is shown above the alignment, and the ACC-DR is marked in green. The amino acid numbering is based on the full-length *Pseudomonas putida* ACC deaminase. The arrows show the location of primers used in this study to amplify the ACC-DR, and 'SNQ' show the active site residues that bind cofactor and sulfate.

I assessed the variation in the ACC-DR from our rhizosphere soil samples using paired-end Illumina sequencing. After quality filtering, clustering, and removal of singletons, over 3.4 million different ACC-DR DNA variants remained, which encode over 450k different ACC-DR protein variants. These numbers are likely inflated due to sequencing errors, but overall the result indicates that the rhizosphere soil contains a high diversity of ACC-DR variants. The seven most abundant ACC deaminase protein variants from this rhizosphere soil comprised 51.5% of the sequences, and phylogenetic analysis indicated that they were encoded by the genera *Burkholderia* and *Pseudomonas* from the Proteobacteria phylum and *Tetrasphaera* and *Promicromonospora* from the Actinobacteria Phyla (Figure 2.3). In accord, previous work has reported that ACC deaminases are expressed by bacteria from the Proteobacteria and Actinobacteria phyla (Glick et al 2007, Onofre-Lemus et al 2009). Importantly, the high level of ACC deaminase protein diversity in this rhizosphere soil sample indicated sufficient diversity to serve as an initial variant pool for a selection assay. I conducted molecular evolutionary analyses on a random subset of sequences from this Illumina-sequenced natural ACC-DR variant pool, and found evidence for purifying selection on the ACC-DR (data not shown).

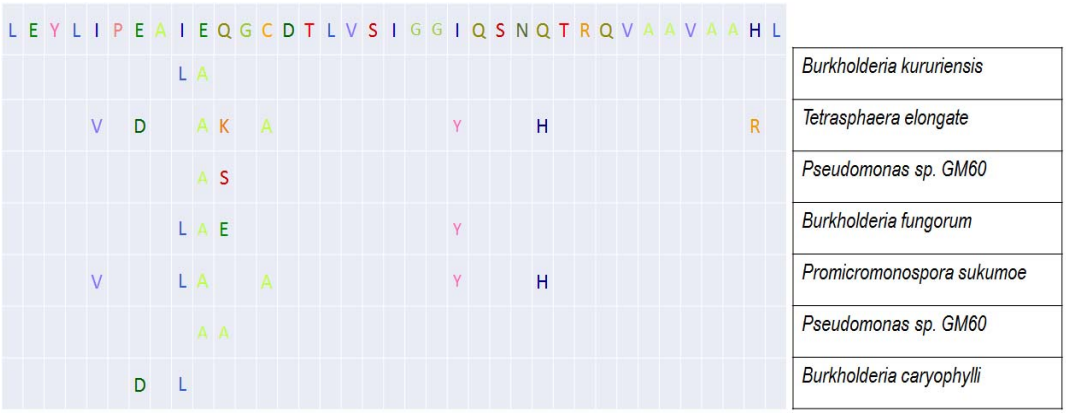

Figure 2.3 The seven most abundant rhizosphere bacterial ACC-DR protein variants. Rhizosphere bacterial ACC deaminase regions were amplified by PCR from the pooled DNA sample of four Missouri maize soil samples, and sequenced with Illumina paired-end sequencing. The first amino acid sequence shows the ACC deaminase region from the *P. cloacae* ACC deaminase. Below this sequence the seven most abundant ACC deaminase protein variants found in this DNA sample are shown. Listed at right are the top protein BLAST hits for these protein variants.

Selection assay allows survival of functional ACC-DR variants only

In order to identify the optimal ACC-DR sequences from the rhizosphere for ACC function in *E. coli*, I sought to utilize a competition assay in which the fittest ACC-DR variants would be selected for and enriched. The competition assay therefore requires that *E. coli* lacking a functional ACC deaminase gene cannot survive (i.e. cheat) the selection process by scavenging nitrogen released by co-occurring strains that do have ACC deaminase activity. To verify that this condition is met, I first competed *E. coli* cells lacking a functional ACC deaminase with *E. coli* cells with a functional ACC deaminase.

Li and Glick have shown previously that *E. coli* cells transformed with a plasmid (p4U2) containing the ACC deaminase gene from *Pseudomonas cloacae* display ACC deaminase activity (Li and Glick 2001) and are able to grow with ACC as the sole nitrogen source. I confirmed that *E. coli* cells, which lack ACC deaminase, fail to grow when ACC is the sole nitrogen source, but are able to grow when transformed

with p4U2 (data not shown). Furthermore, I verified that *E. coli* cells containing the plasmid lacking the ACC deaminase region (p4U2 ΔACC-DR) also fail to grow when ACC is the sole nitrogen source (data not shown).

Next, I mixed in a 1:1 ratio *E. coli*–p4U2 with *E. coli–p4U2ΔACC-DR,* and grew the mixed populations with ACC as the sole nitrogen source (Figure 2.4a). After the first round of selection (see Methods), both *E. coli* types were still present (Figure 2.4b); however, *E. coli*–p4U2ΔACC-DR disappeared after the second round of selection, (Supplementary Figure 3b). This experiment confirmed that the selection assay would not allow the growth of cheater strains lacking ACC deaminase activity.

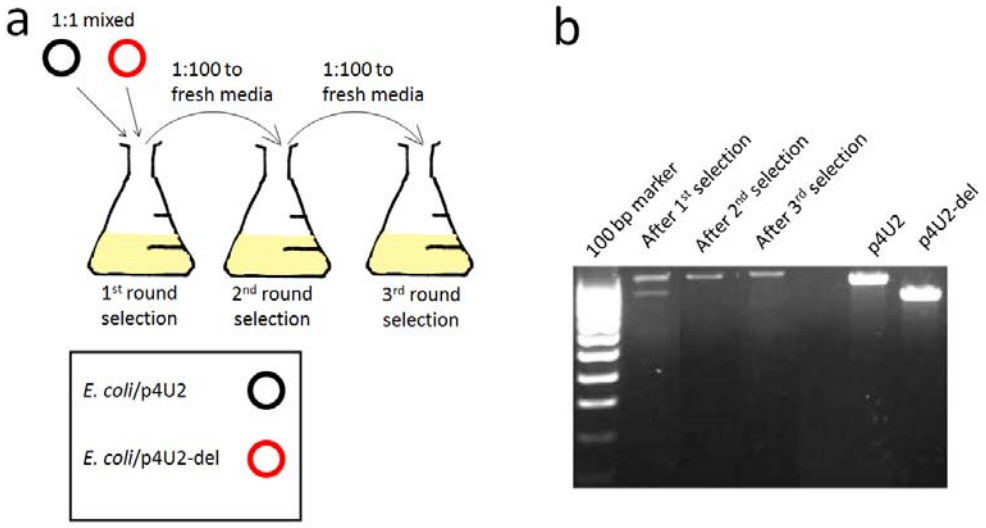

Figure 2.4 Non-functional ACC-DR variant cannot grow on ammonia produced by other ACC-DR variants. (a) *E. coli* containing the P. cloacae ACC deaminase on the plasmid (p4U2, black circle) was mixed in a 1:1 ratio with *E. coli* containing the plasmid lacking the ACC-DR (p4U2-del, red circle). The two variants were grown in the DF/ACC media for three rounds of selection. (b) A region containing the ACC-DR was amplified by PCR from the mixed growth samples. The PCR amplicons from p4U2 and p4U2-del were used as controls.

Construction and selection of ACC-DR variant library

In order to generate *E. coli* libraries with ACC-DR variants, I amplified the bacterial ACC-DR from a rhizosphere DNA sample isolated from a similar rhizosphere soil sample as that described above (see Methods for the description of this second soil). I expected a similar degree of diversity among the rhizosphere DNA samples as ACC deaminase is widespread in rhizosphere bacteria (Glick et al 2007). These variants were pooled from five PCR reaction replicates and cloned into p4U2 by domain swapping. The ACC-DR library was transformed into *E. coli* grown in a minimal salt medium with ACC as the sole nitrogen source (Figure 2.5) to select for successful transformants.

This library construction protocol was conducted three times from the same rhizosphere DNA sample to produce three individual *E. coli* libraries (Libraries A, B, C; one library is shown as an example in Figure 2.5) to maximize the diversity recovered from the rhizosphere. Libraries A, B, and C ACC-DR variant pools contained 891, 742, and 560 ACC-DR DNA variant clusters at 99% similarity, which encoded 310, 268, and 226 ACC-DR protein variants, respectively. In total, the libraries represented 1,220 unique DNA variants encoding 455 protein variants. Two reasons may account for a lower diversity of ACC-DR variants in the *E. coli* libraries A, B, and C: 1) I conducted the selection assay in liquid culture and 2) I sequenced the PCR amplicons of rhizosphere ACC-DR variants for the Illumina sequence directly without cloning them into *E. coli*.

Each library underwent six rounds of selection in triplicate. These triplicate competition assays were referred to as lineages (i.e. for library A, linages A_1, A_2, A_3). For each lineage, I collected samples prior to the selection (i.e. time zero) and samples after each of the 6 rounds of selection (Figure 2.5).

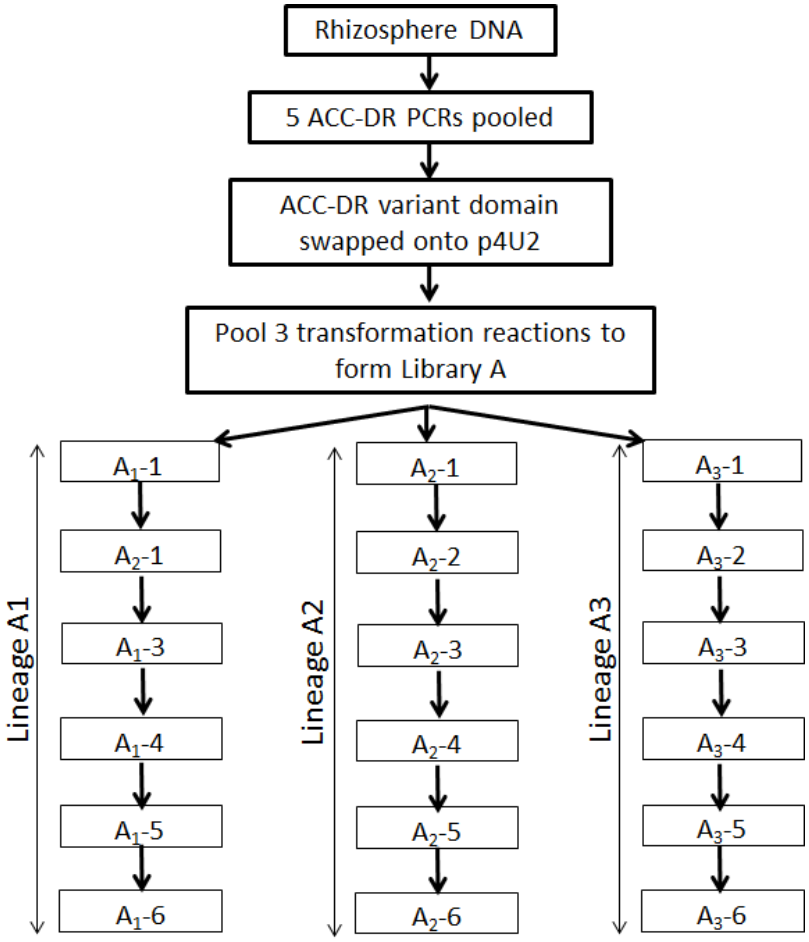

Figure 2.5 Construction of *E. coli* ACC-DR variant libraries and growth-based selection assay. Shown in this figure is the experimental design for one of the *E. coli* Libraries (Library A) with ACC-DR variants, and the growth-based selection assay for this library. The *E. coli* Libraries B and C with ACC-DR variants were constructed and selected in the same way as Library A (see Methods for details of library construction and selection assay).

Effect of selection on the diversity of ACC-DR variants

To gain a coarse overview of the impact of selection on the genetic diversity of the ACC-DR gene variants, I estimated the β-diversity (between-sample diversity) of the ACC-DR pools from Libraries A, B, and C using the unweighted UniFrac distance metric (Lozupone and Knight 2005). The UniFrac metric ranges from 0 to 1 and is based on the unique/shared fraction of a common phylogeny relating the gene variant

sequences, such that any two pools with closely related variants will have a low UniFrac value, while two pools with phylogenetically less related content will have a value closer to 1. Distances are computed for all pair-wise comparisons and principal coordinates analysis (PCoA) of the distance matrix is applied to display the relationships between pools. In all three libraries, the first round of selection clearly impacted the overall diversity (Figure 2.6), and had a stronger influence on the overall variant diversity than subsequent rounds of selection. (Figure 2.6).

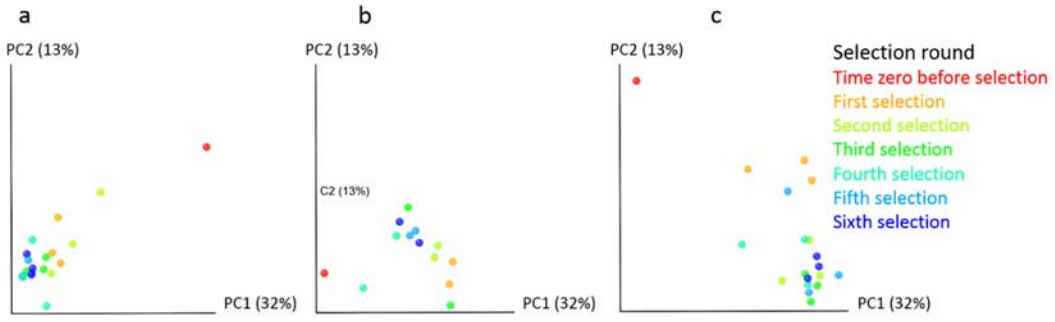

Figure 2.6 ACC-DR variant pools cluster by selection round. The ACC-DR variant pools after each round of selection for the three replicate lineages within each library are clustered by round of selection in a PCoA of the unweighted UniFrac distances between samples. The percentage of variation explained by the principal coordinates is indicated on the axes. The ACC-DR variant pools are colored by a gradient from red to blue, and each point corresponds to an ACC-DR variant pool colored by selection round: red, time zero before selection, orange, cyan, green, aqua, teal, and blue, after the first to the sixth round of selection, respectively. Lineages for the same library are represented by dots of the same color for each round of selection, (a) ACC-DR variant pools in Library A. (b) ACC-DR variant pools in Library B,. Note that library B only contained two lineages. (c) ACC-DR variant pools in Library C.

Purifying selection fixed most essential residues before selection assay

In order to understand the specific effects of the selection assay on the ACC-DR, I began by analyzing previously reported essential amino acid residues (Karthikeyan et al 2004, Ose et al 2003, Yao et al 2000). Most essential residues (G20, Q23, S24, N25, T27, R28, A34, and A35) that are involved in binding cofactor and substrate as

well as ACC deaminase monomer-monomer interaction (Yao et al 2000) were already fixed in the starting libraries with over 90% frequency, suggesting strong selection pressures on ACC deaminase in rhizosphere bacterial populations. To confirm this supposition, I used the codon-based Z-test (Nei and Gojobori 1986), and calculated the Tajima's D value (Tajima 1989) (Table 2.1).

Table 2.1 Results from Tajima's neutrality test on the rhizosphere bacterial ACC-DR variants from Libraries A, B, and C at time zero before selection.

m	S	ps	Θ	Π	D
16433	110	0.964912	0.093824	0.004721	-2.38002
NOTE.-- The analysis involved 16433 nucleotide sequences. Codon positions included were 1st+2nd+3rd. All ambiguous positions were removed for each sequence pair. There were a total of 114 positions in the final dataset. Evolutionary analyses were conducted in MEGA6.					
Abbreviations: m = number of sequences, n = total number of sites, S = Number of segregating sites, ps = S/m, Θ = ps/a1, Π = nucleotide diversity, and D is the Tajima test statistic.					

A negative Tajima's D value of -2.380024 indicated the presence of purifying selection in the starting library, and the subsequent codon-based Z-test results (not shown due to MEGA6 export size limit) also showed a high probability ($p < 0.05$) to reject the null hypothesis of strict-neutrality in favor of the alternative hypothesis of purifying selection. Hence, nature has already selected on the function of the ACC deaminase variants in soil bacteria, and provides us with a starting point to optimize and understand the functionality of the less constrained residues in the ACC deaminase region.

Selection on a divergent essential residue

One essential site previously identified by structural analyses, however, was not fixed in the starting library. In a small proportion of the starting library, residue 26

contained a divergent amino acid that is present in some rhizosphere bacterial ACC deaminase sequences. Karthikeyan et al. have reported that Q26 interacts with the bound sulfate in the active site of the protein (Karthikeyan et al 2004), and is important for ACC function. In each library, both glutamine (Q) and histidine (H) were present at residue 26 in the population prior to selection, but glutamine was fixed or enriched after the first round of selection (Figure 2.7).

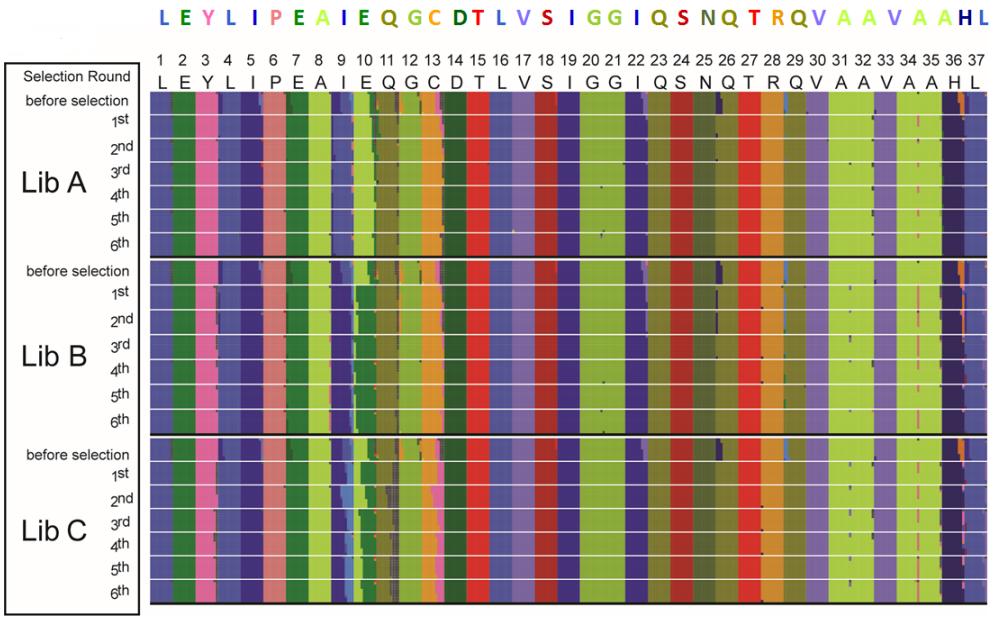

Figure 2.7 Amino acid residue waffle plots for the Libraries A, B, and C. The amino acid residue waffle plots show the frequency of each residue in ACC-DR variant pools at time zero and after each round of selection averaged for the three lineages in each library. Each amino acid residue is represented by a unique color, and the percentage of grids of the same color shows the frequency of that residue at a position. The amino acid residues on top of the waffle plots are color coded, and represent the sequence in the P. cloacae ACC-DR. The number on top of each amino acid residue shows the position of the residue from 1 to 37. Amino acid residues A, C, D, E, F, G, H, I, K, L, M, N, , Q, R, S, T, V, W, and Y are colored as shown in the letters , and the same colors are used to show their relative abundances.

To understand why the H26 ACC-DR variant was quickly excluded from the variant pool by the selection assay, I employed homology modeling to estimate the

structure of this variant. I used the known *Pseudomonas sp. ACP* ACC deaminase crystal structure (PDB ID 1TYZ), which encodes the Q26 ACC-DR variant, as the template to model the structure of the H26 variant. I found that these two structures were very similar when aligned (Figure 2.8). However, because the side chain of the glutamine residue interacts with the bound sulfate ion in the active site of the protein (Karthikeyan et al 2004, Ose et al 2003, Yao et al 2000), a change from an uncharged to a charged amino acid may impact the efficiency of the deaminase in the selection assay. Thus, it appears that while Q is more beneficial for binding ACC in *E. coli,* H may favor an alternative substrate or context in nature.

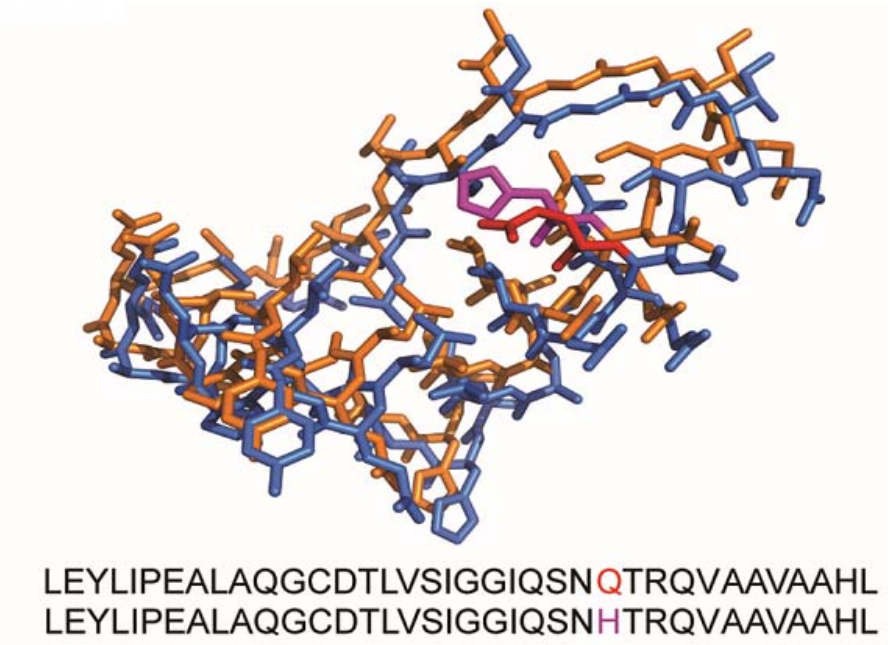

Figure 2.8 Alignment of the 3D-structures for *Pseudomonas. sp. ACP* ACC deaminase and homology-modeled ACC-DR. Alignment of the *P. sp. ACP* ACC-DR, shown in blue, and the homology-modeled ACC-DR based on the *P. sp. ACP* ACC deaminase structure is shown in orange. The Q26 residue in the *P. sp. ACP* ACC-DR is colored in red, and the H26 residue in the homology modeled ACC-DR is colored in purple. The other regions of the full-length ACC deaminase protein structures from these two variants are identical, and are omitted in the alignment.

Selection assay reveals importance of non-essential residues

My selection assay acted on other sites not previously reported to be essential. ACC-DR variants with a leucine (L) at residue 4 were enriched after one or two rounds of selection in all libraries (Figure 2.7). Yao et al. have reported that this residue is located on helix 3, which is in close contact with helix 2, which binds the cofactor in the *Hansenula saturnus* ACC deaminase (Yao et al 2000). Although the underlying structural mechanism is not clear, this functionality may explain why the ACC-DR variants containing L4 were enriched by my selection assay.

Other sites with completely unknown roles in ACC deaminase function were found to fix in the selection assay. Residues I5, E7, G12, C13, I22, Q29, H36 and L37 were fixed at selection round one. Based on the known structure of the yeast ACC deaminase protein structure, which is highly similar to bacterial ACC deaminase structure, residue I22 is on a loop between β-strand C and α-helix 4 of the protein, which is involved in linking the active site cavity to the surface of the protein (Yao et al 2000). The ACC deaminase consists of two domains (Karthikeyan et al 2004), a small domain of unknown function, and the cofactor-binding domain. As components of the small domain, residues I5, E7, G12, C13, Q29, H36, and L37 may help maintain the overall shape of the protein (Yao et al 2000).

Given the proximity of many of residues with each other, I tested the time zero before selection ACC-DR variants for independence among the eight residues to determine if selection at one residue was accompanied by concomitant changes at another site. My results indicated that these residues were significantly associated ($p < 10e-16$). To further elucidate what subsets of the eight residues were likely to be selected together in my assay, I employed statistical coupling analysis (SCA) (Lockless and Ranganathan 1999) to the time zero before selection rhizosphere

bacterial ACC-DR variants (Supplementary Figure 4). SCA calculates the sequence similarity of ACC-DR variants based on the multiple sequence alignment of the time zero before selection ACC-DR variants, and constructs a positional correlation matrix of all residues in the ACC-DR. All residue pairs within the fixed residues (residues 4, 5, 7, 12, 13, 22, 26, 29, 36, and 37) were more correlated than others. Thus, I could not exclude the possibility that these residues went to fixation together because of stronger correlation. However, as these residues are scattered through α-helices 4 and 5, as well as the loops connecting α4, α5 and β-sheet 3 in the protein structure, it is likely that these residues function together to increase the efficiency of ACC-DR variants in an unknown fashion..

While most emphasis on the studies for ACC deaminase has been focused on the PLP-binding domain of ACC deaminase, my results revealed that residues in other parts of the protein, especially the small domain, are also critical for the optimal efficiency of the enzyme. Therefore, my selection assay is able to reveal additional sites that impact the functional performance of ACC deaminase in *E. coli*, and highlights the significance of such assays to unravel the hidden structural info that may play important roles in protein function.

Selection assay identifies neutral sites with no influence on enzyme function

Other sites remained heterogeneous throughout the selection process. Residues 9 and 10, for example, bore a mixture of several residues (predominantly IE and LA) prior to selection in Library A. After the first round of selection, the ACC-DR variants with the LA residues began to dominate the population, although the IE variants were still present in the population at a much smaller frequency (Figure 2.7). Similarly, the IE ACC-DR variants in Library B became dominant after the first round of selection

(Figure 2.7). The ACC-DR variants from Library C contained a mixture of I/L/M and A/E at residues 9 and 10, respectively, prior to selection, and there was no clear winner after six rounds of selection (Figure 2.7). Together these data suggest that LA and IE do not differentially affect the function of the ACC-DR.

To test this hypothesis, I constructed the LA and IE ACC-DR protein variants on the identical background so that the variants only differed at the 9th and 10th residues, and grew *E. coli* cells containing either variant separately on ACC as the sole nitrogen source. I did not observe any significant difference in their growth rates (Figure 2.9), indicating that these two residues impose a neutral influence on the efficiency of ACC deaminase in *E. coli*.

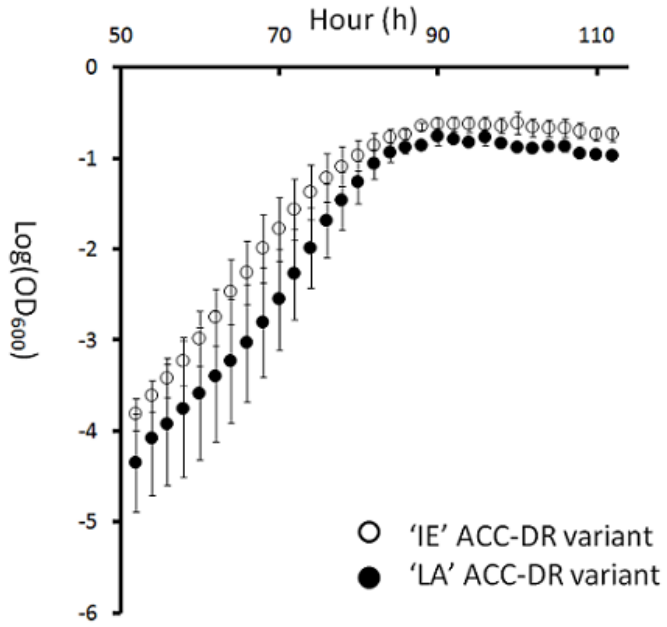

Figure 2.9 Individual growth curves of *E. coli* cells containing 'IE' and 'LA' ACC-DR variants in ACC/DF media. This figure shows the growth curves of *E. coli* cells containing the 'IE' and 'LA' ACC-DR protein variants grown individually in the DF/ACC media. These two ACC-DR protein variants were identical in other residues except for the 'IE' or 'LA' residues at positions 9 and 10. Open circle: *E. coli* containing the 'IE' ACC deaminase protein variant. Closed circle: *E. coli* containing the 'LA' ACC deaminase protein variant. The curves were generated by plotting the natural log of the OD_{600} values of the variants versus time. The error bars represent the standard errors of the mean from five replicates of growth curve experiments for each strain.

Neither the 9 nor 10 residue is known to be involved in the enzymatic actions of the deaminase. Modeling of the two “winning” ACC deaminase variants at positions 9 and 10 showed that the structure of the IE ACC-DR variant was very similar to the LA variant (data not shown). Furthermore, the predicted structures indicated that the IE or LA residues are located on the outside of the protein structure, away from the active site. Their location may explain their neutral behavior under the selection conditions.

Similarly, I found that residue 11 remained heterogeneous throughout the selection assay. Based on its position in the ACC-DR, I predict that this residue has no direct role in deaminase function. Hence, the heterogeneity maintained at sites 9, 10, and 11 may reflect the neutrality of these residues in the selection assay and in the function of ACC.

Selection at DNA level

Similarly, I followed selection of ACC-DR variants at the DNA level (Figure 2.10). As expected, I found that most variation was in the wobble positions of the codons, and that variation in the first and second positions of codons were fixed quickly after the first round of selection. Reflecting the observations at the protein level, amino acid residues that were highly variable throughout the selection assay displayed the persistent polymorphisms in the first and second codon positions after several rounds of selection.

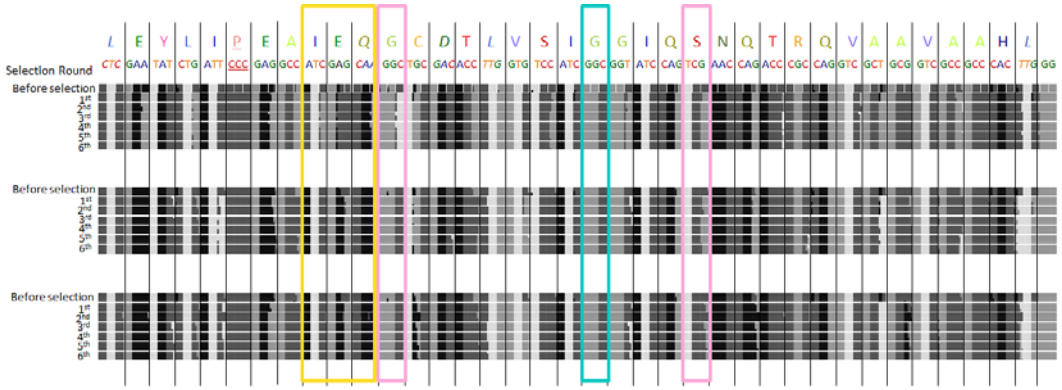

Figure 2.10 DNA base waffle plots for Libraries A, B, and C ACC-DR DNA variant pools at time zero and after each round of selection. The DNA base waffle plots for each Library were made based on the frequency of each base in ACC-DR DNA variant pools at time zero and after each round of selection averaged for the three lineages in each library. Each base is represented by a unique color, and the percentage of grids of the same color shows the frequency of that base at a position. The DNA sequence on top of the waffle plots are color coded, and represent the 'wildtype' DNA sequence in the P. cloacae ACC deaminase region. Bases A, C, G, and T are represented by black, dark grey, light grey, and white, respectively. DNA encoding amino acid residues 4, 9-11, 12 and 26 are shown in the cyan, yellow, and purple rectangles to show examples of fixed, neutral, and essential residues, respectively.

Most essential and important residues, including L4, I5, E7, G12, C13, I22, Q23, S24, Q26, T27, R28, Q29, A34, A35, H36, and L37 (e.g., G12 and S24 in the purple rectangles, Figure 2.10), contained more than one DNA variant for each residue at time zero before selection, and multiple codons encoding the same amino acid residue were fixed in the selection assay in the three libraries, indicating that the selection from nature and my assay acted mainly on the protein level. However, two essential residues, G20 and N25, contained only one dominant DNA variant in all three libraries at time zero before selection, suggesting there is selection from nature on both the protein and the DNA levels. I also observed the enrichment of the codons for residues 9, 10, and 11 (Figure 2.10, the yellow rectangle) by the selection assay.

I also found some rare codons for *E. coli* in the ACC-DR libraries, the most prominent being the codon 'CCC' encoding proline at residue 6, with other examples

such as the codon ‘CTC’ encoding leucine at residue 1, the codon ‘TTG’ encoding leucine at residue 16, and the codon ‘TTG’ encoding leucine at residue 37. These may reflect the soil bacterial origin where the ACC-DR variants were recovered from: for example, the codon ‘CCC’ encoding proline is not a rare codon in *Pseudomonas* or *Burkholderia*.

Comparison with an artificial ACC deaminase region variant pool

To compare the results of the metagenome-derived variants with artificially produced variants, I constructed an artificial ACC-DR variant library generated from doped DNA oligomer synthesis by using one of the winning LA ACC-DR DNA variants as the wildtype backbone, and doping each base with 2.1% non-wildtype nucleotides (see Methods for details). I selected the artificial ACC-DR variant library for six rounds, and sequenced the ACC-DR variants before and after each round of selection as performed for the metagenomic library. Compared to the rhizosphere bacterial ACC-DR variant libraries that started with 1262 unique DNA variants encoding 471 protein variants in total, the artificial ACC-DR variant pool started with 932 unique ACC-DR DNA variant clusters at 99% similarity, which encoded 684 unique ACC-DR protein variants. Thus the artificial library was comprised of a similar number of variants as the rhizosphere bacterial ACC-DR variant library. Using the amino acid waffle plots to track the selection at the amino acid level and the same cut-off values to identify important residues enriched by the selection assay, I found that the L4 residue was important and was fixed after the first round of selection. (Figure 2.11). Similarly, although the H residue competed with Q at the 26th position, Q26 was fixed by the selection. The other fixed and neutral residues observed in the rhizosphere bacterial ACC-DR variant libraries were already the dominant residues in the artificial ACC-DR variant libraries at time zero before the selection assay. Thus, the artificial protein

variant pools yielded the same result as the natural pool.

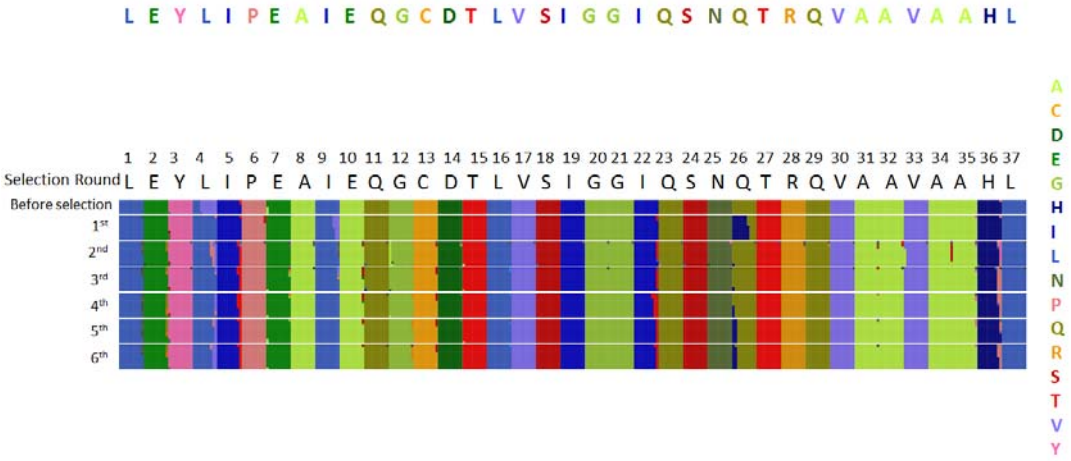

Figure 2.11 Amino acid residue waffle plots for the artificial ACC-DR protein variant pools at time zero and after each round of selection. The amino acid residue waffle plots for the three libraries are shown as in Figure 2.7.

Discussion

Protein structure analysis and optimization has traditionally been an arduous and low-throughput process requiring the generation of purified proteins and point mutation libraries. The discovery of the vast number of uncharacterized gene and protein variants in metagenomes is driving demand for high throughput assays. Here I utilized a metagenomic library and a growth-based selection assay in order to understand the protein sequence-function relationships of ACC-DR (Figure 2.12) and identify the optimal ACC-DR variants in *E. coli*.

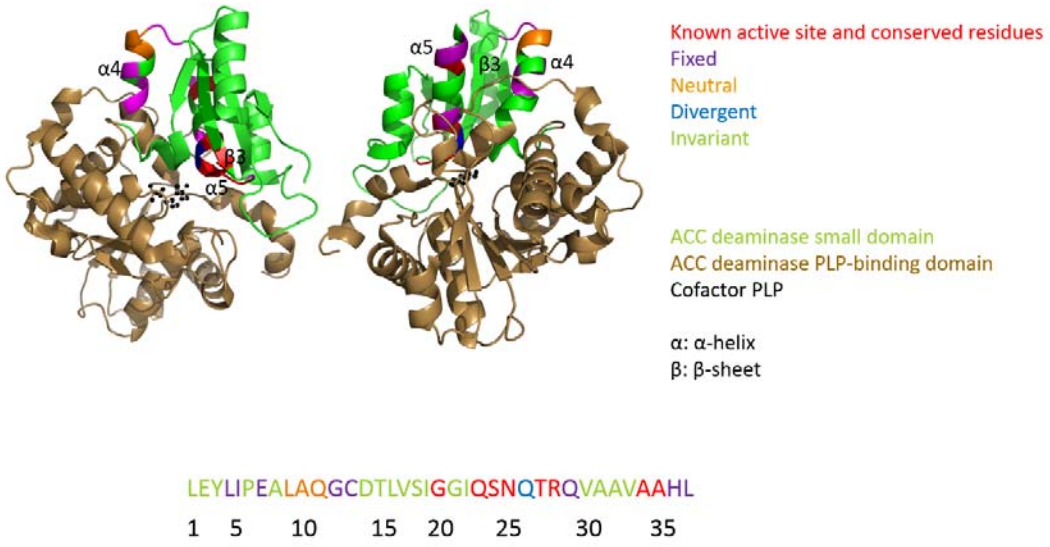

Figure 2.12 Summary of the selection assay results. The ACC deaminase monomer is shown in copper with a back view (left) and front view (right) of the ACC-DR. The essential, fixed, neutral, and divergent residues identified by the selection assay are colored red, purple, orange, and blue, respectively, and the invariant residues in ACC-DR are shown in green. The linear amino acid sequence of ACC-DR with the above-mentioned five types of residues colored accordingly is placed below the protein structure.

In my assay, the first round of selection had the greatest impact on the diversity within the variant pools. These results suggest that other metagenomic protein

variants could be optimized in very few selection cycles. By utilizing the soil metagenome as the initial source of protein variation, I was able to begin my selection assay with most essential residues already fixed due to the purifying selection present in the soil metagenome. Thus, my assay avoided the large sequence space of non-functional protein variants.

Within the sequence space of functional ACC-DR variants, my assay was able to identify the functional, divergent, and neutral regions of the ACC-DR. Of particular interest, my assay revealed novel regions, or novel combination of certain amino acid residues, of the ACC-DR that may be critical to the optimal efficiency of ACC deaminase: the assay enriched for specific residues within the previously structurally identified but less studied small domain of ACC deaminase. The statistical coupling analysis on the enriched residues indicated that these enriched residues may collectively play an important role to enhance the deaminase efficiency. Additionally, the assay uncovered diversification at Q26, an essential residue within the ACC-DR. While the selection assay favored one of two dominant residues, Q or H, at position 26, the presence of this alternative residue in nature suggests that other selective pressures such as the need for flexibility for alternate substrates, or the need to co-evolve with other residues in the protein may be driving diversification at this residue. Finally, my selection assay found a cluster of several residues (positions 9, 10, 11) that were relatively neutral with respect to ACC deaminase function. The ACC-DR artificial protein variant pools yielded similar results as the natural pool, supporting the use of a metagenomic variant pool for mutational analysis and protein optimization.

Overall, this work shows that the generation of protein variant pools from the soil metagenome is able to provide a detailed sketch of the functional regions of a protein domain and thus a starting point for understanding protein structure or optimizing

enzyme performance. Compared to the generation of protein variant pools from artificial libraries, my method is relatively cost-effective, and it focuses on the natural and functional protein variant sequence space. The growth-based selection assay is straightforward and is readily adaptable to other enzymes and expression hosts. Hence, the use of metagenomic libraries in a competition assay has the potential to speed the translation of novel natural products from nature to industry.

REFERENCES

Arnold K, Bordoli L, Kopp J, Schwede T (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22:** 195-201.

Blaha D, Prigent-Combaret C, Mirza MS, Moenne-Loccoz Y (2006). Phylogeny of the 1-aminocyclopropane-1-carboxylic acid deaminase-encoding gene acdS in phytobeneficial and pathogenic Proteobacteria and relation with strain biogeography. *FEMS microbiology ecology* **56:** 455-470.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7:** 335-336.

Dantas G, Sommer MO, Degnan PH, Goodman AL (2013). Experimental approaches for defining functional roles of microbes in the human gut. *Annual review of microbiology* **67:** 459-475.

DeLano W (2002). The PyMOL Molecular Graphics System, Version 1.2 r3pre, Schrödinger, LLC.

Dworkin M, Foster JW (1958). Experiments with Some Microorganisms Which Utilize Ethane and Hydrogen. *J Bacteriol* **75:** 592-603.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and

high throughput. *Nucleic Acids Res* **32:** 1792-1797.

Edgar RC (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26:** 2460-2461.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194-2200.

Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D *et al* (2010). High-resolution mapping of protein sequence-function relationships. *Nature methods* **7:** 741-746.

Fujino A, Ose T, Yao M, Tokiwano T, Honma M, Watanabe N *et al* (2004). Structural and enzymatic properties of 1-aminocyclopropane-1-carboxylate deaminase homologue from Pyrococcus horikoshii. *J Mol Biol* **341:** 999-1013.

Glick BR (2004). Bacterial ACC deaminase and the alleviation of plant stress. *Adv Appl Microbiol* **56:** 291-312.

Glick BR, Todorovic B, Czarny J, Cheng Z, Duan J, McConkey B (2007). Promotion of plant growth by bacterial ACC deaminase. *Critical Reviews in Plant Sciences* **26:** 227-242.

Glick BR, Stearns JC (2011). Making Phytoremediation Work Better: Maximizing a Plant's Growth Potential in the Midst of Adversity. *Int J Phytoremediat* **13:** 4-16.

Karthikeyan S, Zhou QX, Zhao ZB, Kao CL, Tao ZH, Robinson H *et al* (2004). Structural analysis of *Pseudomonas* 1-aminocyclopropane-1-carboxylate deaminase complexes: Insight into the mechanism of a unique pyridoxal-5 '-phosphate dependent cyclopropane ring-opening reaction. *Biochemistry-Us* **43:** 13328-13339.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H *et al* (2007). Clustal W and clustal X version 2.0. *Bioinformatics* **23:** 2947-2948.

Li J, Glick BR (2001). Transcriptional regulation of the Enterobacter cloacae UW4 1-aminocyclopropane-1-carboxylate (ACC) deaminase gene (acdS). *Canadian journal of microbiology* **47:** 359-367.

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

McGarvey KM, Queitsch K, Fields S (2012). Wide variation in antibiotic resistance proteins identified by functional metagenomic screening of a soil DNA library. *Applied and environmental microbiology* **78:** 1708-1714.

Nacke H, Engelhaupt M, Brady S, Fischer C, Tautzt J, Daniel R (2012). Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. *Biotechnology letters* **34:** 663-675.

Nei M, Gojobori T (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution* **3:** 418-426.

Onofre-Lemus J, Hernandez-Lucas I, Girard L, Caballero-Mellado J (2009). ACC (1-aminocyclopropane-1-carboxylate) deaminase activity, a widespread trait in Burkholderia species, and its growth-promoting effect on tomato plants. *Applied and environmental microbiology* **75:** 6581-6590.

Ose T, Fujino A, Yao M, Watanabe N, Honma M, Tanaka I (2003). Reaction intermediate structures of 1-aminocyclopropane-1-carboxylate deaminase - Insight into PLP-dependent cyclopropane ring-opening reaction. *J Biol Chem* **278:** 41069-41076.

Peiffer JA, Spor A, Jin Z, Koren O, Tringe SG, Dangl JL *et al* (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Manuscript in preparation*.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing: Vienna, Austria.

Rice P, Longden I, Bleasby A (2000). EMBOSS: The European molecular biology open software suite. *Trends Genet* **16:** 276-277.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF *et al* (2013).

Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499:** 431-437.

Sheehy RE, Honma M, Yamada M, Sasaki T, Martineau B, Hiatt WR (1991). Isolation, Sequence, and Expression in Escherichia-Coli of the *Pseudomonas* Sp Strain Acp Gene Encoding 1-Aminocyclopropane-1-Carboxylate Deaminase. *J Bacteriol* **173:** 5260-5265.

Suyama M, Torrents D, Bork P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34:** W609-W612.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585-595.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* **30:** 2725-2729.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25:** 1189-1191.

Wickham H (2007). Reshaping data with the reshape package. *J Stat Softw* **21:** 1-20.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer Publishing Company, Incorporated.

Wickham H (2011). The Split-Apply-Combine Strategy for Data Analysis. *J Stat Softw* **40:** 1-29.

Yao M, Ose T, Sugimoto H, Horiuchi A, Nakagawa A, Wakatsuki S *et al* (2000). Crystal structure of 1-aminocyclopropane-1-carboxylate deaminase from Hartsenula saturnus. *J Biol Chem* **275:** 34557-34565.

Zhang J, Rosenberg HF, Nei M (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proceedings of the National Academy of Sciences of the United States of America* **95:** 3708-3713.

Chapter 3

Maize Genotypes are Significantly Associated with Metabolic Genes of Rhizosphere Pseudomonas Isolates

Zhao Jin, Tijana Glavina Del Rio, Tanja Woyke, Susannah G. Tringe, Wei Zhang, Jeff Dangl, Ruth E. Ley

Abstract

Plants and their roots-associated microorganisms interact closely with each other. However, relatively little is known about the impact of plant genetic variation on its rhizosphere bacterial populations, especially whether plants select for taxonomic or functional components of bacteria within a plant species. Here, I sequenced the genomes of 48 *Pseudomonas* isolates from the rhizosphere of two maize genotypes, a sweet corn inbred line and a non-stiff stalk maize inbred line, grown in two field conditions at the same developmental stage. I observed a small but significant association of maize genotypes with the variation in the metabolic genes of these *Pseudomonas* isolates after controlling for the effects from *Pseudomonas* isolate taxonomy and field conditions, while I did not see a significant association of maize genotypes with the variation in the abundance of the OTUs containing the *Pseudomonas* isolates. I identified the corresponding enrichment of metabolic genes in *Pseudomonas* isolates with respect to each maize genotype, including enriched denitrification-related and sugar metabolic genes in *Pseudomonas* isolates from the rhizosphere of the non-stiff stalk and the sweet corn maize inbred lines, respectively. I conducted molecular evolution analyses on the enriched metabolic genes and observed sites under negative selection. I also identified co-occurring OTUs from the same maize rhizosphere where the *Pseudomonas* isolates were cultured; these co-occurring OTUs may be involved in various cooperative activities such as nitrogen fixation and cell-cell communication with the *Pseudomonas* isolates in the maize rhizosphere. These results should facilitate future studies to locate regions of bacterial genomes that are directly controlled by plant genotypes and are involved in plant-microbe interactions, which will ultimately benefit crop breeding.

Introduction

Plants and their roots-associated microorganisms are considered as "superorganisms" (Mendes et al 2011), in which plants interact closely with microbes. Plants release up to 40% of their photosynthates to the rhizosphere, mainly in the form of root exudates (Singh et al 2004), which feed the roots-associated microbes with carbon and energy sources. In turn, microbes decompose soil organic matter, providing essential nutrients to plants (Lugtenberg and Kamilova 2009). In addition, roots-associated microbes also benefit plants in a number of ways, such as protecting plants from infection by soil-borne pathogens (Garbeva et al 2004, Mendes et al 2011), fixing nitrogen (Brencic and Winans 2005, Hsu and Buckley 2009), and promoting root growth by producing phytohormones (Mavrodi et al 2006). Much as plants and their microbes rely on each other, the impact of plant genetic variation on its rhizosphere bacterial populations remains poorly understood.

It is also unclear whether plants select for taxonomic or functional bacterial populations in their rhizosphere. While it is well known that components in plant root exudates attract bacteria to colonize roots (Zhang et al 2014), some studies indicated that different plant species or artificial root exudates mimicking natural maize root exudates selectively assembled rhizosphere bacterial communities of varying taxonomic compositions (Baudoin et al 2003, Grayston et al 1998). Others showed that plants selected for functional bacterial populations (Briones Jr et al 2003, Martinez-Romero 2009). Using a well-balanced study design, I addressed the above-mentioned question from studying the relationships between one plant species and one bacterial genus.

I focused on the impact of maize genotypes on the maize roots-associated *Pseudomonas* populations. Maize is one of the staple food crops in the world, and

harbors extensive natural diversity and tractable genotypic and phenotypic information (McMullen et al 2009). Bacteria in the *Pseudomonas* genus are closely associated with plants, and play a number of important roles such as acquisition of important elements including iron (Philippot et al 2013) and phosphorous (Rodriguez et al 2006), suppression of plant diseases (Raaijmakers and Weller 1998), and induction of plant systemic resistance (De Vleesschauwer and Höfte 2009). The study of maize genetic control on its rhizosphere *Pseudomonas* populations will advance our understanding on what components of the rhizosphere microbiome (i.e. bacterial taxa or traits) plant genetic variation controls.

To investigate the influence of maize genetic variation on its rhizosphere *Pseudomonas* populations, I employed the following study design (Figure 3.1). Two maize genotypes, Mo17, a non-stiff stalk maize inbred line, and Il14h, a sweet corn inbred line (Flint-Garcia et al 2005), were grown in multiple replicates at two different fields located at New York and Illinois as describe previously (Peiffer et al 2013). I cultured the *Pseudomonas* isolates using the rhizosphere soil samples from three replicate plants of each maize genotype grown in each field, and sequenced the genomes of the first four *Pseudomonas* isolates cultured from each rhizosphere soil sample. From a total of 48 *Pseudomonas* isolate genomes, I aimed to find out whether maize genotypes are significantly associated with the variation in the *Pseudomonas* isolate genomes, and whether maize genotypes select the taxonomy or function of the isolates.

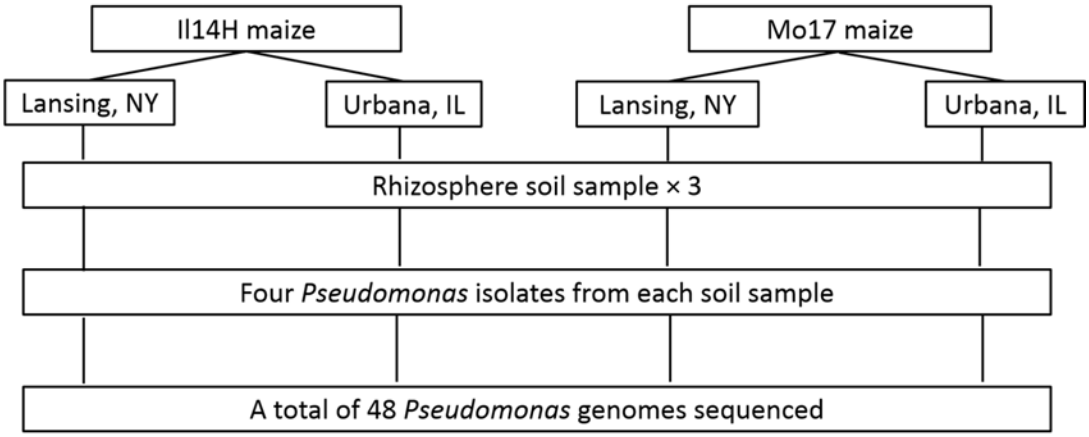

Figure 3.1 Experimental design. Multiple replicates of two maize inbred lines, Il14h, a sweet corn inbred line, and Mo17, a non-stiff stalk corn inbred line, were grown in two fields (Lansing, NY and Urbana, IL). The rhizosphere soil samples from three different maize plants of each maize genotype grown in each field, i.e. a total of 12 soil rhizosphere soil samples were collected to culture the *Pseudomonas* isolates. Four *Pseudomonas* isolates were cultured from each rhizosphere soil sample, so a total of 48 *Pseudomonas* isolate genomes were sequenced.

I found that maize genotypes are not significantly associated with the variation in the relative abundance of *Pseudomonas* isolates. I observed a small but significant association between maize genotypes and the variation in the counts of the metabolic genes from these *Pseudomonas* isolates after controlling for the effects from *Pseudomonas* isolate taxonomy and field conditions. I identified the corresponding enrichment of metabolic genes in *Pseudomonas* isolates with respect to each maize genotype: *Pseudomonas* isolate genomes from the non-stiff stalk maize harbor increased denitrification-related genes, whereas *Pseudomonas* isolates from the sweet corn maize rhizosphere contain more sugar metabolic genes. I conducted molecular evolution analyses in the enriched metabolic genes: while I did not see any site under positive selection, I observed negative selection in some sites in several genes. I also identified co-occurring OTUs from the same maize rhizosphere where the *Pseudomonas* isolates were cultured. These co-occurring OTUs may be involved

in various cooperative activities such as nitrogen fixation and cell-cell communication with the *Pseudomonas* isolates in the maize rhizosphere. Thus, I have shown that the genetic variation from two maize genotypes grown in two different fields at the same developmental stage significantly influence the metabolic abilities of their rhizosphere *Pseudomonas* isolates, providing evidence that maize genotypes select for the functional component of the rhizosphere microbiome.

Methods

Study design

This study was aimed at determining the effect from maize genetic variation across different fields on *Pseudomonas* isolates. The design of this study mimicked that of a two-way ANOVA with two different maize genotypes and two different field conditions. The two maize genotypes, Il14h, a sweet corn inbred line, and Mo17, a non-stiff stalk maize inbred line, were grown in multiple replicates at two different fields located at New York state and Illinois. The rhizosphere soil samples from three replicates of each maize genotype grown at each field at week 12 after planting were collected as previously described (Peiffer et al 2013) by a team of people in my lab. Hence, a total of 12 rhizosphere soil samples were used in this study to culture *Pseudomonas* isolates. The genomes of the first four *Pseudomonas* isolates cultured from each rhizosphere soil sample were sequenced. Thus, a total of 48 *Pseudomonas* isolate genomes were sequenced.

Isolation, growth, and genomic DNA extraction of *Pseudomonas* isolates

From each rhizosphere soil sample, 0.1 grams of soils were washed in 5 mL sterile phosphate buffered saline with 10% glycerol for 1 hour with gentle rocking at room temperature. 100 µL of the wash liquid was plated onto *Pseudomonas* Isolation

Agar (BD Diagnostic Systems, Franklin Lakes, NJ) using a disposable inoculating loop. The plates were incubated at 30 °C until colonies formed. To extract genomic DNA, single colonies were inoculated into 5 mL Lysogeny broth (LB), and grown at 30 °C overnight. The cultures were harvested by centrifugation at 5000 × g for 5 min, and the cells were lysed using the B1 and B2 solutions as described in the Qiagen Genomic DNA Handbook (Qiagen, Valencia, CA). The genomic DNA was precipitated with ethanol and sodium acetate, and pelleted after a centrifugation at 1811 × g for 30 min. PCR and Sanger sequencing of the 16S rRNA genes were used to confirm the identity and purity of the genomic DNA preparations.

Genome sequencing, assembly, annotation, and functional profile

The genomes of *Pseudomonas* isolates were sequenced at the Joint Genome Institute using Illumina technology (Bentley et al 2008). An Illumina standard shotgun library was constructed and sequenced on the Illumina HiSeq 2000 platform. All general aspects of library construction and sequencing are provided on the Joint Genome Institute (JGI) website (http://www.jgi.doe.gov/sequencing/protocols/index.html). The procedures for genome assembly and annotation are similar to those described previously (Reeve et al 2014). For details of genome statistics, see Tables 3.1 and 3.2. The MetaCyc functional profile for the 48 *Pseudomonas* isolates, which enumerated the primary and secondary metabolic pathways from all isolates, was generated on the JGI IMG/ER portal (Markowitz et al 2012).

Table 3.1 Genome statistics for the 24 *Pseudomonas* isolates from the Il14h maize genotype.

	Il14h.Lansing.I1	Il14h.Lansing.I2	Il14h.Lansing.I3	Il14h.Lansing.I4	Il14h.Lansing.I5	Il14h.Lansing.I6
Total Bases	5665646	6436821	5666577	4756206	5589422	5597776
# Genes	5243	5958	5251	4370	5208	5192
GC%	62.07	60.51	62.06	61.59	62.15	62.14
# Scaffolds	24	20	23	52	54	25
	Il14h.Lansing.I7	Il14h.Lansing.I8	Il14h.Lansing.I9	Il14h.Lansing.I10	Il14h.Lansing.I11	Il14h.Lansing.I12
Total Bases	5953767	6102725	5681488	6099553	6100002	5575902
# Genes	5353	5536	5269	5533	5536	5163
GC%	62.45	62.38	62.07	62.41	62.4	62.14
# Scaffolds	28	29	25	33	33	26
	Il14h.Urbana.I1	Il14h.Urbana.I2	Il14h.Urbana.I3	Il14h.Urbana.I4	Il14h.Urbana.I5	Il14h.Urbana.I6
Total Bases	4749093	4717612	4718477	5817007	5723734	6337748
# Genes	4355	4418	4397	5462	5374	5792
GC%	61.58	61.71	61.73	62.03	62.14	59.99
# Scaffolds	52	93	43	29	24	61
	Il14h.Urbana.I7	Il14h.Urbana.I8	Il14h.Urbana.I9	Il14h.Urbana.I10	Il14h.Urbana.I11	Il14h.Urbana.I12
Total Bases	6883327	5822697	4847117	4847839	4850394	4848195
# Genes	6259	5472	4527	4522	4524	4517
GC%	60.72	62.02	62.75	62.76	62.76	62.76
# Scaffolds	78	27	32	24	23	22

Table 3.2 Genome statistics for the 24 *Pseudomonas* isolates from the Mo17 maize genotype.

	Mo17.Lansing.I1	Mo17.Lansing.I2	Mo17.Lansing.I3	Mo17.Lansing.I4	Mo17.Lansing.I5	Mo17.Lansing.I6
Total Bases	6932332	6928939	4692529	5593635	6103916	6092576
# Genes	6412	6409	4370	5178	5575	5563
GC%	61.03	61.02	61.84	62.14	60.3	60.33
# Scaffolds	34	34	37	28	25	28
	Mo17.Lansing.I7	Mo17.Lansing.I8	Mo17.Lansing.I9	Mo17.Lansing.I10	Mo17.Lansing.I11	Mo17.Lansing.I12
Total Bases	5564386	5565480	7022592	7014050	5560711	5564094
# Genes	5185	5184	6437	6415	5182	5184
GC%	62.23	62.21	62.51	62.53	62.24	62.23
# Scaffolds	24	25	99	99	25	27
	Mo17.Urbana.I1	Mo17.Urbana.I2	Mo17.Urbana.I3	Mo17.Urbana.I4	Mo17.Urbana.I5	Mo17.Urbana.I6
Total Bases	4954515	5283927	4952268	4910367	5759395	6454480
# Genes	4594	4863	4630	4532	5370	5967
GC%	63.59	63.11	63.59	63.37	62.73	59.41
# Scaffolds	15	29	17	29	49	43
	Mo17.Urbana.I7	Mo17.Urbana.I8	Mo17.Urbana.I9	Mo17.Urbana.I10	Mo17.Urbana.I11	Mo17.Urbana.I12
Total Bases	5753615	5752811	4698894	4702100	4693207	6568431
# Genes	5363	5360	4379	4376	4371	6164
GC%	62.76	62.76	61.8	61.77	61.83	59.06
# Scaffolds	41	43	26	30	33	32

16S tree and concatenated ribosomal protein phylogeny

16S rRNA gene sequences for the 48 *Pseudomonas* isolates were retrieved using a 16S rRNA gene mining program developed at JGI (Han, J., unpublished). The 16S rRNA genes of the 48 *Pseudomonas* isolates were aligned in PyNAST using the

alignment of *Pseudomonas* OTUs downloaded from Greengenes (DeSantis et al 2006) as the template. The alignment of the isolates and two representative *Pseudomonas* stains from each Greengenes *Pseudomonas* OTU was filtered by QIIME (Caporaso et al 2010), and used for Phylogeny construction using maximum likelihood implemented in PhyML (Guindon et al 2010) with the GTR+γ+I model of evolution and 100 bootstrap resampling. The bootstrap consensus 16S rRNA gene tree was visualized using the interactive Tree of Life (iTOL) (Letunic and Bork 2011), and the bootstrap values greater than 60 were displayed.

Comparative genomics and Pan-genome SNP analysis

The SNPs of the 48 *Pseudomonas* isolates were determined using the wombac program developed by the Victorian Bioinformatics Consortium that aligns bacterial genomes with a reference genome based on bwa and samtools to identify SNPs (http://www.vicbioinformatics.com/software.wombac.shtml). The wombac output file that aligned all the substitution SNPs in the *Pseudomonas* isolates was used to generate a binary table for the presence or absence of each SNP in each isolate genome. The PlasmidFinder-1.2 Server (http://cge.cbs.dtu.dk/services/PlasmidFinder/) was used to confirm that no plasmid sequence was present in the isolate genomes. SNPs were categorized as synonymous or non-synonymous using SnpEff version 3.5 (Cingolani et al 2012).

Abundance of isolates in amplicon sequencing data

Operational taxonomic units (OTUs) for the rhizosphere microbiome 16S rRNA gene V4 region amplicon Illumina MiSeq sequencing data (Jin, Z and Ley, RE., unpublished data) on rhizosphere soil samples collected from 27 maize inbred lines (including Il14h and Mo17) grown at Lansing and Urbana at week 12 after planting were picked using a closed-reference procedure against the May 2013 Greengenes

database at 97% sequence identity in QIIME. The 16S rRNA gene sequences of the 48 *Pseudomonas* isolates were trimmed to the length of the V4 region, and were used to search against the sequences of the OTUs from the above-mentioned 16S rRNA gene V4 region amplicon sequencing data. This search identified what OTUs the *Pseudomonas* isolates belonged to. The number of reads for the *Pseudomonas* isolates-containing OTUs represented their absolute abundance. Due to the uneven numbers of reads for the rhizosphere soil samples containing the *Pseudomonas* isolates OTUs, as well as the unequal total number of reads for each sequencing run containing the rhizosphere soil samples, the absolute abundance of each *Pseudomonas* isolates-containing OTU was normalized as the following: the number of reads for each rhizosphere soil sample containing the *Pseudomonas* isolates OTUs was divided by the total number of reads from the sequencing run containing that rhizosphere soil sample. This ratio was multiplied to the absolute abundance of each *Pseudomonas* isolates-containing OTU to calculate the relative abundance.

Distance-based approaches

The distance-based approaches to identify maize genotype effect were conducted in R 3.0.2 (R Development Core Team 2005) using the 'vegan' package 2.0-9 (Oksanen et al 2013). The 'betadisper' function was first used to check whether the two groups compared had similar multivariate dispersions, so that the assumptions for PERMANOVA test were satisfied. Then the 'adonis' function, which conducted a permutational multivariate analysis of variance on distance matrices, was employed to run PERMANOVA on the Bray-Curtis dissimilarities (Beals 1984) of the pan-genome SNPs counts, the relative abundance of *Pseudomonas* isolates-containing OTUs, or the MetaCyc metabolic pathway profiles. The MetaCyc metabolic pathways that were significantly associated with each maize genotype were identified

using the ‘multipatt’ function in the R ‘indicspecies’ package (Cáceres and Legendre 2009).

Molecular Evolution Analyses

The metabolic genes used for evolutionary analyses are listed in Supplement Table 1. The protein sequences for each gene set were aligned using MUSCLE (Edgar 2004), and were used to align the corresponding DNA sequences using PAL2NAL (Suyama et al 2006). The multiple sequence alignment of each gene set were manually curated before they were provided to jModeltest 2 (Darriba et al 2012) to estimate the best evolution models for phylogeny inference by PhyML. The multiple sequence alignment of each gene set and their phylogenetic tree were supplied to HyPhy (Pond and Muse 2005), and the methods of QuickSelectionDetection.bf, BivariateCodonRateAnalysis.bf, and BranchSiteREL.bf were used to infer natural selection in the genes, as well as the methods of SingleBreakpointRecomb.bf and GARDProcessor.bf to identify any site with recombination. The pairwise genetic distance for the sequences in each gene set was calculated in MEGA6 (Tamura et al 2013).

Co-occurring OTUs and networks

The OTUs tables for the above-mentioned rhizosphere microbiome 16S rRNA gene V4 region amplicon sequencing data were normalized with frequency: in each un-normalized OTU table, the number of reads for each OTU in a given rhizosphere soil sample was divided by the total number of reads for that sample. The frequency-normalized OTU-tables were used to identify co-occurring OTUs with the *Pseudomonas* isolates-containing OTUs. OTUs present in fewer than 3 samples were removed; Pearson correlation coefficients of the OTUs and the p-values of the correlations were calculated using the R Bioconductor (Gentleman et al 2004)

'gpgraph' package (Castelo and Roverato 2006), and the Benjamini and Hochberg multiple correction method (Benjamini and Hochberg 1995) was applied to calculate the false discovery rates. OTUs that had a correlation coefficient over 0.4 with the *Pseudomonas* isolates-containing OTUs and a correlation q-value lower than 0.01 were considered as a true co-occurring OTU. The correlation coefficients between the identified co-occurring OTUs and the *Pseudomonas* isolates-containing OTUs were imported into Cytoscape (Shannon et al 2003) to generate network graphs.

Results and Discussion

16S Phylogeny of *Pseudomonas* isolates

To infer the phylogeny of the 48 *Pseudomonas* isolates, I constructed a bootstrapped maximum-likelihood 16S rRNA gene phylogenetic tree of the 48 *Pseudomonas* isolates using the 16S rRNA genes of two representative strains from each *Pseudomonas* OTU in the Greengenes database, and *P. stutzeri str. SWI26* as the outgroup. The phylogenetic tree shows that four *Pseudomonas* isolates from the Mo17 maize inbred line grown in Urbana, IL form a clade, clade I, which is distant from the other big clade, clade II, which contains the remaining 44 isolates (Figure 3.2). Within the big clade, 29 *Pseudomonas* isolates group within one big cluster, whereas the other 15 isolates form five smaller clusters (marked by letters). Within the big cluster, cluster A, the *Pseudomonas* isolates do not group by maize genotypes or field conditions; This distribution suggests that the standard cultivation procedure used in this study favor the isolation of close *Pseudomonas* species from rhizosphere soil samples of two maize genotypes grown in two different fields, whereas the small bootstrap values for branching patterns within cluster A also suggest that these *Pseudomonas* isolates have very similar 16S rRNA genes. Among

the other five smaller clusters of *Pseudomonas* isolates, clusters C, D, and F are composed of isolates from the same maize genotype and field, whereas cluster E contains isolates from two maize genotypes of the same field. Overall, there is no consistent clustering of *Pseudomonas* isolates by maize genotypes or by fields. The genus *Pseudomonas* has been divided into two intragenic clusters, 'IGC *P. aeruginosa*' and 'IGC *P. fluorescens*', as suggested by analyses of 16S rRNA and housekeeping gene sequences from over a hundred *Pseudomonas* species (Kampfer and Glaeser 2012). Based on the 16S rRNA gene phylogeny, the 4 *Pseudomonas* isolates in clade I belong to the IGC *P. aeruginosa* intragenic cluster, while the remaining 44 isolates in clade II belong to the IGC *P. fluorescens* intragenic cluster.

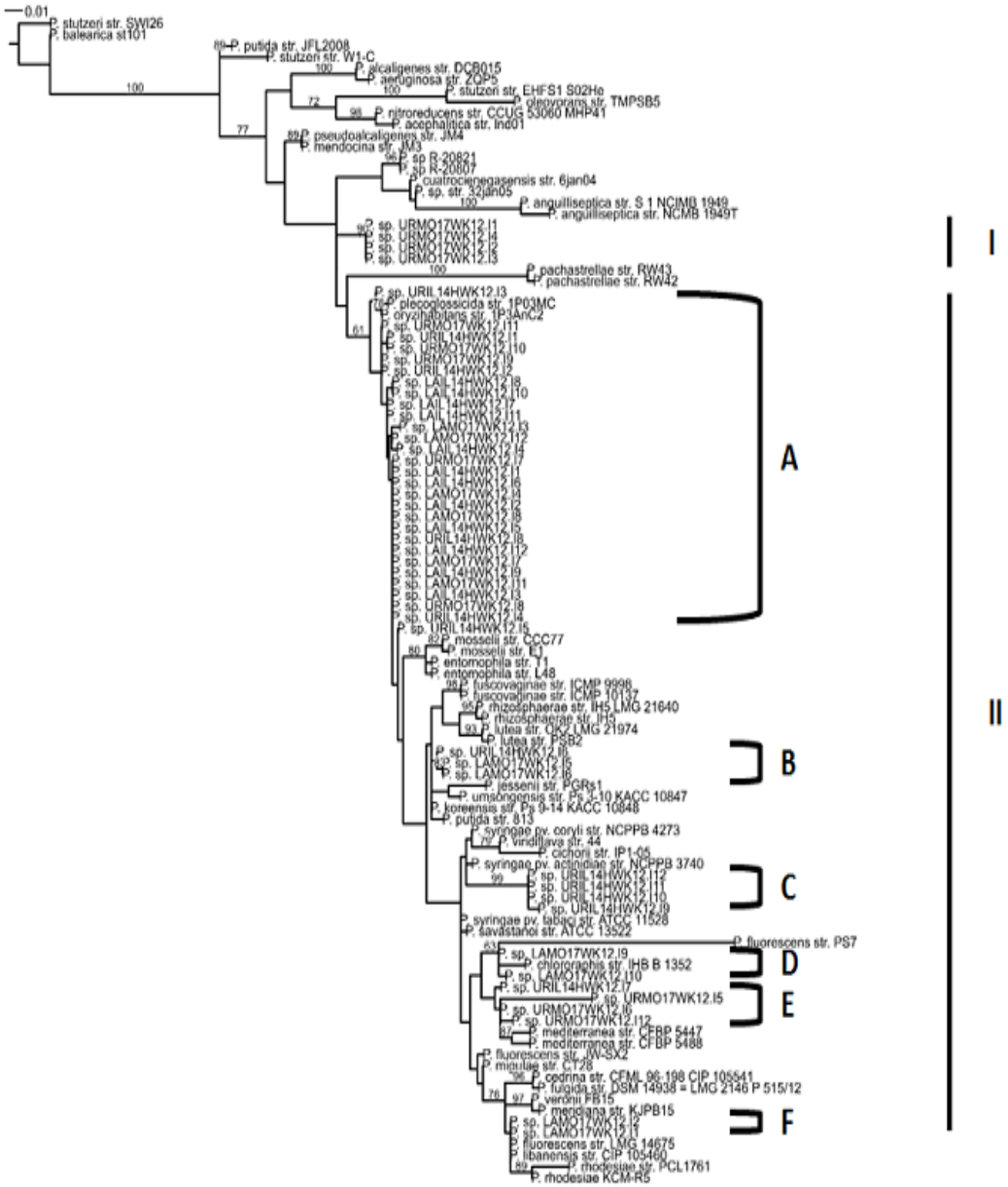

Figure 3.2 16S rRNA gene phylogeny of 48 *Pseudomonas* isolates. The phylogenetic tree was built using the 16S rRNA gene sequences from the 48 *Pseudomonas* isolates and two representative sequences from each *Pseudomonas* OTU in the Greengenes May 2013 database (DeSantis et al 2006). Bootstrap values greater than 60% are displayed on the branches. *P. stutzeri str. SWI26* was used as the outgroup. Clade I includes the four *Pseudomonas* isolates from the Mo17 maize inbred line grown in Urbana, IL. Clade II includes the remaining 44 *Pseudomonas* isolates, with clusters A – F representing the clusters formed by the 44 isolates within Clade II.

Abundance of Pan-genome SNPs is not significantly associated with maize genotypes

To further investigate the genetic diversity of the *Pseudomonas* isolates, I studied the pan-genome single nucleotide polymorphisms (SNPs) of the isolates. I first confirmed that the 48 *Pseudomonas* isolate genomes do not contain any plasmid sequence. Using the complete genome of *P. entomophila str L48* as the reference genome, I identified 69350 pan-genome SNPs shared by all 48 *Pseudomonas* isolates, with any two isolates sharing 61.07 ± 0.9820% of their SNPs on average. All 48 *Pseudomonas* isolate genomes harbor 6480429 synonymous pan-genome SNPs and 1647943 non-synonymous pan-genome SNPs in total, and 135008 ± 50821 synonymous pan-genome SNPs and 34332 ± 9697 non-synonymous pan-genome SNPs per isolate genome on average.

To identify whether maize genotypes influence the pan-genome SNP abundance of their rhizosphere *Pseudomonas* isolates, I tested the statistical association of maize genotypes with the differences in the abundance of *Pseudomonas* isolate pan-genome SNPs. I generated a binary table based on the presence or absence of each SNP in each *Pseudomonas* isolate, and conducted the permutational multivariate analysis of variance (PERMANOVA) (Anderson 2001) on the Bray-Curtis dissimilarity of the SNP table. I found that fields do not significantly influence the pan-genome SNP abundance. After controlling for field conditions, I did not observe a significant association between maize genotypes and the differences in the abundance of *Pseudomonas* isolate pan-genome SNPs ($P > 0.05$). While I did not observe a significant association between maize genotypes and the abundance of the pan-genome SNPs in the *Pseudomonas* isolates, polymorphic sites in genes involved in maize-*Pseudomonas* interactions are more likely to reflect influence and selection

from maize genotypes.

Abundance of *Pseudomonas* isolates-containing OTUs is not significantly associated with maize genotypes

To examine whether maize genotypes influence the abundance of the OTUs containing the *Pseudomonas* isolates in the rhizosphere soil samples, I tested the statistical association of maize genotypes and the abundance of the OTUs containing the *Pseudomonas* isolates from the maize rhizosphere. I searched the maize rhizosphere microbiome 16S rRNA gene Illumina amplicon data (Jin, Z and Ley, R. E., unpublished data) using the V4 region of the 16S genes from the *Pseudomonas* isolates to find the OTUs containing the *Pseudomonas* isolates. I identified eight OTUs that the 48 *Pseudomonas* isolates belong to (Table 3.3), which are all members of the *Pseudomonas* genus (not shown due to table size limit). Because the microbiomes of the 12 rhizosphere soil samples used to culture the *Pseudomonas* isolates were sequenced in three different Illumina runs, I normalized the absolute abundance of the OTUs containing the *Pseudomonas* isolates with the number of reads per sample and the total number of reads per Illumina run. PERMANOVA on the Bray-Curtis dissimilarity for the relative abundance of the OTUs containing the *Pseudomonas* isolates show that fields are significantly associated with the variation in the abundance of the OTUs containing the *Pseudomonas* isolates ($P < 0.05$), and contribute to 44.735% of the total variation. After controlling for the effect from field conditions, maize genotypes are not significantly associated with the abundance of the OTUs containing the *Pseudomonas* isolates.

Table 3.3 The eight OTUs that contain the 48 *Pseudomonas* isolates. The OTUs were identified by searching the maize rhizosphere microbiome 16S rRNA gene Illumina amplicon data (Jin, Z and Ley, R. E., unpublished data) using the V4 region of the 16S genes from the *Pseudomonas* isolates. Columns 1 and 3 show the names of the *Pseudomonas* isolates, columns 2 and 4 show the OTU IDs that contain the *Pseudomonas* isolates. The OTU IDs are Greengenes OTU numbers.

Isolate names	OTUs isolates belong to	Isolate names	OTUs isolates belong to
Il14h.Lansing:I1	845178	Il14h.Urbana:I1	845178
Il14h.Lansing:I2	845178	Il14h.Urbana:I2	845178
Il14h.Lansing:I3	845178	Il14h.Urbana:I3	845178
Il14h.Lansing:I4	845178	Il14h.Urbana:I4	845178
Il14h.Lansing:I5	845178	Il14h.Urbana:I5	845178
Il14h.Lansing:I6	845178	Il14h.Urbana:I6	4456889
Il14h.Lansing:I7	845178	Il14h.Urbana:I7	1109251
Il14h.Lansing:I8	817734	Il14h.Urbana:I8	845178
Il14h.Lansing:I9	845178	Il14h.Urbana:I9	817209
Il14h.Lansing:I10	817734	Il14h.Urbana:I10	817209
Il14h.Lansing:I11	845178	Il14h.Urbana:I11	817209
Il14h.Lansing:I12	845178	Il14h.Urbana:I12	817209
Mo17.Lansing:I1	4435982	Mo17.Urbana:I1	845178
Mo17.Lansing:I2	4435982	Mo17.Urbana:I2	845178
Mo17.Lansing:I3	845178	Mo17.Urbana:I3	845178
Mo17.Lansing:I4	845178	Mo17.Urbana:I4	845178
Mo17.Lansing:I5	4456889	Mo17.Urbana:I5	1109251
Mo17.Lansing:I6	4456889	Mo17.Urbana:I6	1109251
Mo17.Lansing:I7	845178	Mo17.Urbana:I7	845178
Mo17.Lansing:I8	845178	Mo17.Urbana:I8	845178
Mo17.Lansing:I9	4451011	Mo17.Urbana:I9	845178
Mo17.Lansing:I10	4451011	Mo17.Urbana:I10	845178
Mo17.Lansing:I11	845178	Mo17.Urbana:I11	845178
Mo17.Lansing:I12	845178	Mo17.Urbana:I12	557974

A number of studies have shown that field conditions, such as soil moisture, pH values, and temperature could influence the abundance of certain plant growth promoting bacteria taxa (Gaiero et al 2013). We have also found previously that the physiochemical properties of the Midwest field was significantly different from the

New York State field, and that field conditions explained most of the variation in the α- and β-diversity of the rhizosphere microbiome from 27 maize inbred lines (Peiffer et al 2013). Therefore, it is not surprising that I observed that field conditions contribute significantly to the variation in the abundance of the OTUs containing the *Pseudomonas* isolates.

Multiple investigations have shown that plant genotypes affected the abundance of the overall bacterial communities (Peiffer et al 2013) as well as individual bacterial taxa (Costa et al 2006, Depret and Laguerre 2008, Fromin et al 2001), whereas other studies, such as one on the plant symbiotic nitrogen-fixing *Sinorhizobium sp.* associated with Medicago concluded that the host plant diversity was not related to the diversity of *S. sp.* isolates: in this study, the diversity of *S. sp.* isolates from 20 different Medicago genotypes was similar to that of *S. sp.* isolates from 20 Medicago plants of the same genotype (Bailly et al 2006). Previous research that focused on rhizosphere *Pseudomonas* have shown that genetically different wheat and potato lines influenced the relative abundance of taxonomic and functional *Pseudomonas* populations, respectively (Dias et al 2013, Meyer et al 2013). For this study, it is possible that maize genotypes do not influence the diversity of *Pseudomonas* in the rhizosphere significantly, but are related to the differences in the functions of the *Pseudomonas* isolates.

Maize genotypes are significantly associated with *Pseudomonas* isolate function profiles

To assess whether maize genotypes have a significant influence on the functions of the rhizosphere *Pseudomonas* isolates, I focused on the relationship between maize genotypes and the differences in metabolic profiles for the *Pseudomonas* isolates. I generated the MetaCyc (Caspi et al 2008) function profiles

of the *Pseudomonas* isolates, which enumerate genes involved in non-redundant and experimentally verified primary and secondary metabolic pathways. PERMANOVA on the Bray-Curtis dissimilarity for the MetaCyc function profiles of the isolates shows that field conditions are significantly associated with the differences in the counts of genes in the metabolic profiles of the *Pseudomonas* isolates ($P < 0.01$), and contribute to 11.141% of the variation. After controlling for field conditions, I observed a significant association between maize genotypes and the variation in the MetaCyc function profiles of the *Pseudomonas* isolates ($P < 0.05$), and that maize genotypes explain 5.658% of the variation.

To investigate whether maize genotypes contribute significantly to the variation in the function profiles of the *Pseudomonas* isolates after controlling for the effects from the taxonomy of *Pseudomonas* isolates and field conditions, I included in the PERMANOVA analysis the taxonomy of the isolates based on what OTUs they belong to. The OTU taxonomy of the *Pseudomonas* isolates significantly explains a large part of the variation in the counts of the metabolic genes of the *Pseudomonas* isolates (47.891%, $P < 0.01$). After controlling for the OTU taxonomy effect, field conditions also contribute significantly to the differences in the function profiles of the isolates (7.905%, $P < 0.01$). After controlling for the effects from OTU taxonomy and fields, I still found that maize genotypes explain a small but significant proportion of the variation in the counts of the metabolic genes of the *Pseudomonas* isolates (2.694%, $P < 0.05$). I did not observe a significant contribution from the interactions between *Pseudomonas* OTUs, fields, or maize genotypes. This suggests that the function profile of one *Pseudomonas* isolate from a given maize genotype is not dependent on the particular OTU the isolate belongs to, or the particular field where the isolate was cultured.

Numerous studies have supported the notion that plant genotypes select for functional bacterial populations. Previously, researchers have identified the selection from different rice cultivars on the ammonia-oxidizing bacteria in the rhizosphere (Briones et al 2002). Studies comparing the auxin-producing (Picard and Bosco 2005) and 2,4-diacethylphloroglucinol-producing (2,4-DAPG-producing) (Picard and Bosco 2006) *Pseudomonas* strains from parental lines and heterozygous offspring lines concluded that hybrid maize inbred lines with different genotypes from their parents were able to recruit more antiphytopathogenic *Pseudomonas* strains to their rhizosphere. Also, five wheat cultivars differed in their ability to enrich 2,4-DAPG-producing *Pseudomonas* strains; more specifically, one wheat cultivar exclusively enriched *P. fluorescens* containing one type of 2,4-DAPG-producing key gene, whereas another wheat cultivar recruited the majority of *P. fluorescens* containing another type of 2,4-DAPG-producing key gene (Mazzola et al 2004). It is not surprising here that I found that maize genotypes are significantly associated with the differences in the function profiles of their rhizosphere *Pseudomonas* isolates. This implies that these two maize genotypes select for their rhizosphere *Pseudomonas* from the perspective of functions, or that these maize genotypes interact specifically with *Pseudomonas* containing certain metabolic capacities. Notably, although most of the variation in the function profiles of the *Pseudomonas* isolates is explained by the OTU taxonomy of the isolates and field conditions, the OTU a *Pseudomonas* isolate belongs to, or the field where the isolate was cultured does not determine the function profile of the particular *Pseudomonas* isolate, as suggested by the non-significant interactions between the OTU taxonomy of the isolates and maize genotypes, or field conditions and maize genotypes. These results all demonstrate that although small, the maize genotype effect observed in this study is true and significant, and is

independent of the effects from the OTU taxonomy of the *Pseudomonas* isolates or the field conditions.

To identify the *Pseudomonas* function profiles that are significantly associated with each maize genotype, I employed the indicator species approach (Tables 3.4 and 3.5). The *Pseudomonas* isolates from the Mo17 maize rhizosphere are enriched in genes involved in the metabolic pathways for denitrification (Table 3.4). *Pseudomonas* heavily participate in the denitrification process (Cheneby et al 2004). A previous study has shown that the Mo17 maize secreted more citrate than other maize inbred lines, which led to a slightly more acidic environment at its rhizosphere root cap (Piñeros et al 2005). Interestingly, artificial root exudates containing the highest amount of organic acid applied to maize roots resulted in significantly higher activity of nitrate reducers (Henry et al 2008). Therefore, the slightly more acidic rhizosphere may explain enriched denitrification genes in *Pseudomonas* isolates from the Mo17 maize rhizosphere. Note that I am aware that the MetaCyc function profiles list genes involved in metabolic pathways, which may not include complete metabolic pathways. I have verified that the enriched nitrate reduction I (denitrification) pathway contains all the enzymes (EC 1.7.2.1, EC 1.7.2.4, EC 1.7.2.5, and EC 1.7.99.4) for the complete pathway. Previous studies also showed that root morphology varied by maize genotypes, with Il14h having longer root systems than Mo17 (Kumar et al 2012), and that maize root mucilage affected the diversity of denitrification bacterial population (Mounier et al 2004). Thus, another possible reason for enrichment of denitrification genes in *Pseudomonas* isolates cultured from the Mo17 maize rhizosphere is that the distinct root morphology and root mucilage of the Mo17 maize attract more denitrifying *Pseudomonas* to the rhizosphere.

Table 3.4 The *Pseudomonas* metabolic genes associated with the Mo17 maize genotype. The genes were identified using the 'indicator species' approach. Column 1 shows the IDs of MetaCyc pathways containing the enriched metabolic genes, column 2 shows the information for the pathways, and column 3 shows whether the pathways are complete with all enzymes or not.

Group Mo17		
MetaCyc pathway IDs	**Pathways**	**Enzymes involved in each pathway**
PWY-66	GDP-L-fucose biosynthesis I (from GDP-D-mannose)	EC 4.2.1.47, missing EC 1.1.1.271
DENITRIFICATION-PWY	nitrate reduction I (denitrification)	EC 1.7.2.1, EC 1.7.2.4, EC 1.7.2.5, and EC 1.7.99.4
PWY-6748	nitrate reduction VII (denitrification)	EC 1.7.2.1, EC 1.7.2.4 and EC 1.7.99.4, missing EC 1.7.5.2
PWY0-1338	polymyxin resistance	EC 1.1.1.305, EC 2.1.2.13, EC 2.4.2.43, EC 2.6.1.87, and EC 2.7.8.30
HCAMHPDEG-PWY	3-phenylpropanoate and 3-(3-hydroxyphenyl)propanoate degradation to 2-oxopent-4-enoate	EC 1.13.11.16 and EC 1.14.13.127, missing EC 1.14.12.19, EC 1.3.1.87, and EC 3.7.1.14.
PWY-6690	cinnamate and 3-hydroxycinnamate degradation to 2-oxopent-4-enoate	EC 1.13.11.16 and EC 1.14.13.127, missing EC 1.14.12.19, EC 1.3.1.87, and EC 3.7.1.14.
PWY-5641	2-nitrotoluene degradation	EC 3.7.1-, missing EC 1.13.11.2
TOLUENE-DEG-DIOL-PWY	toluene degradation to 2-oxopent-4-enoate (via toluene-cis-diol)	EC 3.7.1-, missing EC 1.13.11.2, EC 1.14.12.11, and EC 1.3.1.19
PWY-1501	mandelate degradation I	EC 1.2.1.28 and EC 4.1.1.7, missing EC 1.1.99.31, EC 1.2.1.7, and EC 5.1.2.2
PWY-5648	2-nitrobenzoate degradation II	EC 1.14.12.1
PWY-6079	anthranilate degradation I (aerobic)	EC 1.14.12.1
PWY-6444	benzoate biosynthesis II (CoA-independent, non-β-oxidative)	EC 1.2.1.28, missing EC 4.3.1.24
PWY-6446	benzoate biosynthesis III (CoA-dependent, non-β-oxidative)	EC 1.2.1.28
PROPIONMET-PWY	methylmalonyl pathway	EC 5.1.99.1, missing EC 5.4.99.2 and EC 6.4.1.3
TRPCAT-PWY	tryptophan degradation I (via anthranilate)	EC 1.13.11.11, EC 3.5.1.9, and EC 3.7.1.3

In addition, the *Pseudomonas* isolates from the Mo17 maize rhizosphere are enriched in the genes for degrading anthranilate and tryptophan (Table 3.5). The Il14h maize produces higher amount of 2, 4-Dihydroxy-7-methoxy-I, 4-benzoxazin-3-one (DIMBOA) (Butrón et al 2010). The *Pseudomonas* isolates from the Mo17 maize rhizosphere could help to decrease the DIMBOA levels by degrading the DIMBOA precursors, as anthranilate has been shown to be incorporated into DIMBOA (Kumar and Chilton 1994), and tryptophan and DIMBOA share common steps and intermediates for their synthesis (Melanson et al 1997). I have also verified that the anthranilate degradation I (aerobic) and tryptophan degradation I (via anthranilate) pathways contain the enzymes for the complete pathways.

Table 3.5 The *Pseudomonas* metabolic genes associated with the Il14h maize genotype. The columns are the same as in Table 3.4.

Group Il14h		
MetaCyc pathway IDs	**Pathways**	**Enzymes involved in each pathway**
PWY-6714	GDP-L-fucose biosynthesis I (from GDP-D-mannose)	EC 4.2.1.47, missing EC 1.1.1.271
PWY-6714	L-rhamnose degradation III	EC 1.1.1.-, EC 4.2.1.90, missing EC 1.1.1.173 and EC 3.1.1.65
PWY-7136	beta myrcene degradation pathway	EC 1.1.1.-, missing EC 1.2.1.86, EC 4.2.1.127, EC 5.4.4.4
GAMMAHEXCHLORDEG-PWY	gamma-hexachlorocyclohexane degradation	EC 1.1.1.-, missing EC 1.3.1.32, EC 3.8.1.5, and EC 4.5.1.-
PWY-6080	4-ethylphenol degradation (anaerobic)	EC 1.1.1.-, missing EC 1.17.99 and EC 1.3.7.9
PWY-6391	meso-butanediol biosynthesis I	EC 1.1.1.-
PWY-6392	meso-butanediol biosynthesis II	EC 1.1.1.-
PWY-5451	acetone degradation I (to methylglyoxal)	EC 1.1.1.-, missing EC 1.1.1.80, EC 1.14.14.1 and EC 4.1.1.4
PWYQT-4450	aliphatic glucosinolate biosynthesis, side chain elongation cycle	EC 1.1.1.-, missing EC 2.3.3.- and EC 5.4.4.-
7ALPHADEHYDROX-PWY	cholate degradation (bacteria, anaerobic)	EC 1.1.1.-, missing EC 3.1.2.26, EC 4.2.1.106, and EC 6.2.1.7
PWY-5848	cinchona alkaloids biosynthesis	EC 1.1.1.-
DENITRIFICATION-PWY	nitrate reduction I (denitrification)	EC 1.7.2.1, EC 1.7.2.4, EC 1.7.2.5, and EC 1.7.99.4
PWY-5519	D-arabinose degradation III	EC 1.1.1.-, missing EC 1.2.1.26, EC 3.1.1.30 and EC 4.2.1.5
PWY-6491	D-galacturonate degradation III	EC 1.1.1.-
PWY-6501	D-glucuronate degradation II	EC 1.1.1.-
GALACTITOLCAT-PWY	galactitol degradation	EC 1.1.1.-, missing EC 2.7.1.144 and EC 4.1.2.40
PWY-6678	geraniol and nerol degradation	EC 1.1.1-, missing EC 1.2.1.86
PWY-6518	glycocholate metabolism (bacteria)	EC 1.1.1.-, missing EC 1.1.1.159, EC 1.1.1.176, EC 1.1.1.201, EC 1.1.1.238,and EC 3.5.1.24
PWY-2601	isethionate degradation	EC 1.1.1.-
P302-PWY	L-sorbose degradation	EC 1.1.1.-, missing EC 1.1.1.140
LACTOSEUTIL-PWY	lactose degradation II	EC 1.1.1.-, missing EC 1.1.99.13 and EC 3.2.1.23
PWY-5453	methylglyoxal degradation III	EC 1.1.1.-
PWY-5271	phaseic acid biosynthesis	EC 1.1.1.-, missing EC 1.14.13.93
PWY-5410	traumatin and (Z)-3-hexen-1-yl acetate biosynthesis	EC 1.1.1.-, missing EC 1.13.11.12 and EC 2.3.1.195
PWY-5516	xylose degradation II	EC 1.1.1.-
PWY-5782	2-keto-L-gulonate biosynthesis	EC 1.1.1.- and EC 1.1.99.21, missing EC 1.1.99.32
PWY-6704	L-ascorbate degradation IV	EC 1.1.1.- and EC 1.1.1.264

On the other hand, in the *Pseudomonas* isolates cultured from the rhizosphere of the sweet corn maize inbred line Il14h, several sugar-metabolism related genes are associated with this maize genotype. The maize sweet corn inbred line Il14h harbors a mutation at the sugary1 gene, which results in higher sucrose and glucose concentration and lower starch production in the endosperm (James et al 1995). Relatively little is known for the influence of sugary1 mutation on sugar concentrations in stalks and roots in adult sweet corn plants, and one study has proposed that during the movement of sugars from stalks to kernels as sweet corn plants developed from the tassel formation stage to the milk stage, which includes flowering time (LANCASHIRE et al 1991), the levels of sucrose increased while the levels of fructose and glucose dropped in the ninth stalk internode (Russo and Smith 1999). It is well known that plants release root exudates into their rhizosphere, which mediate interactions between plants and their roots-associated microbiome (Huang et al 2014), and sugars make up 65% of the maize root exudates (Aira et al 2010). The relatively higher amount of sucrose in the sweet corn is likely to be released into the rhizosphere in the form of root exudates, and may attract *Pseudomonas* containing higher number of genes involved in the metabolism of sugars. Moreover, it has been shown previously that *P. chlororaphis O6* could produce meso-butanediol (Han et al 2006). I also observed that the *Pseudomonas* isolates from the Il14h maize rhizosphere enrich genes for meso-butanediol biosynthesis pathways.

Molecular Evolution

To investigate whether the enriched metabolic genes associated with each maize genotype in the *Pseudomonas* isolates are under the selection from their respective maize genotype, I conducted molecular evolution analysis to infer natural selection on the enriched metabolic gene sequences from the *Pseudomonas*

genomes. I tested the nitrate reduction related genes enriched in the *Pseudomonas* isolates from the Mo17 maize rhizosphere, and the sugar metabolic genes enriched in the *Pseudomonas* isolates from the Il14h maize rhizosphere (see Table 3.6 for a list of genes used to infer selection). I aligned the gene sequences based on the alignment of their protein sequences, and inferred natural selection using several methods integrated in HyPhy (Pond and Muse 2005). I did not identify any positively selected site, whereas I observed that multiple sites in several genes are under negative selection. I also verified that there is no recombination in the gene sequences used to infer natural selection.

Table 3.6 List of *Pseudomonas* genes used for the molecular evolution analyses to infer natural selection. Sub-tables a, b, and c lists the denitrification genes, sugar metabolic genes, and regulatory genes for denitrification used for the molecular evolution analyses that are associated with the Mo17, Il14h, and Mo17 maize inbred lines, respectively. Column 1 shows the names of the genes, and columns 2 and 3 show whether positive or negative is detected in the genes.

Table 3.6a		
Mo17 enriched metabolic genes		
Genes	**Positive selection**	**Negative selection**
assimilatory nitrate reductase (NADH) alpha subunit apoprotein	No	Yes
anthranilate 1,2-dioxygenase, small subunit	No	Yes
4-amino-4-deoxy-L-arabinose transferase and related glycosyltransferases of PMT family	No	Yes
Table 3.6b		
Il14h enriched metabolic genes		
Genes	**Positive selection**	**Negative selection**
Short-chain alcohol dehydrogenase of unknown specificity	No	Yes
Dehydrogenases with different specificities	No	No
Table 3.6c		
genes regulating nitrate reduc.		
Genes	**Positive selection**	**Negative selection**
respiratory nitrate reductase chaperone NarJ	No	Yes
periplasmic nitrate reductase chaperone NapD	No	Yes
Signal transduction histidine kinase, nitrate/nitrite-specific	No	Yes

I then asked whether regulatory genes for the enriched metabolic genes associated with each maize genotype are under the selection from maize. Previous studies examining positive selection in *Streptococcus* genomes have found that the two-component signal transduction kinase genes regulating virulence gene

expression were under positive Darwinian selection (Anisimova et al 2007). So I proposed that maybe the genes regulating metabolic gene expression are under the selection from maize, and changes in the regulatory genes lead to the enrichment of metabolic genes in the *Pseudomonas* isolates. I sought to infer natural selection in the known regulatory genes for denitrification, including the nitrate/nitrite-specific signal transduction histidine kinase genes NarXL (Sparacino-Watkins et al 2014) and the nitrate reductase chaperone genes (Grahl et al 2012, Sparacino-Watkins et al 2014). I did not identify any site under positive selection from maize, whereas I observed negative selection at some sites in the regulatory genes (Table 3.6).

I also estimated the evolutionary divergence between the gene sequences that were included in the analyses to infer natural selection. The estimated pairwise genetic distance of the genes ranges from 0.03 to 0.2, indicating a moderate level of divergence between these genes. As the divergence level increases, the relative number of non-synonymous mutation decreases, leading to higher non-synonymous mutation to synonymous mutation ratio (d_N/d_S) (Forsdyke 2007); As a d_N/d_S ratio over 1 indicates positive selection, sequences with higher divergence level are more likely to show positive selection if the selection is present. That I did not identify any positively selected site in the moderately divergent sequences indicates that there is no molecular adaptation in these genes, and that the variation in the counts of the metabolic genes enriched in each maize genotype is more likely to have been influenced by the negative selection pressure from maize genotypes on the *Pseudomonas* isolates. This supports my PERMANOVA results that maize genotypes select on function abilities of the *Pseudomonas* isolates. This result is also in line with previous conclusions that purifying selection is pervasive in functional genetic elements in bacterial genomes and that bacterial genomes evolve under negative

selection (Petersen et al 2007, Tümmler and Cornelis 2005). It is also possible that certain maize genes involved in recruiting the *Pseudomonas* isolates or some compounds in the maize root exudates play key roles in the differences in the metabolic abilities of the *Pseudomonas* isolates cultured from the rhizosphere of two different maize genotypes.

Co-occurring OTUs with *Pseudomonas* isolates

To further elucidate the functions of the *Pseudomonas* isolates and their relationships with the other rhizosphere bacteria, I identified the OTUs that co-occur with the *Pseudomonas* isolates using the maize rhizosphere microbiome 16S rRNA gene V4 region amplicon data (Jin, Z and Ley, R.E., unpublished data) (Figure 3.3-3.5). The 48 *Pseudomonas* isolates belong to 8 *Pseudomonas* OTUs (Table 3.3), and I identified from 3 to over 1000 co-occurring OTUs for each OTU containing the *Pseudomonas* isolates (not listed due to table size limit). Among the co-occurring OTUs with the *Pseudomonas* isolates, I observed OTUs from the *Achromobacter*, *Arthrobacter*, *Bacillus*, *Paenibacillus*, and *Stenotrophomonas* families that have been shown previously to be diazotrophic isolates from the wheat rhizosphere along with *Pseudomonas* (Venieraki et al 2011), and OTUs from the *Bradyrhizobium* family, which have been shown to co-operate with *Pseudomonas* to enhance nitrogen fixation in legumes (Barea et al 2005); These bacteria may collaborate with *Pseudomonas* on nitrogen fixation in the maize rhizosphere. I also identified co-occurring OTUs of the *Acidovorax*, *Agrobacterium*, *Rhizobium*, *Sphingomonas*, and *Variovorax* families that have been shown to produce the same quorum sensing signaling molecules as *Pseudomonas* (D'Angelo-Picard et al 2005), and OTUs from the *Acinetobacter* family which could degrade the signaling molecules produced by *Pseudomonas* (Chan et al 2011); These bacteria may interact with the *Pseudomonas*

isolates via inter-genus communicating (Dubern and Diggle 2008). In addition, I found that the *Pseudomonas* isolates from both maize genotypes and both fields all have non-random associations with some other *Pseudomonas* OTUs. This suggests that *Pseudomonas* OTUs in the maize rhizosphere may interact with each other in ways similar to how different 2,4-DAPG-producing *Pseudomonas* strains interacted in the wheat rhizosphere (Landa et al 2003).

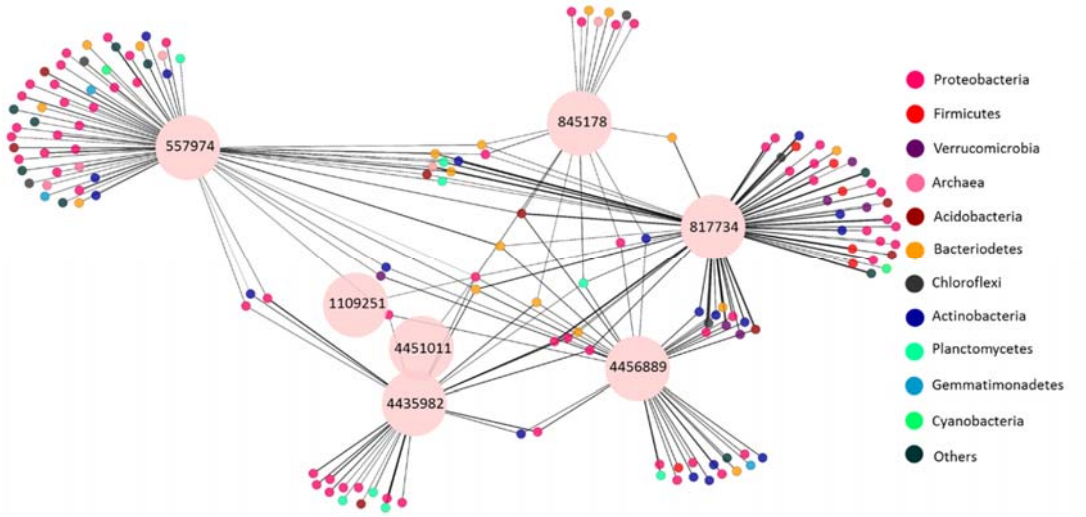

Figure 3.3 Network of co-occurring OTUs with the OTUs containing the *Pseudomonas* isolates in Lansing. The seven big nodes with numbers show the OTUs containing the *Pseudomonas* isolates and the OTU IDs. The smaller nodes were co-occurring OTUs with the OTUs containing the *Pseudomonas* isolates, and were colored based on their phyla. The edge between an OTU containing the *Pseudomonas* isolates and any of its co-occurring OTU represents a strong (Spearman's $\rho > 0.4$) and significant (Q-value < 0.01) correlation. The width of each edge is proportional to the Spearman's ρ between each node pairs.

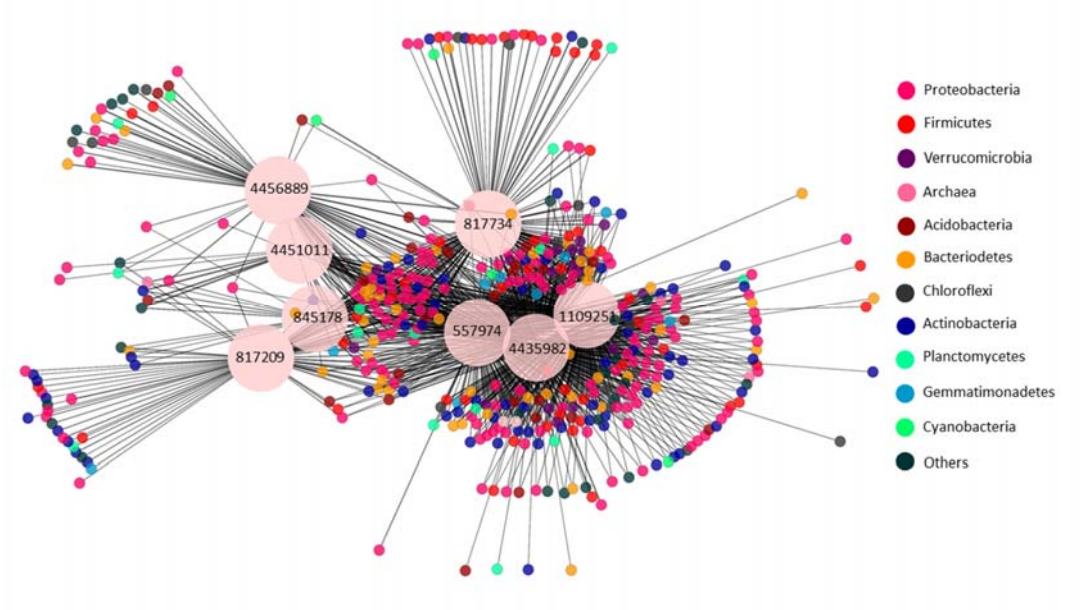

Figure 3.4 Network of co-occurring OTUs with OTUs containing the *Pseudomonas* isolates in the first plate of Urbana. The seven big nodes with numbers show the OTUs containing the *Pseudomonas* isolates and the OTU IDs. The smaller nodes were co-occurring OTUs with the OTUs containing the *Pseudomonas* isolates, and were colored based on their phyla. The edge between an OTU containing the *Pseudomonas* isolates and any of its co-occurring OTU represents a strong (Spearman's $\rho > 0.4$) and significant (Q-value < 0.01) correlation. The width of each edge is proportional to the Spearman's ρ between each node pairs.

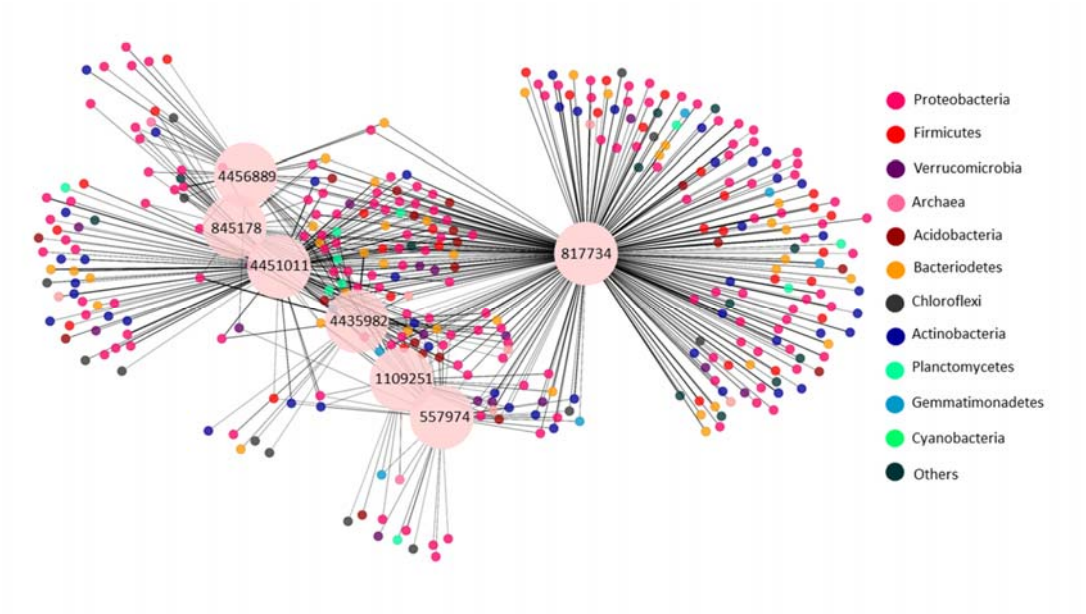

Figure 3.5 Network of co-occurring OTUs with OTUs containing the *Pseudomonas*

isolates in the second plate of Urbana. The seven big nodes with numbers show the OTUs containing the *Pseudomonas* isolates and the OTU IDs. The smaller nodes were co-occurring OTUs with the OTUs containing the *Pseudomonas* isolates, and were colored based on their phyla. The edge between an OTU containing the *Pseudomonas* isolates and any of its co-occurring OTU represents a strong (Spearman's $\rho > 0.4$) and significant (Q-value < 0.01) correlation. The width of each edge is proportional to the Spearman's ρ between each node pairs.

To identify the co-occurring OTUs unique to each maize genotype across two fields, I focused on one OTU that contains the *Pseudomonas* isolates from both maize genotypes and both fields. I compared the co-occurring OTUs for this OTU from the same maize genotype grown in both fields, and identified the shared co-occurring OTUs (Table 3.7). Comparison of the shared co-occurring OTUs from each maize genotype shows that bacterial OTUs from the *Chitinophagaceae* and *Bacteriovoracaceae* families, the MND1 order of Betaproteobacteria, the RB41 order of Acidobacteria, and the *Rhodoplanes* and *Frankia* genera are unique co-occurring OTUs to the *Pseudomonas* isolates from the Il14h maize rhizosphere, whereas an bacterial OTU from the *Edomicrobium* genus is a unique co-occurring OTU to the *Pseudomonas* isolates from the Mo17 maize rhizosphere. The common co-occurring OTUs to the *Pseudomonas* isolates from both maize genotypes include bacterial OTUs from the *Sphingomonadaceae*, *Syntrophobacteraceae*, *Gemmataceae*, and *Chitinophagaceae* families. Among these co-occurring OTUs, bacteria from the *Bacteriovoracaceae* family have been reported to prey on *Pseudomonas* (Davidov et al 2006). Although there is no full knowledge for how the other co-occurring OTUs interact with the *Pseudomonas* isolates in the rhizosphere of each maize genotype, some of these OTUs have been reported as members of the maize rhizosphere (Bouffaud et al 2014, Chauhan et al 2011, Garcia-Salamanca et al 2013, Li et al 2014), and the understanding of their roles and functions will benefit from future studies on maize-rhizosphere microbiome interactions and functional potential of

microbes in the maize rhizosphere microbiome.

Table 3.7 Unique and shared co-occurring OTUs with the OTUs containing the *Pseudomonas* isolates for each maize genotype. Sub-tables a and b list the co-occurring OTUs unique to the Mo17 and Il14h maize genotype, respectively. Sub-table c lists the co-occurring OTUs shared by the two maize genotypes. OTU IDs are Greengenes OTU numbers.

Table 3.7a	
unique to mo17	
OTU IDs	**Taxonomy of co-occurring OTUs**
4433035	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhizobiales; f__Hyphomicrobiaceae; g__Pedomicrobium; s__
Table 3.7b	
unique to Il14h	
OTU IDs	**Taxonomy of co-occurring OTUs**
114076	k__Bacteria;, p__Bacteroidetes;, c__[Saprospirae];, o__[Saprospirales];, f__Chitinophagaceae;, g__;, s__
11544	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Frankiaceae; g__Frankia; s__
185100	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Bdellovibrionales; f__Bacteriovoracaceae; g__; s__
4440262	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__MND1; f__; g__; s__
4450676	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhizobiales; f__Hyphomicrobiaceae; g__Rhodoplanes; s__
697997	k__Bacteria; p__Acidobacteria; c__[Chloracidobacteria]; o__RB41; f__; g__; s__
Table 3.7c	
shared	
OTU IDs	**Taxonomy of co-occurring OTUs**
1085229	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Sphingomonadaceae; g__; s__
2254354	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Sphingomonadaceae; g__Sphingomonas; s__
4260136	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Syntrophobacterales; f__Syntrophobacteraceae; g__; s__
4433032	k__Bacteria; p__Planctomycetes; c__Planctomycetia; o__Gemmatales; f__Gemmataceae; g__; s__
853114	k__Bacteria; p__Bacteroidetes; c__[Saprospirae]; o__[Saprospirales]; f__Chitinophagaceae; g__; s__
929398	k__Bacteria; p__Bacteroidetes; c__[Saprospirae]; o__[Saprospirales]; f__Chitinophagaceae; g__; s__

Conclusion

In this study I characterized the taxonomy of the 48 *Pseudomonas* isolates cultured from two different maize genotypes grown in two fields at the same developmental stage. Although I did not observe a significant maize genotype effect on the pan-genome SNPs or the abundance of the OTUs containing the *Pseudomonas* isolates, I showed that maize genotypes significantly contribute to the variation in the counts of metabolic genes of the *Pseudomonas* isolates. I also identified the metabolic genes from the *Pseudomonas* isolates that are associated with each maize genotype. I conducted molecular evolution analyses in the enriched metabolic genes and observed genes under negative selection. In addition, I identified co-occurring OTUs from the same maize rhizosphere where the *Pseudomonas* isolates were cultured; these co-occurring OTUs may be involved in various cooperative activities such as nitrogen fixation and cell-cell communication with the *Pseudomonas* isolates in the maize rhizosphere.

The experimental design permitted me to test the influence of maize genotypes on its rhizosphere *Pseudomonas* isolates across different field conditions, and to assess the degree to which these maize-*Pseudomonas* interactions depend upon the maize genotypes as well as the field conditions and the taxonomy of the isolates. Sequencing of the *Pseudomonas* isolates also allowed me to examine the whole genomes of the isolates and to specify the metabolic genes that imply maize genotype selection effect, providing a wide range of candidates that will benefit future studies on plant-microbiome interactions and crop breeding.

REFERENCES

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

Anderson MJ (2001). A new method for non‐parametric multivariate analysis of variance. *Austral Ecology* **26:** 32-46.

Anisimova M, Bielawski J, Dunn K, Yang Z (2007). Phylogenomic analysis of natural selection pressure in Streptococcus genomes. *BMC evolutionary biology* **7:** 154.

Bailly X, Olivieri I, De Mita S, CLEYET‐MAREL JC, Béna G (2006). Recombination and selection shape the molecular diversity pattern of nitrogen‐fixing Sinorhizobium sp. associated to Medicago. *Molecular Ecology* **15:** 2719-2734.

Barea J-M, Pozo MJ, Azcon R, Azcon-Aguilar C (2005). Microbial co-operation in the rhizosphere. *Journal of experimental botany* **56:** 1761-1778.

Baudoin E, Benizri E, Guckert A (2003). Impact of artificial root exudates on the bacterial community structure in bulk soil and maize rhizosphere. *Soil Biology and Biochemistry* **35:** 1183-1192.

Beals EW (1984). Bray-Curtis ordination: an effective strategy for analysis of

multivariate ecological data. *Advances in Ecological Research* **14:** 55.

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)***:** 289-300.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG *et al* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53-59.

Bouffaud ML, Poirier MA, Muller D, Moenne-Loccoz Y (2014). Root microbiome relates to plant host evolution in maize and other Poaceae. *Environ Microbiol*.

Brencic A, Winans SC (2005). Detection of and response to signals involved in host-microbe interactions by plant-associated bacteria. *Microbiol Mol Biol Rev* **69:** 155-194.

Briones AM, Okabe S, Umemiya Y, Ramsing N-B, Reichardt W, Okuyama H (2002). Influence of different cultivars on populations of ammonia-oxidizing bacteria in the root environment of rice. *Applied and environmental microbiology* **68:** 3067-3075.

Briones Jr AM, Okabe S, Umemiya Y, Ramsing N-B, Reichardt W, Okuyama H (2003). Ammonia-oxidizing bacteria on root biofilms and their possible contribution to N use efficiency of different rice cultivars. *Plant and Soil* **250:** 335-348.

Butrón A, Chen Y, Rottinghaus G, McMullen MD (2010). Genetic variation at bx1 controls DIMBOA content in maize. *Theoretical and applied genetics* **120:** 721-734.

Cáceres MD, Legendre P (2009). Associations between species and groups of sites: indices and statistical inference. *Ecology* **90:** 3566-3574.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7:** 335-336.

Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, Latendresse M *et al* (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic acids research* **36:** D623-D631.

Castelo R, Roverato A (2006). A robust procedure for Gaussian graphical model search from microarray data with p larger than n. *The Journal of Machine Learning Research* **7:** 2621-2650.

Chan K-G, Atkinson S, Mathee K, Sam C-K, Chhabra SR, Cámara M *et al* (2011). Characterization of N-acylhomoserine lactone-degrading bacteria associated with the Zingiber officinale (ginger) rhizosphere: Co-existence of quorum quenching and quorum sensing in Acinetobacter and Burkholderia. *BMC microbiology* **11:** 51.

Chauhan PS, Chaudhry V, Mishra S, Nautiyal CS (2011). Uncultured bacterial diversity in tropical maize (Zea mays L.) rhizosphere. *J Basic Microbiol* **51:** 15-32.

Cheneby D, Perrez S, Devroe C, Hallet S, Couton Y, Bizouard F *et al* (2004). Denitrifying bacteria in bulk and maize-rhizospheric soil: diversity and N2O-reducing abilities. *Canadian journal of microbiology* **50:** 469-474.

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L *et al* (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6:** 80-92.

Costa R, Gomes N, Peixoto RS, Rumjanek N, Berg G, Mendonça-Hagler L *et al* (2006). Diversity and antagonistic potential of< i> *Pseudomonas*</i> spp. associated to the rhizosphere of maize grown in a subtropical organic farm. *Soil Biology and Biochemistry* **38:** 2434-2447.

D'Angelo‐Picard C, Faure D, Penot I, Dessaux Y (2005). Diversity of N‐acyl homoserine lactone‐producing and‐degrading bacteria in soil and tobacco rhizosphere. *Environmental Microbiology* **7:** 1796-1808.

Darriba D, Taboada GL, Doallo R, Posada D (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature methods* **9:** 772-772.

Davidov Y, Friedjung A, Jurkevitch E (2006). Structure analysis of a soil community of predatory bacteria using culture‐dependent and culture‐independent methods reveals

a hitherto undetected diversity of Bdellovibrio-and-like organisms. *Environmental Microbiology* **8:** 1667-1673.

De Vleesschauwer D, Höfte M (2009). Rhizobacteria-induced systemic resistance. *Advances in botanical research* **51:** 223-281.

Depret G, Laguerre G (2008). Plant phenology and genetic variability in root and nodule development strongly influence genetic structuring of Rhizobium leguminosarum biovar viciae populations nodulating pea. *New Phytol* **179:** 224-235.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* **72:** 5069-5072.

Dias AC, Dini-Andreote F, Hannula SE, Andreote FD, Pereira ESMC, Salles JF *et al* (2013). Different selective effects on rhizosphere bacteria exerted by genetically modified versus conventional potato lines. *PLoS One* **8:** e67948.

Dubern J-F, Diggle SP (2008). Quorum sensing by 2-alkyl-4-quinolones in *Pseudomonas* aeruginosa and other bacterial species. *Molecular bioSystems* **4:** 882-888.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32:** 1792-1797.

Flint-Garcia SA, Thuillet AC, Yu J, Pressoir G, Romero SM, Mitchell SE *et al* (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *Plant J* **44:** 1054-1064.

Forsdyke DR (2007). Positive Darwinian Selection: Does the Comparative Method Rule? *Journal of Biological Systems* **15:** 95-108.

Fromin N, Achouak W, Thiéry J, Heulin T (2001). The genotypic diversity of *Pseudomonas* brassicacearum populations isolated from roots of Arabidopsis thaliana: influence of plant genotype. *FEMS microbiology ecology* **37:** 21-29.

Gaiero JR, McCall CA, Thompson KA, Day NJ, Best AS, Dunfield KE (2013). Inside the root microbiome: Bacterial root endophytes and plant growth promotion. *American journal of botany* **100:** 1738-1750.

Garbeva P, van Veen JA, van Elsas JD (2004). Microbial diversity in soil: selection microbial populations by plant and soil type and implications for disease suppressiveness. *Annu Rev Phytopathol* **42:** 243-270.

Garcia-Salamanca A, Molina-Henares MA, van Dillewijn P, Solano J, Pizarro-Tobias P, Roca A *et al* (2013). Bacterial diversity in the rhizosphere of maize and the surrounding carbonate-rich bulk soil. *Microb Biotechnol* **6:** 36-44.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S *et al* (2004). Bioconductor: open software development for computational biology and

bioinformatics. *Genome biology* **5:** R80.

Grahl S, Maillard J, Spronk CA, Vuister GW, Sargent F (2012). Overlapping transport and chaperone-binding functions within a bacterial twin-arginine signal peptide. *Molecular microbiology* **83:** 1254-1267.

Grayston SJ, Wang S, Campbell CD, Edwards AC (1998). Selective influence of plant species on microbial diversity in the rhizosphere. *Soil Biology and Biochemistry* **30:** 369-378.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* **59:** 307-321.

Han SH, Lee SJ, Moon JH, Park KH, Yang KY, Cho BH *et al* (2006). GacS-dependent production of 2R, 3R-butanediol by *Pseudomonas* chlororaphis O6 is a major determinant for eliciting systemic resistance against Erwinia carotovora but not against *Pseudomonas* syringae pv. tabaci in tobacco. *Molecular Plant-Microbe Interactions* **19:** 924-930.

Henry S, Texier S, Hallet S, Bru D, Dambreville C, Cheneby D *et al* (2008). Disentangling the rhizosphere effect on nitrate reducers and denitrifiers: insight into the role of root exudates. *Environmental Microbiology* **10:** 3082-3092.

Hsu SF, Buckley DH (2009). Evidence for the functional significance of diazotroph

community structure in soil. *ISME J* **3:** 124-136.

Huang X, Chaparro JM, Reardon KF, Zhang R, Shen Q, Vivanco JM (2014). Rhizosphere interactions: root exudates, microbes and microbial communities. *Botany*.

James MG, Robertson DS, Myers AM (1995). Characterization of the maize gene sugary1, a determinant of starch composition in kernels. *The Plant Cell Online* **7:** 417-429.

Kampfer P, Glaeser SP (2012). Prokaryotic taxonomy in the sequencing era--the polyphasic approach revisited. *Environ Microbiol* **14:** 291-317.

Kumar B, Abdel-Ghani AH, Reyes-Matamoros J, Hochholdinger F, Lübberstedt T (2012). Genotypic variation for root architecture traits in seedlings of maize (Zea mays L.) inbred lines. *Plant Breeding* **131:** 465-478.

Kumar P, Chilton WS (1994). Incorporation of anthranilate-d< sub> 4</sub> into DIMBOA in maize. *Tetrahedron letters* **35:** 3247-3250.

LANCASHIRE PD, Bleiholder H, Boom T, Langelüddeke P, Stauss R, WEBER E *et al* (1991). A uniform decimal code for growth stages of crops and weeds. *Annals of Applied Biology* **119:** 561-601.

Landa BB, Mavrodi DM, Thomashow LS, Weller DM (2003). Interactions between

strains of 2, 4-diacetylphloroglucinol-producing *Pseudomonas* fluorescens in the rhizosphere of wheat. *Phytopathology* **93:** 982-994.

Letunic I, Bork P (2011). Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic acids research* **39:** W475-W478.

Li X, Rui J, Mao Y, Yannarell A, Mackie R (2014). Dynamics of the bacterial community structure in the rhizosphere of a maize cultivar. *Soil Biology and Biochemistry* **68:** 392-401.

Lugtenberg B, Kamilova F (2009). Plant-growth-promoting rhizobacteria. *Annu Rev Microbiol* **63:** 541-556.

Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y *et al* (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic acids research* **40:** D115-D122.

Martinez-Romero E (2009). Coevolution in Rhizobium-legume symbiosis? *DNA Cell Biol* **28:** 361-370.

Mavrodi DV, Blankenfeldt W, Thomashow LS (2006). Phenazine compounds in fluorescent *Pseudomonas* spp. biosynthesis and regulation. *Annu Rev Phytopathol* **44:** 417-445.

Mazzola M, Funnell DL, Raaijmakers J (2004). Wheat cultivar-specific selection of 2,

4-diacetylphloroglucinol-producing fluorescent *Pseudomonas* species from resident soil populations. *Microbial Ecology* **48:** 338-348.

McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q *et al* (2009). Genetic properties of the maize nested association mapping population. *Science* **325:** 737-740.

Melanson D, Chilton M-D, Masters-Moore D, Chilton WS (1997). A deletion in an indole synthase gene is responsible for the DIMBOA-deficient phenotype of bxbx maize. *Proceedings of the National Academy of Sciences* **94:** 13345-13350.

Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JH *et al* (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* **332:** 1097-1100.

Meyer JB, Song-Wilson Y, Foetzki A, Luginbuhl C, Winzeler M, Kneubuhler Y *et al* (2013). Does wheat genetically modified for disease resistance affect root-colonizing pseudomonads and arbuscular mycorrhizal fungi? *PLoS One* **8:** e53825.

Mounier E, Hallet S, Cheneby D, Benizri E, Gruet Y, Nguyen C *et al* (2004). Influence of maize mucilage on the diversity and activity of the denitrifying community. *Environmental Microbiology* **6:** 301-312.

Oksanen J, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R *et al* (2013). vegan: Community Ecology Package. R package version 2.0-9.

Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL *et al* (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proceedings of the National Academy of Sciences of the United States of America* **110:** 6548-6553.

Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R (2007). Genes under positive selection in Escherichia coli. *Genome research* **17:** 1336-1343.

Philippot L, Raaijmakers JM, Lemanceau P, van der Putten WH (2013). Going back to the roots: the microbial ecology of the rhizosphere. *Nature Reviews Microbiology* **11:** 789-799.

Piñeros MA, Shaff JE, Manslank HS, Alves VMC, Kochian LV (2005). Aluminum resistance in maize cannot be solely explained by root organic acid exudation. A comparative physiological study. *Plant Physiol* **137:** 231-241.

Picard C, Bosco M (2005). Maize heterosis affects the structure and dynamics of indigenous rhizospheric auxins - producing *Pseudomonas* populations. *FEMS microbiology ecology* **53:** 349-357.

Picard C, Bosco M (2006). Heterozygosis drives maize hybrids to select elite 2, 4-diacethylphloroglucinol - producing *Pseudomonas* strains among resident soil populations. *FEMS microbiology ecology* **58:** 193-204.

Pond SLK, Muse SV (2005). HyPhy: hypothesis testing using phylogenies. *Statistical methods in molecular evolution*. Springer. pp 125-181.

R Development Core Team (2005). R: A language and environment for statistical computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2013. url: http://www. R-project. org.

Raaijmakers JM, Weller DM (1998). Natural plant protection by 2, 4-diacetylphloroglucinol-producing *Pseudomonas* spp. in take-all decline soils. *Molecular Plant-Microbe Interactions* **11:** 144-152.

Reeve WG, Ballard R, Drew E, Tian R, Bräu L, Goodwin L *et al* (2014). Genome sequence of the Medicago-nodulating Ensifer meliloti commercial inoculant strain RRI128. *Standards in Genomic Sciences* **9**.

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

Russo VM, Smith T (1999). 464 Distribution of Sugars in Sweet Corn. *HortScience* **34:** 524.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D *et al* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction

networks. *Genome research* **13:** 2498-2504.

Singh BK, Millard P, Whiteley AS, Murrell JC (2004). Unravelling rhizosphere-microbial interactions: opportunities and limitations. *Trends Microbiol* **12:** 386-393.

Sparacino-Watkins C, Stolz JF, Basu P (2014). Nitrate and periplasmic nitrate reductases. *Chemical Society Reviews* **43:** 676-706.

Suyama M, Torrents D, Bork P (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34:** W609-W612.

Tümmler B, Cornelis P (2005). Pyoverdine receptor: a case of positive Darwinian selection in *Pseudomonas* aeruginosa. *Journal of bacteriology* **187:** 3289-3292.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013). MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular biology and evolution* **30:** 2725-2729.

Venieraki A, Dimou M, Pergalis P, Kefalogianni I, Chatzipavlidis I, Katinakis P (2011). The genetic diversity of culturable nitrogen-fixing bacteria in the rhizosphere of wheat. *Microbial Ecology* **61:** 277-285.

Zhang N, Wang D, Liu Y, Li S, Shen Q, Zhang R (2014). Effects of different plant root exudates and their organic acid components on chemotaxis, biofilm formation and

colonization by beneficial rhizosphere-associated bacterial strains. *Plant and Soil* **374:** 689-700.

Chapter 4

Heritability of Maize Rhizosphere Microbiome at Different Maize Developmental Stages

Zhao Jin, Antonio González, Jose Navas, Wei Zhang, Jeffery L. Dangl, Edward S. Buckler, Rob Knight, Ruth E. Ley

Abstract

Characterizing the heritability of the plant rhizosphere microbiome advances our understanding on the relationship between plant genetic control and the variation in the rhizosphere microbiome. Root exudates change at different plant growth stages, reflecting the variation of plant genetic control over time. I describe here a longitudinal study on the heritability of the maize rhizosphere microbiome conducted using 27 diverse maize lines grown in three different fields over the entire maize growing season. I estimated the proportion of variation in the beta diversity of the rhizosphere microbiome samples at each time point explained by maize genotypes, fields, and genotype by field interactions. I found that the maize genotype effect starts to increase at week 2 after planting, suggesting that the maize genetic control is taking effect. I observed the strongest maize genotype effect around flowering time. I also identified some potential heritable taxa as well as OTUs whose abundances vary over maize developmental stages. In addition, I observed increased species loss starting at week 2, which corresponds to the time point when maize genetic control starts to take effect, whereas species loss peaks at flowering time when maize imposes the strongest genetic control on the rhizosphere microbiome. The results from this study will expand our knowledge on the dynamics of the rhizosphere microbiome and how plant genotypes interact with environmental factors to cultivate the rhizosphere microbiome. These results will also benefit future studies that incorporate heritable plant-microbiome interactions into genetic models for plant breeding.

Introduction

Understanding the heritability of the microbiome is an important aspect to characterize the relationship between host genetic control and variation in the composition of the microbiome. Heritability refers to the proportion of phenotypic variation in a population accounted for by genetic variation of individuals (Spor et al 2011). Treating the microbiome as a quantitative trait, the heritability of the microbiome answers to what extent is host genetic variation related to the variation in the microbiome.

The rhizosphere microbiome is critical to the health and development of plants (Berendsen et al 2012), yet the heritability of the rhizosphere microbiome is not well characterized. An earlier investigation on the intraspecific heritability of root microbial communities from *Populus angustifolia* showed that intraspecific plant genotypic variation explained over half of the variation in microbial biomass nitrogen levels and the microbial community composition (Schweitzer et al 2008). Our recent study on the maize rhizosphere microbiome discovered that a small but significant proportion of the variation in the rhizosphere microbiome is heritable (Peiffer et al 2013). While research on vertebrate gut microbiome can provide us with additional insights into understanding the heritability and heritable components of microbiomes (Benson et al 2010, Hansen et al 2011, Meng et al 2014, Nelson et al 2011, Spor et al 2011, Turnbaugh et al 2009), further understanding on the heritability of plant rhizosphere microbiome is lacking.

Plant developmental stages also influence the rhizosphere microbiome besides the impact from plant genetic control and environmental factors. Changes in time may underlie the changes in many soil physiochemical properties, such as moisture (Baskan et al 2013), nitrogen availability (Cain et al 1999), or C/N ratio

(Zhang et al 2011). More specifically, the composition of root exudates changes as plants age (Baudoin et al 2002, Chaparro et al 2013), reflecting the variation of plant genetic control over time. Previous studies conducted on the Arabidopsis rhizosphere microbiome have identified that relative abundances of some bacterial taxa and functional genes in the rhizosphere microbiome followed temporal patterns in response to Arabidopsis developmental stages (Chaparro et al 2013, Chaparro et al 2014). Other studies carried out in different plant rhizosphere microbiomes have obtained similar conclusions that plant developmental stages affect the composition of the rhizosphere microbiome (Inceoglu et al 2010, Li et al 2014, Mougel et al 2006, van Overbeek and van Elsas 2008).

Using diverse maize inbred lines and a well-designed longitudinal study, I aimed to comprehend how maize genetic control on the rhizosphere microbiome changes over maize developmental stages. Our previous survey measured the heritability of the maize rhizosphere microbiome at flowering time, which was deemed a transition point in maize development from release of ample carbon resources to the rhizosphere to decrease in carbon flow to the rhizosphere due to maize reproduction (Peiffer and Ley 2013). Here I followed the rhizosphere microbiome of 27 maize inbred lines grown at three different fields over 20 weeks, and conducted time-series analysis on the heritability of the rhizosphere microbiome. I estimated the proportion of variation in the beta diversity of the rhizosphere microbiome samples at each time point explained by maize genotypes, fields, and genotype by field interactions. I found that the maize genotype effect starts to increase at week 2 after planting, suggesting that the maize genetic control is taking effect. I observed the strongest maize genotype effect around flowering time. I also identified some potential heritable taxa as well as OTUs whose abundances vary over maize

developmental stages. In addition, I observed increased species loss starting at week 2, which corresponds to the time point when maize genetic control starts to take effect, whereas species loss peaks at flowering time when maize imposes the strongest genetic control on the rhizosphere microbiome. My research revealed the changes in the heritability of the rhizosphere microbiome over maize developmental stages, and found potential heritable components of the rhizosphere microbiome. I also disentangled the temporal dynamics of species turnover and loss in the maize rhizosphere microbiome, and showed agreement between this temporal pattern and the dynamics of the maize genotype effect over time.

Materials and methods

Study design

This study was aimed at determining the heritability of diverse maize lines over maize developmental stages. The maize germplasm and planting of the maize inbred lines have been described previously by a team of people in my lab (Peiffer et al 2013). Briefly, 27 maize inbred lines were planted in a randomized complete block design in three different fields (two conventional managed, one organic) located at upstate New York. The rhizosphere soil samples of all maize inbred plots were collected every week from week 1 after planting to week 15 after planting, as well as week 20 after planting (Figure 4.1).

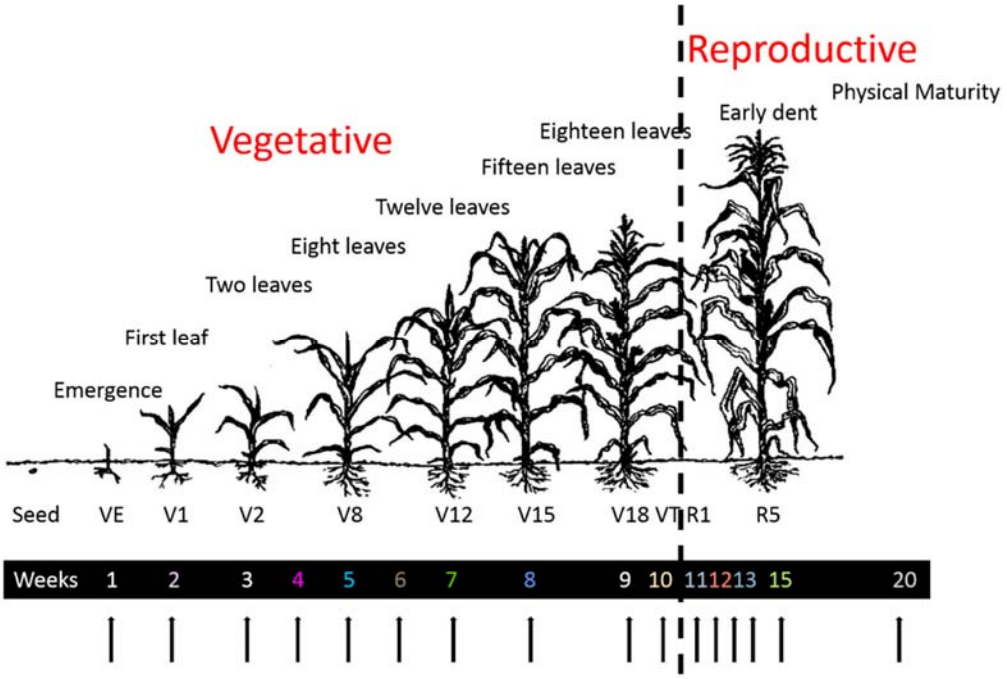

Figure 4.1 Sampling time of maize rhizosphere microbiome. VE to VT are vegetative maize developmental stages, and R1 to R5 are reproductive maize developmental stages. Weeks are converted from approximate days after seedling. Arrows indicate the weeks of rhizosphere soil samples taken. Figure adapted from www.smallgrains.org/springwh/Jun04/crop/crop.htm.

DNA extraction and 16S rDNA V4 region PCR amplification

Total genomic DNA was extracted from the rhizosphere soil samples as described previously by a team of people in my lab (Peiffer et al 2013). The partial 16S rRNA gene PCR amplification for all DNA samples was conducted following the Illumina MiSeq iTags working protocol with refined staggered primers (Joint Genome Institute, unpublished). Specifically, the 515F forward PCR primer (5'-AATGATACGGCGACCACCGAGATCTACAC *TCTTTCCCTACA* **GTGCCAGCMGCCGCGGTAA**-3', underlined, italic, and bold bases represent Illumina adapter, primer pad, and 16S rRNA gene V4 region forward primer, respectively) and the 806R reverse PCR primer (5'-

CAAGCAGAAGACGGCATACGAGAT XXXXXXXXXXXX NNN

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

GGACTACHVGGGTWTCTAAT-3', underlined, italic, and bold bases represent Illumina adapter, primer pad, and 16S rRNA gene V4 region forward primer, respectively. String of Xs represents the unique barcode for each reverse primer, and NNN represents the one to three staggered bases.)

The 16S rRNA gene V4 region PCR amplification was conducted as follows: Each rhizosphere soil DNA sample was amplified in duplicates in 50 µl PCR reactions, which contained 26 µl PCR water, 20 µl 5 PRIME HotMasterMix (5 PRIME, Inc., Gaithersburg, MD), 1 µl forward primer at 10 µM, 1 µl reverse primer at 10 µM, and 2 µl of rhizosphere DNA template. The PCR reaction mixes were set up manually, and the rhizosphere DNA templates were added using the Eppendorf epMotion liquid handling robotic workstation (Eppendorf North America, Hauppauge, NY). Thermal cycling consisted of an initial denaturation at 94 °C for 3 min, 35 cycles of denaturation at 94°C for 45 s, annealing at 50°C for 1 min, and elongation at 72°C for 1.5 min, followed by a final extension at 72°C for 10 min. The duplicate PCR reactions for each rhizosphere DNA sample were combined, and purified using the Agencourt AMPure XP PCR purification beads (Beckman Coulter, Indianapolis, IN). The purified PCR amplicons were quantified using the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, Grand Island, NY), and pooled in equimolar ratios into a single sample with a final concentration of ~ 20 ng/µl.

The 16S rRNA gene V4 region PCR amplicons were sequenced using Illumina technology (Bentley et al 2008) on the Illumina MiSeq platform at the Joint Genome Institute with the following sequencing primers: TCTTTCCCTACA GTGCCAGCMGCCGCGGTAA (read 1),

GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT (read 2), and GATCGGAAGAGCACACGTCTGAACTCCAGTCAC (index).

Analysis of 16S rRNA gene V4 region sequences

The analysis of the 16S rRNA gene V4 region sequences were conducted by Antonio González Peña and Jose Navas at Dr. Rob Knight's lab in University of Colorado Boulder. The 16S rRNA gene V4 region sequences were analyzed using the QIIME software package (Quantitative Insights into Microbial Ecology) (Caporaso et al 2010) version 1.8.0-dev with default settings. Pre-filtering on a total of 600,756,830 reads kept 99.72% of the sequences. Open-reference OTU picking was conducted on a total of 448,805,476 reads using the Greengenes (DeSantis et al 2006) August 2013 taxonomy as the reference at 97% sequence identity. 28359 OTUs were picked for 4405 samples, with 10% of the OTU table being non-zero values. The minimal and maximal reads per sample were 1 and 593118, respectively, with a median of 85917 reads and a standard deviation of 69734 reads. The unweighted and weighted UniFrac distances (Lozupone and Knight 2005) were calculated using the OTU table that was rarified (sub-sampled) at 10k and 20k read depths. After filtering out the Midwest and bulk soil samples, 3990 samples remained. The OTU table was converted to relative abundance in R 3.1.0 (R Core Team 2014). The absolute and relative abundance OTU tables were split by maize inbreds and maize ages, as well as collapsed into family-level OTUs in QIIME.

Statistical analyses

Permutation-based multiple regression was conducted on the family-level relative-abundance OTU table using the R package 'lmPerm' (Wheeler 2010). The R program 'EDGE' (Storey 2007) was employed to identify OTUs whose abundances vary with time. Partition of the beta diversity into turnover and nestedness

components was done using the R package ‘betapart’ (Baselga et al 2013). The R package ‘vegan’ (Oksanen et al 2013) and the function ‘capscale’ were used to estimate variation in beta diversity explained by maize inbreds, fields, and maize inbreds by field interaction using bootstrapped partial canonical analysis of principal coordinates. Heatmaps were generated using the R packages ‘gplots’ (Warnes et al 2014) and ‘RColorBrewer’ (Neuwirth 2011). The Poisson generalized linear models were fitted using the ‘glm’ function in the R base package ‘stats’, the negative binomial generalized linear models were fitted using the ‘glm.nb’ function in the R package ‘MASS’ (Venables and Ripley 2002), the zero-inflated generalized linear models were fitted using the R ‘pscl’ package (Zeileis et al 2007), the general and generalized linear mixed models were fitted using the R ‘lme4’ package (Bates et al 2014), and the linear mixed models with kinship matrix were fitted using the ‘lmekin’ function in the R ‘coxme’ package (Therneau 2012). The Bray-Curtis similarity metric was calculated using the R package ‘vegan’, and hierarchical clustering of OTUs was conducted using the R function ‘hclust’ in the R base package ‘stats’. The mantel test and Procrustes analysis were performed using the R ‘ade4’ (Dray and Dufour 2007) and ‘vegan’ packages, respectively. Statistical learning to define maize developmental stages or predict maize genotypes was done in the R package ‘pamr’ (Hastie et al 2013).

Results and Discussion

Influence of host genotypes on maize rhizosphere microbiome over time

To investigate the influence from maize genotypes on the rhizosphere microbiome over time, I first estimated the proportion of variation in the beta diversity (between sample diversity) of the rhizosphere microbiome samples explained by

maize genotypes, fields, and genotype by field interactions at each time point. I employed a bootstrapped partial canonical analysis of principle coordinates (CAP) to the beta diversity of rhizosphere microbiome samples collected from weeks 2, 4, 5, 6, 7, 8, 10, 11, 12, 13, 15, and 20 after planting, and assessed the variation accounted for by each term of interest after the effect of other terms had been 'partialled out'. This approach has been successfully applied to studying the diversity and heritability of the maize rhizosphere microbiome at flowering time previously (Peiffer et al 2013).

I conducted the CAP analysis on the unweighted and weighted UniFrac distances calculated at 10k and 20k rarefaction depths, and plotted the proportion of variation explained by maize genotypes, field, and genotype by field interactions at each time point. I showed that rarefaction depth does not influence the partitioning of variation (Figures 4.2 and 4.3).

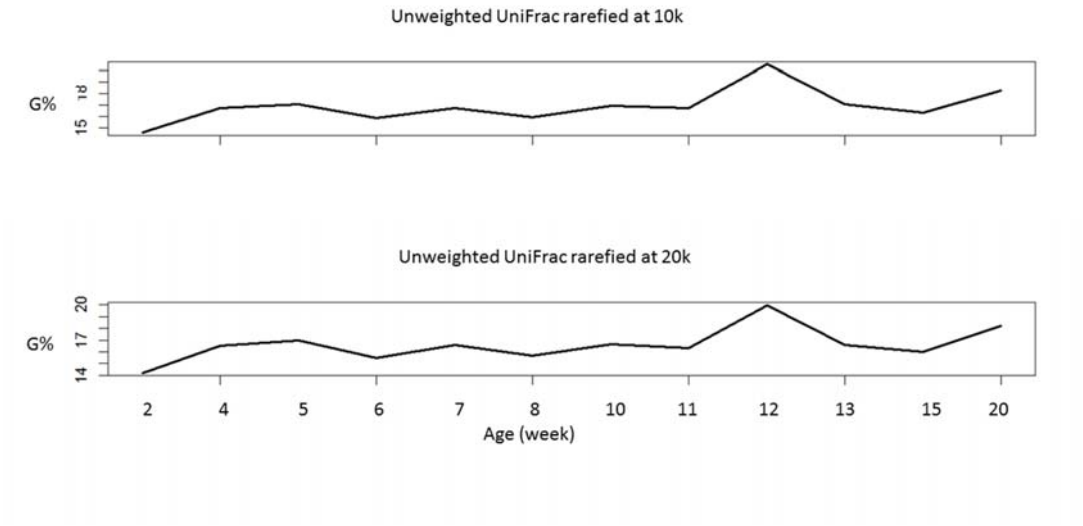

Figure 4.2 Proportion of variation in unweighted UniFrac distances rarefied at 10k and 20k explained by maize genotypes. X-axis, maize developmental time points in ages. Y-axis, percentage of variation in beta diversity explained by maize genotypes.

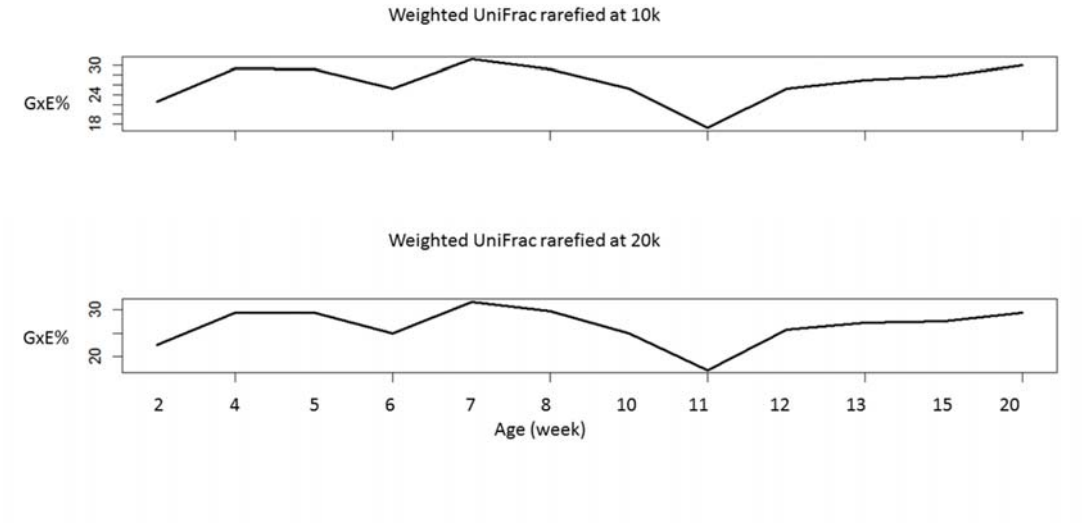

Figure 4.3 Proportion of variation in weighted UniFrac distances rarefied at 10k and 20k explained by maize genotype by field interactions. X-axis, maize developmental time points in ages. Y-axis, percentage of variation in beta diversity explained by genotype by field interactions.

The CAP analysis on the unweighted UniFrac distances shows that fields in general explain less variation than maize genotypes and genotype by field interactions. However, the variation in beta diversity explained by fields decreases sharply from week 2 to week 4, suggesting that maize genotypes starts to influence the rhizosphere microbiome after maize emergence (Figure 4.4). I observed an increase in maize genotype effect at the beginning of the vegetation growth stage at weeks 2 and 4, a slight decrease in these effects in the middle of the vegetation growth stage at week 6, and a near plateau from week 6 to week 11, a week before flowering. The maize genotype effect peaks at flowering time, indicating that root exudates are changing under the influence of maize genetic control at week 12. Interestingly, maize genotype effect still increases moderately after flowering and denting at week 15. Previous studies on the Arabidopsis root microbiome at different

Arabidopsis growth stages showed that towards the end of the growth cycle, the rhizosphere microbiomes from different Arabidopsis genotypes tended to be more alike, and that the rhizosphere microbiomes became more similar to bulk soil bacterial communities (Micallef et al 2009). While several other studies on Arabidopsis supported changes in root exudate contents over time in Arabidopsis (Chaparro et al 2013, Chaparro et al 2014), relatively less is known about changes in the maize root exudate profiles during maize development. As it is generally regarded that maize plants release less carbon sources in the root exudates as they divert more carbon into kernels, future studies monitoring maize root exudate profiles over time are needed to fully understand whether maize genetics still tightly control their rhizosphere microbiome in older maize plants. The temporal pattern of the proportion of variation in the rhizosphere microbiome beta diversity explained by maize genotype and field interaction shows a similar trend to that of the genotype effect with slight differences, indicating that the maize genotype effect may be dependent on which field the maize plant is grown in. In addition, the smaller proportion of variation explained by field than that explained by maize genotype and genotype by field interactions is consistent with our previous findings that the three New York state fields were similar despite different ways of management; the variation in the rhizosphere microbiome beta-diversity is thus less dependent on the differences between fields.

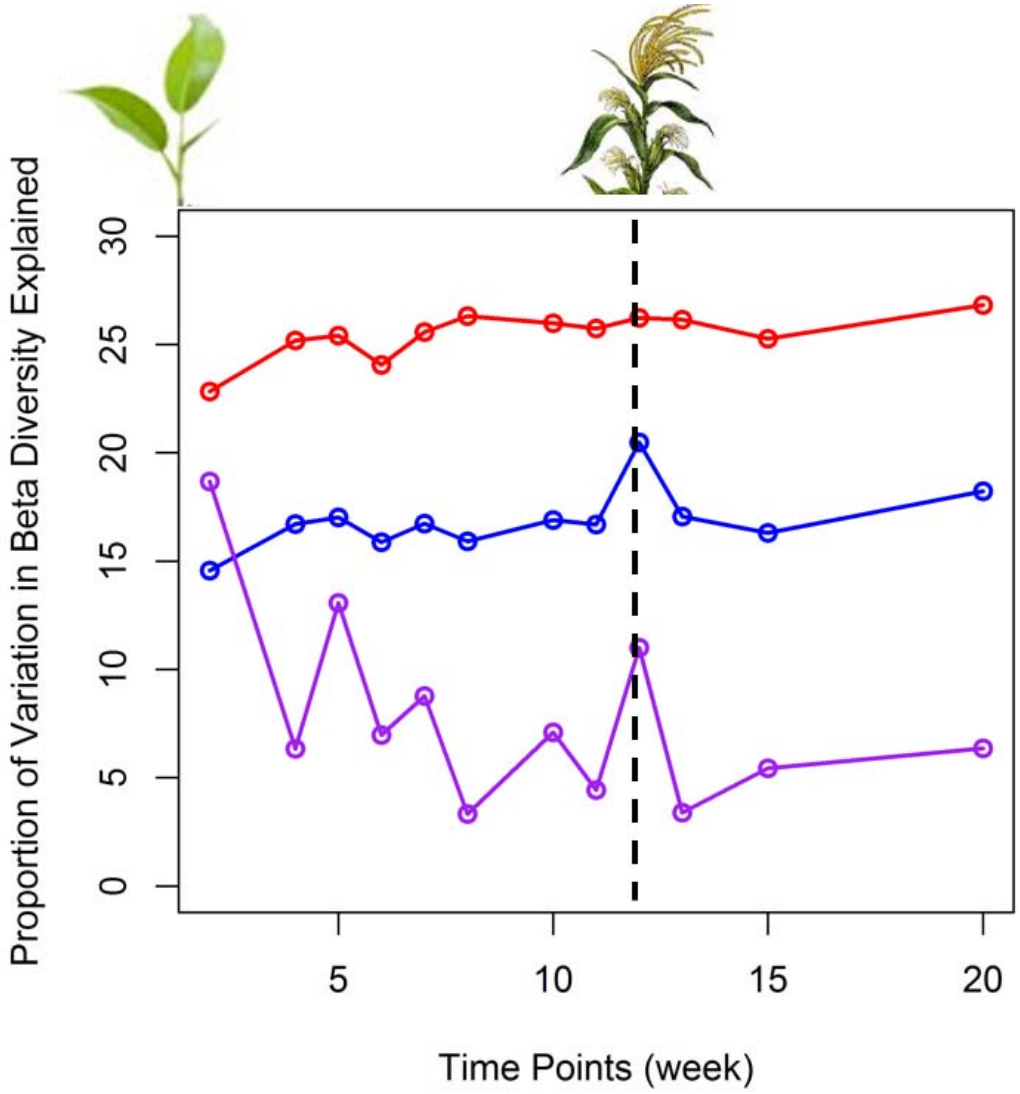

Figure 4.4 Proportion of variation in unweighted UniFrac distances rarefied at 10k explained by maize genotypes (blue), fields (purple), and genotype by field interactions (red). X-axis, maize developmental time points in ages. Y-axis, percentage of variation in beta diversity explained by each factor.

The CAP analysis on the weighted UniFrac distances shares a similar trend to that for the unweighted UniFrac distances, although there are several differences (Figure 4.5). For both the unweighted and weighted UniFrac distances, the maize genotype effect increases at week 2 to week 5, although the increase takes a longer time and does not drop until week 6 for unweighted UniFrac distances. The genotype

effect both peaks at week 12 for the unweighted and weighted UniFrac distances, and increases slightly after week 15. The weighted UniFrac is based on the abundances of taxa and is less sensitive to rare taxa, which may explain the observed small incongruities.

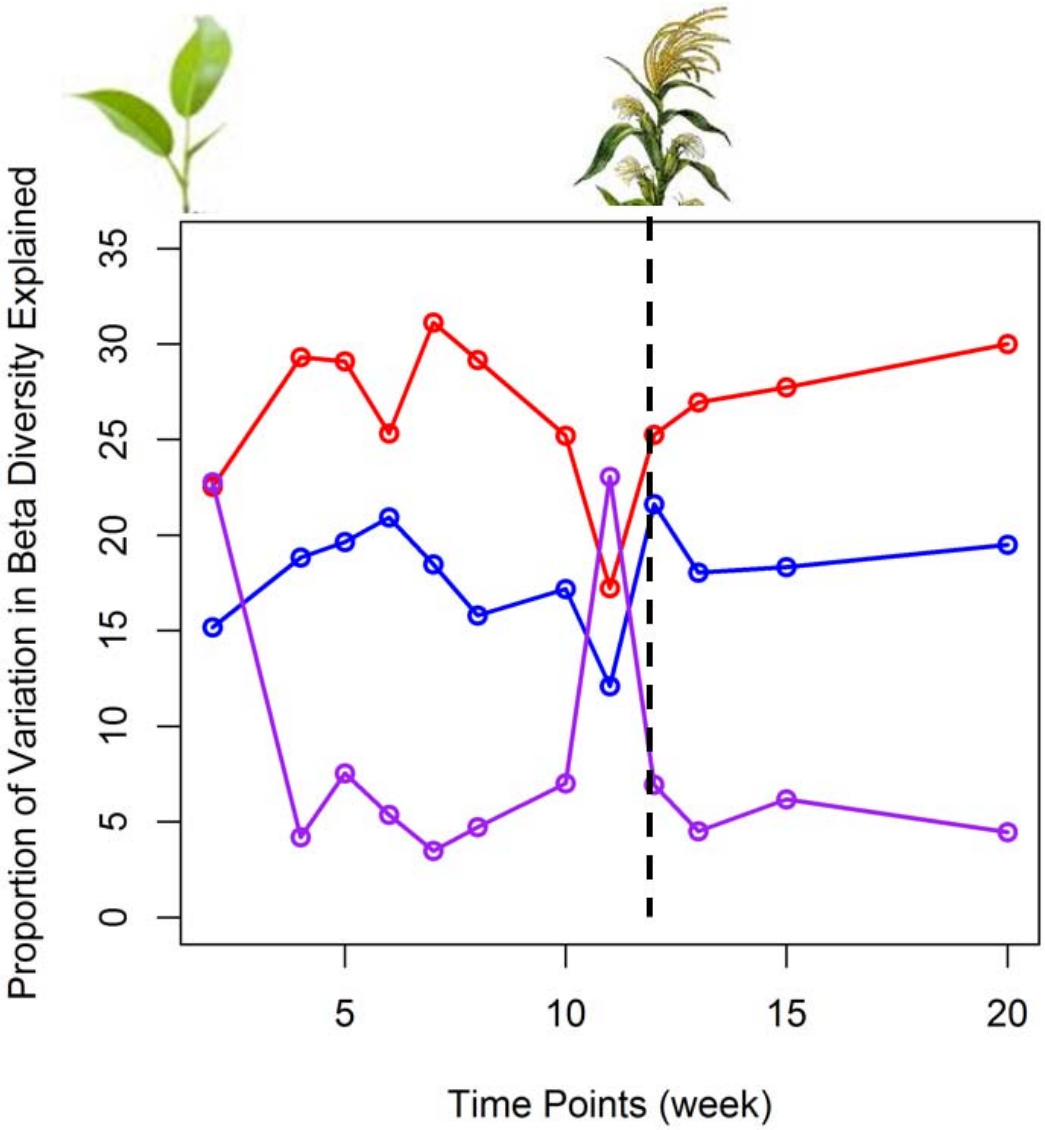

Figure 4.5 Proportion of variation in weighted UniFrac distances rarefied at 10k explained by maize genotypes (blue), fields (purple), and genotype by field interactions (red). X-axis, maize developmental time points in ages. Y-axis, percentage of variation in beta diversity explained by each factor. Note that because rarefaction depth does not influence the temporal pattern, only the results for the UniFrac distances rarefied at 10k are shown here.

To find out whether the maize genetic effect on the rhizosphere microbiome abundance changes with maize development, I first sought to identify the heritable components of the rhizosphere microbiome and how much variation in the microbiome abundance is attributable to maize genotypes at each time point. I took the approach of modeling OTU abundances with regard to maize genotypes, field, genotype by field interactions and other factors, such as shipment and sequencing run. I employed a number of methods to explore the relationship of these factors with the variation in the rhizosphere microbiome OTU abundances, as described below.

To deal with the non-normality in the OTU count data, I first used permutation-based multiple regression to model OTU abundance. Because the original OTU table is sparse, with 10% counts being non-zeroes, I collapsed the OTU table at family level (hereafter referred to as the L6 data). The L6 data of all the maize rhizosphere microbiome samples contain 1261 family-level taxa. I regressed the absolute counts of each taxa in samples with maize genotypes, field, genotype by field interactions, and other covariates, using the total reads per sample as the offset, or regressed the relative abundance of each taxa with these factors without the offset. I corrected for multiple testing, and took the intersect of the significant hits from both calculations. I defined a taxa as having a truly significant maize genotype main effect when the genotype by field interactions are insignificant. I identified 63 taxa as having a significant maize genotype effect, and plotted the relative abundance of these taxa with maize genotypes at all times and by each time point (Figures 4.6 and 4.7). However, I did not observe a distinct pattern in the relative abundances of these taxa with regard to maize genotypes (Figure 4.6A). I also collapsed the 27 maize genotypes into 6 maize subgroups based on their genetic diversity (Liu et al 2003); this did not improve the pattern of the taxa relative abundance across maize

subgroups (Figure 4.6B). Plotting the relative abundances of these taxa with maize genotypes at each time point seems to show the variation in the relative abundances of the taxa across the maize genotypes better, although the patterns for many taxa are still not discernable (Figure 4.7).

Figure 4.6 Heatmaps of potential heritable family-level taxa with maize genotypes (left) and subgroups (right) for all rhizosphere microbiome samples at all time points. Color bar on top of left graph shows the 27 maize genotypes from left to right: B73, B97, CML103, CML227, CML247, CML322, CML333, CML52, CML69, Hp301, Il14H, Ki11, Ki3, Ky21, M162W, M37W, Mo17, Mo18W, MS71, NC350, NC358, Oh43, Oh7B, P39, Tx303, and Tzi8. Color bar on top the right graph shows the 6 maize subgroups from left to right: mixed, non-stiff stalk (nss), popcorn, sweet corn, stiff stalk (ss), and tropical (ts).

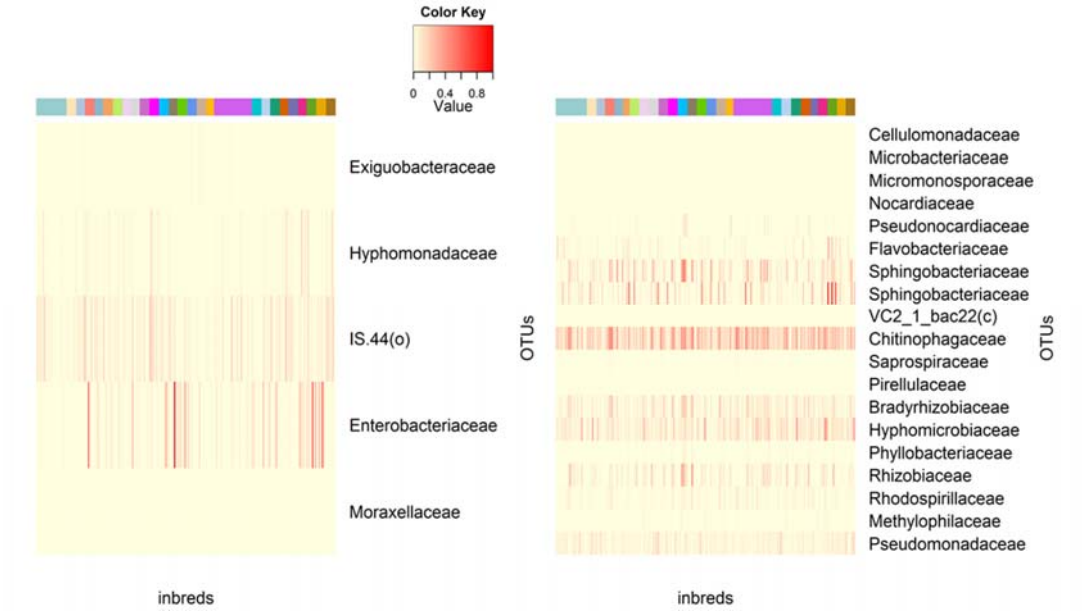

Figure 4.7 Heatmaps of potential heritable family-level taxa with maize genotypes for rhizosphere microbiome samples at weeks 7 and 15. These are shown as an example for visualizing the patterns of potential heritable taxa at each time point. Color bars on top show the 27 maize genotypes as in Figure 4.6A.

I further probed the relationship between OTU abundance and maize genotypes, field and genotype by field interactions using Poisson, negative binomial, and zero-inflated Poisson (ZIP) generalized linear models (GLMs) to account for over-dispersion and zero-inflation in the count data. I modeled both the L6 data, and, after removing any OTU that was present in less than 75% of all the samples to trim excess zeroes, the p75 OTU data (the p75 OTU data contained 785 OTUs). For each of the 12 time points, I stacked the counts of all taxa/OTUs from all samples. I fitted the absolute abundance of the taxa/OTUs to maize genotypes, field, and genotype by field interactions, and diagnosed the fit. There were very significant over-dispersion for Poisson GLMs at all time points. Analysis of variance (ANOVA) for the negative binomial GLMs with and without the genotype by field interactions confirms that the interaction term is significant. The test for goodness-of-fit on the negative binomial

models shows that the negative binomial GLMs is a better fit than the Poisson GLMs for the L6 and p75 abundance data. However, the residual versus fitted plots indicate heteroscedasticity, and the normal quantile-quantile plots show that the residuals deviated from normality at all time points (see Figure 4.8 for an example). The ZIP GLMs, which assume that the zeroes come from a separate process from the non-zero counts, model the excess zeroes independently from a Poisson count model. The residual versus fitted plots indicate heteroscedasticity, and the normal quantile-quantile plots show that the ZIP model residuals have even more departure from normality compared to the negative binomial GLMs at all time points (see Figure 4.8 for an example).

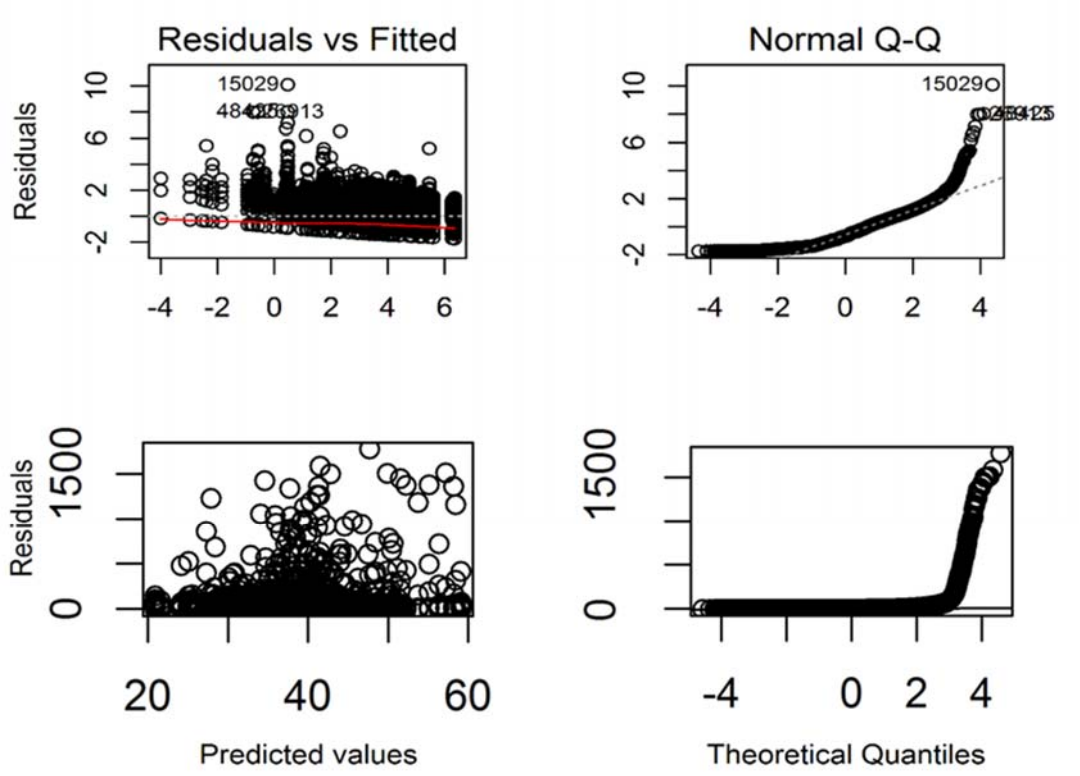

Figure 4.8 Residual versus Fitted (left) and normal quantile-quantile plots (right) for residual of negative binomial (top) and Poisson zero-inflated (bottom) GLMs. These are example diagnostic plots for the negative binomial and ZIP GLMs fitted to the L6 and p75 OTU data at each time point.

To account for the correlation in the OTU abundance data, I fitted general linear mixed models to the relative abundance of the L6 and p75 OTU data at each time point. The field was considered as a fixed effect, whereas the 27 maize genotypes were regarded as a random effect. The general linear mixed models failed to converge within the iteration limit of the models for 8 time points for the L6 data (11 time points for the p75 data), or showed that maize genotypes and/or genotype by field interactions had zero variance in the total variance of the random effects for 9 time points for the L6 data (8 time points for the p75 data). Logarithm or square root transformation did not improve the model fitting. I also applied Poisson generalized linear mixed models (GLMMs) to the absolute abundance of the L6 and p75 data using the total reads per sample as the offset. The Poisson GLMMs also failed to converge for all 12 time points, and showed that maize genotypes/genotype by field interactions had zero variance in the total variance of the random effects for both the L6 and p75 data. Previously, general linear mixed models have been used to detect OTUs differentially enriched in the Arabidopsis rhizosphere relative to bulk soil (Lundberg et al 2012) with the OTU abundance in bulk soil as the fixed effect, and Arabidopsis genotypes as a random effect. The incorporation of the bulk soil OTU abundance fixed effect likely improved the model fitting in this case.

To take into consideration the relationship among the maize genotypes, I explored the possibility of including the maize kinship matrix (Peiffer et al 2013) in modeling OTU abundance using a general linear mixed model. I treated field as a fixed effect, maize genotypes as a random effect, and included the maize kinship matrix as the correlation structure for the random effect. There are two caveats associated with this approach. First, the normal quantile-quantile plots show that the residuals deviate from normality; second, it is not possible to include the interaction

term in the model, which may partly explain the identified lack of fit; thus, the genotype by field interaction cannot be estimated.

I conducted mantel test and Procrustes analysis to examine the relationship between the maize kinship matrix and the beta diversity of the rhizosphere microbiome samples. As the maize genotype effect is strongest in week 12, I focused on comparing the kinship matrix to the unweighted and weighted UniFrac distances for week 12 samples. The kinship matrix was inflated to the same size as that of the UniFrac matrix, and Monte Carlo permutations were applied to the test. The mantel test results cannot reject the null hypothesis that the two matrices are unrelated ($p > 0.1$) for both the unweighted and weighted UniFrac distances. Procrustes analysis with permutations on the maize kinship matrix and the unweighted or weighted UniFrac distances at week 12 cannot reject the null hypothesis that the two configurations are random, either ($p > 0.5$). These results are consistent with our previous conclusion that the maize kinship matrix could not explain the variation in the maize rhizosphere microbiome beta diversity (Peiffer et al 2013).

Influence of time on maize rhizosphere microbiome

Besides exploring the maize genotype effect on the rhizosphere microbiome over time, I also investigated the influence of time on the maize rhizosphere microbiome in general.

I first aimed to model the effect of time along with maize genotypes, field, genotype by field interactions, and other factors, to determine whether time has a significant influence on the variations in OTU abundance. I included time as a factor in the negative binomial GLM to model the stacked p75 OTU data for all samples from all time points. I compared this model to another model without the time factor, and found that time does not have a significant effect on the variation in the overall

OTU abundance. A recent study on maize rhizosphere microbiome in a single maize cultivar sampling four different maize growth stages has observed changes in the relative abundances of certain Proteobacteria or Bacteroidetes taxa in relation to maize developmental stages (Li et al 2014). An earlier research showed that the diversity of culturable rhizosphere bacterial populations did not differ significantly as maize went through five developmental stages, although the abundance of some bacterial taxa varied with time (Cavaglieri et al 2009). Therefore, although I did not find a significant effect of time on the overall rhizosphere microbiome OTU abundance, maize developmental stages may influence the abundance of a subset of the microbiome.

I analyzed the relative OTU abundance for the rhizosphere microbiome samples from each of the 27 maize genotypes to identify the subset of the maize rhizosphere microbiome that vary in abundances over time. I took the intersect of the OTUs whose abundances vary over time from all 27 maize genotypes, and plotted the relative abundance of the 10 most abundant OTUs over the maize developmental time course (Figure 4.9 and Table 4.1). Interestingly, 3 OTUs from the Pseudomonadales order are highly enriched at later time points from week 8 to week 20, suggesting a shift in the maize root exudates at week 8 that attracts more Pseudomonadales to the rhizosphere. Pseudomonadales are r-strategists that populate in nutrition-rich niches (Smit et al 2001). It has been found that as potato plants aged, the amount of ^{13}C-labeled carbon released from roots increased over time, and that *Pseudomonas* and Burkholderia in the potato rhizosphere enriched more ^{13}C-labeled carbon than other bacteria (Dias et al 2013). In addition, two OTUs from the Sphingobacteriaceae family in the Bacteroidetes phylum are slightly enriched in later maize growth stages after week 10. A recent study on Arabidopsis

rhizosphere microbiome has shown that the abundance of Bacteroidetes peaked at flowering time (Chaparro et al 2014), whereas previously we have also found that Sphingobacteriales were enriched in the maize rhizosphere at flowering time (Peiffer et al 2013). Thus, my results are in agreement with previous findings. Notably, this method mostly identified OTUs with varying abundances at later maize growth time points, and is also missing OTUs whose abundances are depleted over time.

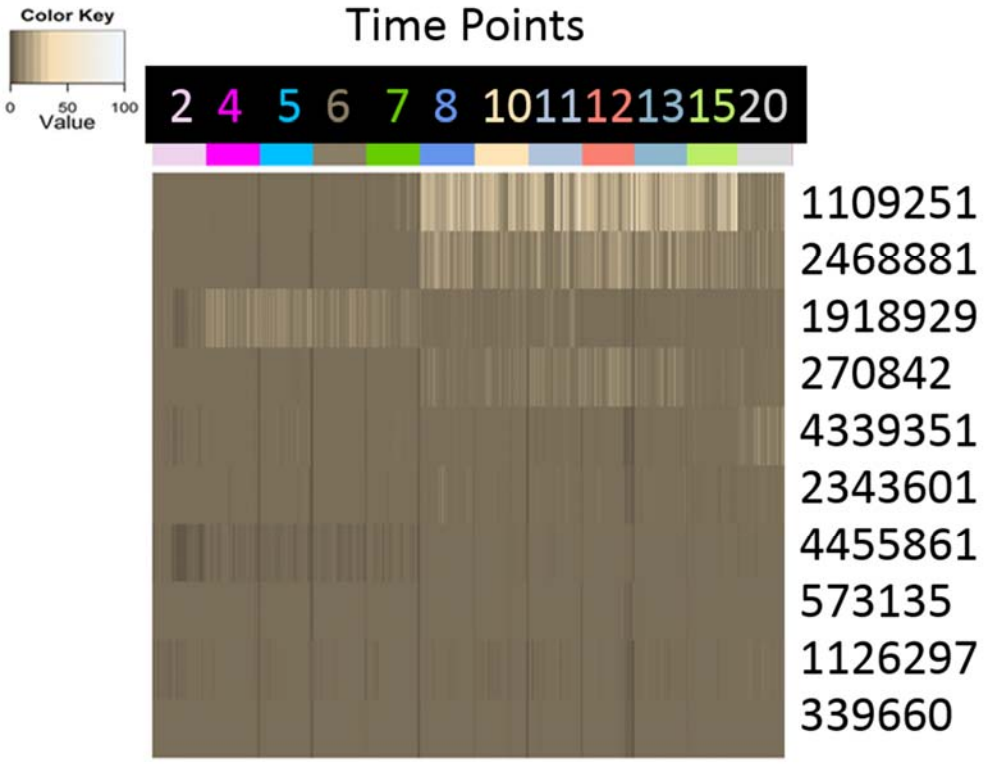

Figure 4.9 Heatmap of OTUs whose abundances vary by time with maize developmental time points for all rhizosphere microbiome samples. Color bar on top shows maize developmental time points from left to right by week. Numbers right next to the heatmap are Greengenes OTU numbers.

Table 4.1 Taxonomy of the OTUs whose abundances varied with maize developmental stages.

1109251	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__*Pseudomonas*; s__
2468881	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Moraxellaceae; g__; s__
1918929	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria; g__; s__
270842	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__*Pseudomonas*; s__
4339351	k__Bacteria; p__Bacteroidetes; c__Sphingobacteriia; o__Sphingobacteriales; f__Sphingobacteriaceae; g__Pedobacter; s__
2343601	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria; g__; s__
4455861	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pseudomonadales; f__Pseudomonadaceae; g__*Pseudomonas*; s__
573135	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhizobiales; f__Bradyrhizobiaceae; g__Bradyrhizobium; s__
1126297	k__Bacteria; p__Bacteroidetes; c__Sphingobacteriia; o__Sphingobacteriales; f__Sphingobacteriaceae; g__Pedobacter; s__

I estimated the effect of time on the unweighted UniFrac distances while controlling for the effect from other factors using CAP analysis. A permutation-based ANOVA on the CAP model reveals that although small (0.3%), the effect of time is

significant ($p < 10e-16$) for the variation in the beta diversity of all rhizosphere microbiome samples. This model-based approach is consistent with the principal coordinates analysis for the unweighted UniFrac distance (Figure 4.10), which shows the clustering patterns of the rhizosphere microbiome samples by time.

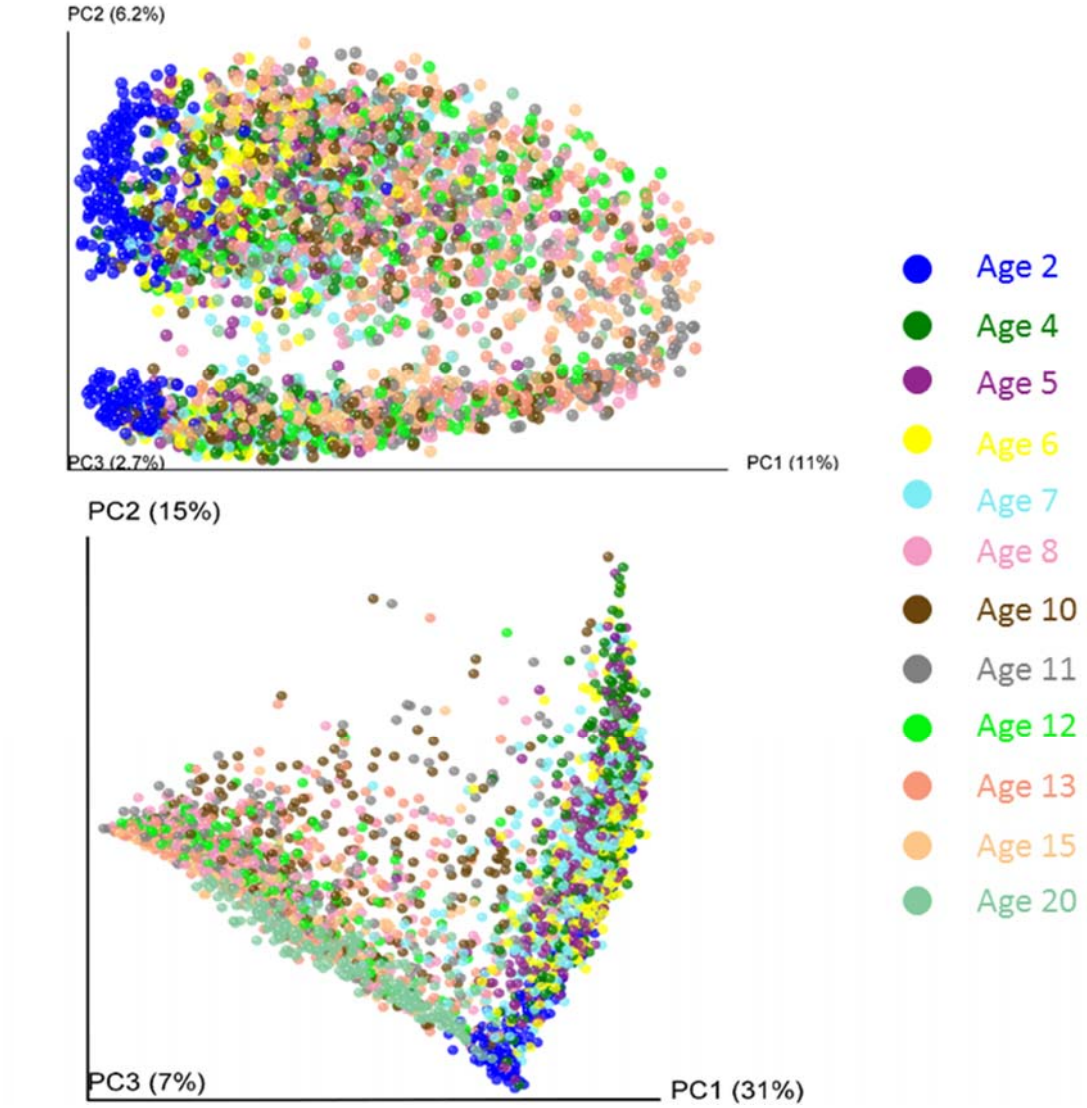

Figure 4.10 Maize rhizosphere microbiome samples clustered using PCoA of the unweighted (top) and weighted (bottom) UniFrac distances. The percentage of variation explained by the principal coordinates is indicated on the axes.

To further investigate the temporal patterns of the maize rhizosphere

microbiome, I sought to explore the variation of the co-occurring OTUs over time. I split the OTU abundance data to each time point, and analyzed the co-occurrence of OTUs by calculating the Pearson correlation of each OTU pair and correcting for multiple comparison. The number of OTU pairs that have significant correlations are huge due to the large number of OTUs. To focus on fewer taxa groups that may exhibit clearer patterns, I used the relative abundance of the L6 and p75 OTU data for all rhizosphere microbiome samples from all time points, and employed hierarchical clustering to detect possible temporal patterns. Hierarchical clustering of the Bray-Curtis similarities of OTUs has been successfully applied to capture the succession patterns of the apple flower microbiome time series data (Shade et al 2013). Hierarchical clustering on the L6 data does not separate the taxa well; over 99% of the total taxa are grouped into one cluster. On the contrary, the p75 OTU data cluster into 6 major groups (Figure 4.11). However, I did not observe any discernible pattern when I plotted the 10 most abundant OTUs in each cluster (see Figure 4.12 for two examples). This suggests that the temporal co-occurrence patterns of the maize rhizosphere microbiome may be more complicated than that can be captured by hierarchical clustering, as the apple flower microbiome starts from a few taxa occupying an almost sterile environment when flowers first open, whereas the maize rhizosphere undergoes constant exchange and competition under field conditions. Alternatively, it is possible that the temporal dynamics of taxa/OTUs may be different for each maize genotype or within each field.

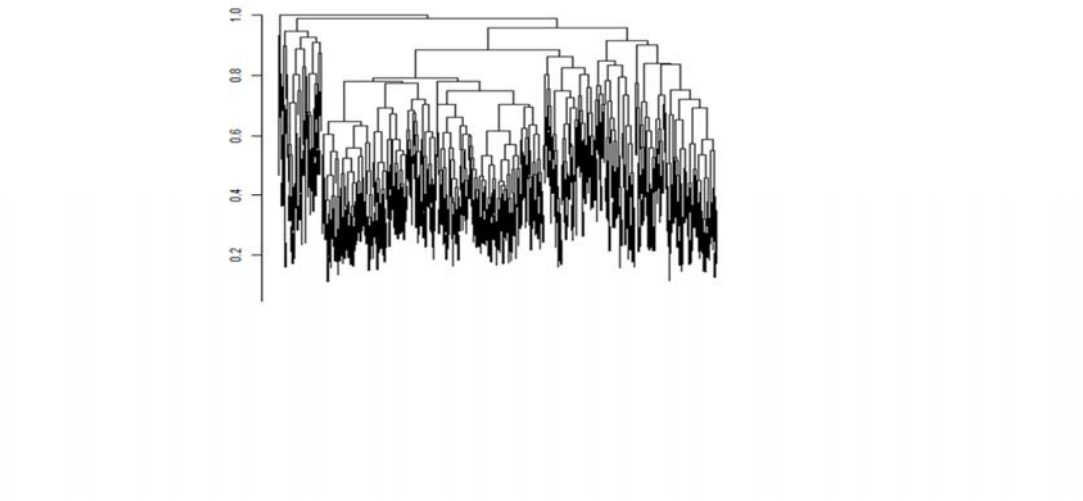

Figure 4.11 Hierarchical clustering (complete linkage-based Bray-Curtis similarities among OTUs defined at 97% sequence) for the relative abundance of the p75 OTU data. Y-axis: within-cluster Bray-Curtis similarity.

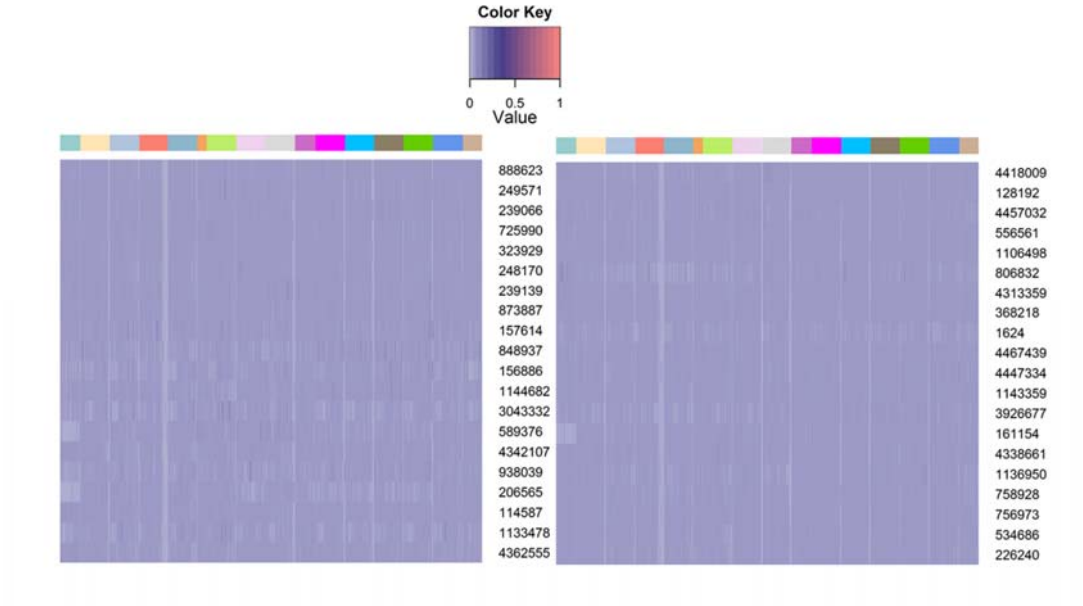

Figure 4.12 Heatmaps of top twenty abundant OTUs in the first two clusters (left to right) from Figure 4.11 with maize developmental time points. Color bar on top shows maize developmental time points from left to right: weeks 1, 10, 11, 12, 13, 14, 15, 16, 2, 20, 3, 4, 5, 6, 7, 8, and 9.

To examine the dynamics of changes in community composition in the rhizosphere microbiome during maize development, I partitioned the beta diversity into the species turnover and nestedness components (Baselga 2010). Species turnover and nestedness refer to the two opposing effects of species replacement and loss, respectively. The calculation was conducted on the presence-absence OTU data, and a monotonic transformation of the Sørensen beta diversity metric was divided into the parts that were due to addition of new community members and changes of constant community members. I found that more species replacement occurs in the maize rhizosphere microbiome than species loss, which is consistent with most of the microbiome members being rare OTUs. I observed an initial species loss at week 2 to week 4 after maize emergence; this corresponds to the increase in maize genetic effect when maize plants starts to select microbial species to establish their rhizosphere microbiome. From week 4 to week 8, members in the rhizosphere microbiome undergo active replacement, suggesting the gradual cultivation of maize on their rhizosphere microbiome is dynamic and involves recruitment of new microorganisms to the rhizosphere, and that old members that no long adapt to changes in the maize root exudates are excluded from the rhizosphere. Species replacement slows down as maize transitions from early to late vegetative growth stages, and species loss is fastest at week 12. This corresponds to the strongest maize genotype effect at week 12. Flowering is controlled by many small-effect quantitative trait locus (Buckler et al 2009). These genetic effects may underlie a relatively bigger change in root exudates at flowering time, which poses a stricter selection on members in the rhizosphere microbiome. Species turnover slows down after week 13, suggesting that the rhizosphere microbiome becomes more similar in later developmental stages.

My results on partitioning the variation in the beta diversity of maize rhizosphere microbiome samples attributable to maize genotypes and partitioning of the beta diversity into species replacement and loss components both indicate that weeks 2 and 4, week 8, and weeks 12 and 13 may be important maize growth time points that have a big impact on the rhizosphere microbiome. Therefore, I aimed to divide maize developmental stages on the basis of the heritable components of the rhizosphere microbiome. I employed statistical learning to discriminate the maize developmental stages using potential heritable family-level taxa identified by the permutation-based regression approach, or OTUs whose abundances vary by time. However, the error rates for the predictions are very high (overall error rates > 80%). I also explored predicting maize genotypes or subgroups using potential heritable family-level taxa. These estimates suffered the same high error rates as the predictions for maize developmental stages. Statistical learning has been used previously to classify ecological data in relation to time (Gilbert et al 2012, Koren et al 2012); however, the high-dimensional data generated by next-generation sequencing may require novel statistical learning approaches such as those described recently (Blagus and Lusa 2013, Gaynanova et al 2014, Lin and Chen 2013).

Conclusion

A number of research articles on microbiome time-series data have emerged during the last several years with the advent of next-generation sequencing technologies (for a review on mammalian microbiome time-series research, see Gerber 2014). Such studies investigated the dynamics of the composition and function of the microbial communities in close association with mammalian and plant hosts, and revealed the important temporal dynamics of the microbiomes. Research on plant microbiomes showed that plant growth stages influenced the diversity or function of the rhizosphere (Chaparro et al 2014, Li et al 2014), endophytic (Lundberg et al 2012, Shi et al 2014), or phyllosphere (Jackson and Denney 2011, Maignien et al 2014, Shade et al 2013) microbiomes.

However, relatively less is known about the dynamics of plant genetic control on the rhizosphere microbiome over plant developmental stages. Compared to studies on single time-point microbiome data, our longitudinal study will answer when the plant genotype effect is strongest over plant developmental stages, and unravel a moving picture of the plant-microbiome interactions during the maize life cycle.

While the search for an optimal modeling approach to determine the heritability of the maize rhizosphere OTUs/taxa and pinpoint truly heritable components is underway, I found that negative binomial GLMs fit the OTU abundance data better than other modeling methods, and that incorporating the maize kinship matrix into modeling OTU abundance may improve the model with the relationship among the maize inbreds being accounted for.

My current analyses also showed that the maize genotype effect is strongest at week 12. In addition, the first two weeks, the mid-life cycle weeks, along with week 12, are possible important time points in maize development that are related to shifts

in the rhizosphere microbiome. Novel statistical learning approaches (Fisher and Mehta 2014, Gaynanova et al 2014), as well as advanced computational analyses for microbiome time-series data (Gerber et al 2012, Marino et al 2014, Stein et al 2013) may be applied to longitudinal microbiome dataset to improve our understanding on the dynamics of the heritability of the maize rhizosphere microbiome and the general temporal-patterns of the rhizosphere microbiome in relation to maize development.

In addition, in the previous chapters of this dissertation, I have shown that the maize rhizosphere microbiome harbors an enormous rich reservoir of functional proteins, and that maize genotypes are associated with certain metabolic genes in members of the rhizosphere microbiome. Recent studies in Arabidopsis and other plants have investigated the functional capacity of the rhizosphere microbiomes over time (Chaparro et al 2013, Chaparro et al 2014, Uksa et al 2014). The implement of Phylogenetic Investigation of Communities by Reconstruction of Unobserved States (PICURSt) (Langille et al 2013) to our maize rhizosphere microbiome OTU abundance data will help reveal whether there are functional heritable components of the rhizosphere microbiome and whether they shift with maize developmental stages.

REFERENCES

Baselga A, Orme D, Villeger S, Bortoli DJ, Leprieur F (2013). betapart: Partitioning beta diversity into turnover and nestedness components. R package version 1.3. http://CRAN.R-project.org/package=betapart

Baselga A (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography* **19:** 134-143.

Baskan O, Kosker Y, Erpul G (2013). Spatial and temporal variation of moisture content in the soil profiles of two different agricultural fields of semi-arid region. *Environmental monitoring and assessment* 185: 10441-10458.

Bates D, Maechler M, Bolker B, Walker S (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-6. http://CRAN.R-project.org/package=lme4

Baudoin E, Benizri E, Guckert A (2002). Impact of growth stage on the bacterial community structure along maize roots, as determined by metabolic and genetic fingerprinting. *Applied Soil Ecology* 19: 135-145.

Benson AK, Kelly SA, Legge R, Ma F, Low SJ, Kim J *et al* (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences of the United States of America* **107:** 18933-18938.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG *et al* (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456:** 53-59.

Berendsen RL, Pieterse CM, Bakker PA (2012). The rhizosphere microbiome and plant health. *Trends in plant science* **17:** 478-486.

Blagus R, Lusa L (2013). Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC bioinformatics* **14:** 64.

Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C *et al* (2009). The genetic architecture of maize flowering time. *Science* **325:** 714-718.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7:** 335-336.

Cain ML, Subler S, Evans JP, Fortin MJ (1999). Sampling spatial and temporal variation in soil nitrogen availability. *Oecologia* 118: 397-404.

Cavaglieri L, Orlando J, Etcheverry M (2009). Rhizosphere microbial community structure at different maize plant growth stages and root locations. *Microbiological research* **164:** 391-399.

Chaparro JM, Badri DV, Bakker MG, Sugiyama A, Manter DK, Vivanco JM (2013). Root exudation of phytochemicals in Arabidopsis follows specific patterns that are

developmentally programmed and correlate with soil microbial functions. *PloS one* **8:** e55731.

Chaparro JM, Badri DV, Vivanco JM (2014). Rhizosphere microbiome assemblage is affected by plant development. *The ISME journal* **8:** 790-803.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K *et al* (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and environmental microbiology* **72:** 5069-5072.

Dias AC, Dini-Andreote F, Hannula SE, Andreote FD, Pereira ESMC, Salles JF *et al* (2013). Different selective effects on rhizosphere bacteria exerted by genetically modified versus conventional potato lines. *PloS one* **8:** e67948.

Dray S, Dufour A-B (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of statistical software* **22:** 1-20.

Fierer N, Bradford MA, Jackson RB (2007). Toward an ecological classification of soil bacteria. *Ecology* 88: 1354-1364.

Fisher CK, Mehta P (2014). Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. *arXiv preprint arXiv:14020511*.

Gaynanova I, Booth JG, Wells MT (2014). Simultaneous sparse estimation of

canonical vectors in the p>> N setting. *arXiv preprint arXiv:14036095*.

Gerber GK, Onderdonk AB, Bry L (2012). Inferring dynamic signatures of microbes in complex host ecosystems. *PLoS computational biology* **8:** e1002624.

Gerber GK (2014). The dynamic microbiome. *FEBS letters*.

Gilbert JA, Steele JA, Caporaso JG, Steinbruck L, Reeder J, Temperton B *et al* (2012). Defining seasonal marine microbial community dynamics. *The ISME journal* **6:** 298-308.

Hansen EE, Lozupone CA, Rey FE, Wu M, Guruge JL, Narra A *et al* (2011). Pan-genome of the dominant human gut-associated archaeon, Methanobrevibacter smithii, studied in twins. *Proceedings of the National Academy of Sciences of the United States of America* **108 Suppl 1:** 4599-4606.

Hastie T, Balasubramanian R, Narasimhan B, Chu G (2013). pamr: Pam: prediction analysis for microarrays. R package version 1.54.1. http://CRAN.R-project.org/package=pamr

Inceoglu O, Salles JF, van Overbeek L, van Elsas JD (2010). Effects of plant genotype and growth stage on the betaproteobacterial communities associated with different potato cultivars in two fields. *Applied and environmental microbiology* **76:** 3675-3684.

Jackson CR, Denney WC (2011). Annual and seasonal variation in the phyllosphere bacterial community associated with leaves of the southern Magnolia (Magnolia grandiflora). *Microbial ecology* **61:** 113-122.

Koren O, Goodrich JK, Cullender TC, Spor A, Laitinen K, Backhed HK *et al* (2012). Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* **150:** 470-480.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA *et al* (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature biotechnology* **31:** 814-821.

Li X, Rui J, Mao Y, Yannarell A, Mackie R (2014). Dynamics of the bacterial community structure in the rhizosphere of a maize cultivar. *Soil Biology and Biochemistry* **68:** 392-401.

Lin WJ, Chen JJ (2013). Class-imbalanced classifiers for high-dimensional data. *Briefings in bioinformatics* **14:** 13-26.

Liu K, Goodman M, Muse S, Smith JS, Buckler E, Doebley J (2003). Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* **165:** 2117-2128.

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

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S *et al* (2012). Defining the core Arabidopsis thaliana root microbiome. *Nature* **488:** 86-90.

Maignien L, DeForce EA, Chafee ME, Eren AM, Simmons SL (2014). Ecological succession and stochastic variation in the assembly of Arabidopsis thaliana phyllosphere communities. *mBio* **5:** e00682-00613.

Marino S, Baxter NT, Huffnagle GB, Petrosino JF, Schloss PD (2014). Mathematical modeling of primary succession of murine intestinal microbiota. *Proceedings of the National Academy of Sciences of the United States of America* **111:** 439-444.

Meng H, Zhang Y, Zhao L, Zhao W, He C, Honaker CF *et al* (2014). Body weight selection affects quantitative genetic correlated responses in gut microbiota. *PloS one* **9:** e89862.

Micallef SA, Channer S, Shiaris MP, Colon-Carmona A (2009). Plant age and genotype impact the progression of bacterial community succession in the Arabidopsis rhizosphere. *Plant signaling & behavior* **4:** 777-780.

Mougel C, Offre P, Ranjard L, Corberand T, Gamalero E, Robin C *et al* (2006). Dynamic of the genetic structure of bacterial and fungal communities at different developmental stages of Medicago truncatula Gaertn. cv. Jemalong line J5. *The New phytologist* **170:** 165-175.

Nelson KE, Mullany P, Warburton P, Allan E (2011). *Metagenomics of the Human Body*. Springer.

Neuwirth E (2011). RColorBrewer: ColorBrewer palettes. R package version 1.0-5. http://CRAN.R-project.org/package=RColorBrewer

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2013). vegan: Community Ecology Package. R package version 2.0-10. http://CRAN.R-project.org/package=vegan

Peiffer JA, Ley RE (2013). Exploring the maize rhizosphere microbiome in the field: A glimpse into a highly complex system. *Communicative & integrative biology* **6:** e25177.

Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL *et al* (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proceedings of the National Academy of Sciences of the United States of America* **110:** 6548-6553.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing.

Shade A, McManus PS, Handelsman J (2013). Unexpected diversity during community succession in the apple flower microbiome. *mBio* **4**.

Shi Y, Yang H, Zhang T, Sun J, Lou K (2014). Illumina-based analysis of endophytic bacterial diversity and space-time dynamics in sugar beet on the north slope of Tianshan mountain. *Applied microbiology and biotechnology*.

Smit E, Leeflang P, Gommans S, van den Broek J, van Mil S, Wernars K (2001). Diversity and seasonal fluctuations of the dominant members of the bacterial soil community in a wheat field as determined by cultivation and molecular methods. *Applied and environmental microbiology* **67:** 2284-2291.

Spor A, Koren O, Ley R (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature reviews Microbiology* **9:** 279-290.

Stein RR, Bucci V, Toussaint NC, Buffie CG, Ratsch G, Pamer EG *et al* (2013). Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS computational biology* **9:** e1003388.

Storey JD (2007). The optimal discovery procedure: a new approach to simultaneous significance testing. *J Roy Stat Soc B* **69:** 347-368.

Therneau T (2012). coxme: Mixed Effects Cox Models. R package version 2.2-3. http://CRAN.R-project.org/package=coxme

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE *et al* (2009). A core gut microbiome in obese and lean twins. *Nature* **457:** 480-484.

Uksa M, Fischer D, Welzl G, Kautz T, Köpke U, Schloter M (2014). Community structure of prokaryotes and their functional potential in subsoils is more affected by spatial heterogeneity than by temporal variations. *Soil Biology and Biochemistry*.

van Overbeek L, van Elsas JD (2008). Effects of plant genotype and growth stage on the structure of bacterial communities associated with potato (Solanum tuberosum L.). *FEMS microbiology ecology* **64:** 283-296.

Venables WN, Ripley BD (2002). Modern Applied Statistics with S. Springer.

Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2014). gplots: Various R programming tools for plotting data. R package version 2.13.0. http://CRAN.R-project.org/package=gplots

Wheeler B (2010). lmPerm: Permutation tests for linear models. R package version 1.1-2. http://CRAN.R-project.org/package=lmPerm

Zhang CH, Wang ZM, Ju WM, Ren CY (2011). Spatial and temporal variability of soil C/N ratio in Songnen Plain maize belt. *Huan jing ke xue* **32:** 1407-1414.

Zeileis A, Kleiber C, Jackman S (2007). Regression models for count data in R. Journal of Statistical Software 27(8). http://www.jstatsoft.org/v27/i08/.