THE UNIVERSITY OF CHICAGO

SPATIOTEMPORAL PATTERNS IN THE COLONIZATION OF A*RABIDOPSIS THALIANA* BY NATIVE BACTERIAL COMMUNITIES AND PATHOGENS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY

KATHLEEN ROSE BEILSMITH

CHICAGO, ILLINOIS

DECEMBER 2020

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

While I hope this dissertation offers some insight to those who follow me in trying to understand how plant microbiomes assemble and influence their hosts, a scientific mindset requires recognizing, as others have noted before, that all this work may represent merely a more exact answer to the wrong question or a collection of facts still waiting to be set in order. All that is certain is that a new generation will follow me in attempting to answer these questions. In that light, I want to acknowledge that some of the most valuable and lasting work of my time at Chicago is not in the chapters that follow. Rather, it took place in the classroom.

Practical lab courses are at the heart of teaching the tools of scientific inquiry to young scholars. I want to thank Alison Hunter, Christine Andrews, and Mike LaBarbera for showing me how to manage this type of instruction effectively. I am grateful to Cathy, Tracy, and Linda in the UChicago Writing Program for their insights on teaching students how to critically evaluate arguments while also creating an inclusive classroom where students feel confident speaking their minds. Panel discussions, workshops, and pedagogy courses from the Chicago Center for Teaching have helped me to further reflect on my teaching and better share my knowledge. Finally, I am grateful to Ashley, Khari, and Guy for their work on the Paradigm Shift outreach pilot at the University and to the ASPIRA and iMentor programs for letting me work with students in the Chicago area during graduate school.

When working with students, I reflect on the public school teachers who laid a solid foundation

for my science education. My earliest memories of school include Mrs. Ward carefully placing pH indicator slips into the chubby hands of me and a dozen other kindergarteners as we marched outside to examine playground puddles after each rainstorm. In this way, science was introduced to me – before I could even read – as a means by which people can answer questions about the world together. If we aspire to create a more scientifically literate country, then we must continue to support teachers like Mrs. Ward and to create incentives for brilliant and compassionate people like her to lead classrooms across America.

My specific interest in ecology was kindled by annual Envirothon competitions supported by my state's Department of Conservation. My high school's team was led by our biology teacher Mr. Barton and our chemistry teacher Dr. Bowman. These dedicated teachers drove us to public parks to take nature hikes and participate in stream clean-ups, teaching us to identify trees and sample water quality. The competitions required us to learn a great deal of material but also, importantly, required us to apply our knowledge to solve problems of natural resource management. We were also challenged to propose and defend those solutions to a variety of audiences. If we want a more sustainable society, then it is important to recognize the role of environmental science and ecology education as an investment that drives young people to become advocates and innovators in this area.

My path to a college degree in science was opened by programming at my local public university. Graduate students from the University of Missouri used our public school and several others in the area as sampling sites in a project on urban bird ecology, recruiting students to help

net birds and monitor nest boxes. The Students and Teachers As Research Scientists program allowed me to spend a summer conducting research in a lab and learning about careers in science. Nicholas Barber supervised my training and encouraged me to pursue science after high school. If we aspire to produce a more diverse cohort of PhDs, then we must continue to create and fund enriching programs like these and work to make them accessible to all students.

While our public institutions can improve student access to science education, success for those reaching the graduate level depends critically on consistent support and thoughtful teaching and mentoring by faculty and staff. Several faculty and staff have, in different ways, made a continuing career in scientific research more accessible to me. Joy Bergelson made space for me in her lab as a work-study student in college and later as a graduate student, consistently presenting me with opportunities to work on new datasets and collaborate with peers at UChicago and beyond. In her lab, my work was supported on a daily basis by Tim Morton's boundless enthusiasm and tireless work to manage the required equipment and supplies. Greg Dwyer and Stefano Allesina taught courses in theoretical ecology and scientific computing that enabled me to eventually expand my skills from field and bench ecology to computational biology. That transition was spurred by Sam Seaver, who has been an encouraging and engaged mentor and friend to me over the past two years as I finished the work in this dissertation.

I reserve my final words of thanks for my first teachers, my parents and siblings, as well as my favorite classmate and co-instructor, Michael Florian.

– Kat Beilsmith, 2020

INTRODUCTION

I. How do plant microbiomes assemble?

Plants provide multiple habitats for microbial life. Their roots recruit microorganisms to the surrounding soil, or rhizosphere, by exuding carbon compounds. Above ground, in the phyllosphere, their photosynthesizing tissues are colonized by a subset of microbes from the rhizosphere as well as those deposited by the the surrounding air and rainfall. While many microbes live as epiphytes on the surface of plant tissues, others reside within host tissues as endophytes. Endophytic microbes can benefit the host or be agents of disease. Regardless of whether they help or harm, these microorganisms can influence the growth and nutrient cycling of their plant hosts and are therefore variables of interest in both agricultural and ecological studies. While a great deal of foundational work has characterized the collections of microbes inhabiting plant tissues around the world, a synthesis with principles of community ecology is now required in order to understand how these plant microbiomes assemble. With a better understanding of how microbiomes form, we hope to learn how to cultivate or engineer plants to recruit microbial partners that can enhance growth or deter pathogens.

This dissertation addresses basic questions about microbiome assembly in plants-- the same questions that ecologists seek to answer about how microbial communities form in the soil, on our skin, and in our guts. Namely,

- How do host and environmental factors lead to habitat filtering in the microbiome?
- How repeatable is the process of microbiome assembly?
- How do interactions among microbes shape the microbiome over time?
- How can we find the host genes that shape the microbiome?
- How does microbiome assembly change across host generations?

To address these questions, I present results from descriptive studies on the native microbiome of the model plant species *Arabidopsis thaliana* in the midwestern United States as well as controlled experiments using isolates of its *Pseudomonas* pathogens from that environment. The following section provides some necessary background on this study system.

II. Bacterial colonists and pathogens of *A. thaliana*

i. The model plant *Arabidopsis thaliana*

Arabidopsis thaliana, a small weed of the mustard family Brassicaceae, grows in disturbed habitats or sandy, rocky soils where nutrients are limited (Mitchell-Olds 2001, Mitchell-Olds and Schmitt 2006). The work that follows involves two *A. thaliana* genetic backgrounds: the midwestern haplogroup HPG-1 (Chapters 1 and 2) and the reference lineage Columbia-0 (Col-0) (Chapters 4 and 5). Whereas the Col-0 lineage completes its entire life cycle within about six weeks, midwestern *A. thaliana* germinate in the fall, overwinter at the rosette stage, and then rapidly flower in the spring. After their flowering stems emerge, the plants typically self-

pollinate (Tsuchimatsu et al. 2017) and produce thousands of genetically identical seeds that shed when the siliques holding them rupture.

The same hardiness, self-fertilization, copious seed production, and short generations that allow *A. thaliana* to pioneer harsh niches have established it as a model species for discovering the genetic basis of plant responses to environmental pressures (Krämer 2015). Its reference genome, together with large libraries of genotyped inbred lines, permit genome-scale studies on the role of natural variation at these loci (Atwell et al. 2010, Brachi et al. 2010). Recently, these resources have been leveraged to discover the genes associated with biotic interactions between *A. thaliana* and the microbes inhabiting its tissues (Horton et al. 2014, Frachon et al. 2019). Chapter 3 discusses the design of genome-wide association studies (GWAS) to find the plant genes shaping different features of the *A. thaliana* microbiome.

ii. The *A. thaliana* microbiome

Wild *A. thaliana* are colonized by rich assemblages of microorganisms including bacteria, fungi, oomycetes, and protists. *A. thaliana* tissues often harbor a taxonomically narrow set of Actinomycetales, Burkholderiales, and Flavobacteriales bacteria compared to the surrounding soil (Schlaeppi et al. 2014). *A. thaliana* fungal microbiomes are enriched for Pleosporales, Helotiales, Olpidiales, and Cantharellales, while classifiable oomycota are dominated by Phythiales, Peronosporales, and Saprolegniales in natural samples (Maciá-Vicente et al. 2020). Cercozoa protists, which prey on other microbes, are also found in *A. thaliana* leaves (Ploch et

al. 2016). Although these surveys show a variety of microbiota capable of flourishing in the plant niche, this dissertation will focus on only the bacterial component of the *A. thaliana* microbiome.

Plant-associated microbes can be surveyed by sequencing the variable regions of broadly conserved marker genes, such as 16S ribosomal RNA (16S rRNA) in bacteria (Gohl et al. 2016). DNA nucleotide sequences from the marker gene are amplified by polymerase chain reaction (PCR) to produce templates for sequencing by synthesis (SBS). Unique nucleotide sequences at the marker gene are taken to indicate the presence of distinct microbial lineages, which are classified based on their similarity to those in public databases. The abundance of each microbial taxon relative to others in a sample is calculated from the number of sequencing reads assigned to it. Relative abundance is quantified as opposed to absolute abundance because microbial cell counts are difficult to recover from plant samples and the number of marker gene copies can vary between microbial genomes.

iii. *Pseudomonas* pathogens of *A. thaliana*

A fraction of *A. thaliana*'s microbial colonists are pathogenic, causing yellowing or necrotic lesions on leaf tissue. Among the best-characterized pathogens of Brassicaceae are members of the *Pseudomonas* bacterial species complex. Pathogenic lineages of *Pseudomonas* are widespread in natural populations of *A. thaliana* (Bartoli et al. 2018, Frachon et al. 2019). In addition, they afflict multiple wild and cultivated plant species (Morris et al. 2019). For example, *P. syringae* is responsible for the reduction in tomato yields caused by bacterial speck disease

(Scofield et al. 1996). The last two chapters of this dissertation use data from *A. thaliana* infected with isolates of *Pseudomonas* pathogens.

During infection, several properties of plant tissues are altered by pathogen activities and the immune responses they trigger. Once physical barriers like cutin and cell junctions are breached (Nawrath et al. 2013, Serrano et al. 2014), plants rely on immune receptors to detect microbes. Immune receptors bind exogenous patterns associated with microbes, such as bacterial flagellin (Chinchilla et al. 2006), and initiate intracellular signaling that limits pathogen growth by altering gene expression, hormone levels, and reactive oxygen species in plant tissues. At sites where pathogens are detected, appositions reinforce the cell wall and guard cells close the stomata (Underwood 2012, Malinovsky et al. 2014, Melotto et al. 2017). *Pseudomonas* pathogens, in turn, inject plant cells with phytotoxins and secreted effector proteins that interfere with host immune responses (Nomura et al. 2011, Hurley et al. 2014, Zhang et al. 2010). Although plant immune-related proteins and pathogen effectors have long been used as case studies in host-pathogen evolution, metabolism is now understood as another key interface in pathogenesis. Increasingly, genome-scale datasets reveal how metabolism is reconfigured in response to plant infection. Chapters 4 and 5 focus on this metabolic facet of *A. thaliana*'s encounters with *Pseudomona*s pathogens.

III. Outline of the results in this dissertation

In Chapter 1, I use a unique dataset to describe spatiotemporal variation in the endophytic

microbiome of *A. thaliana* grown in the midwest. Whereas previous studies describe the microbiome in just one plant tissue or at just one developmental stage, this dataset profiles the bacteria that colonize every tissue habitat emerging in the plant throughout its life cycle. I find differences in microbiome composition between tissues that likely reflect the co-occurrence of ecologically similar species resulting from selection on different bacterial traits in these plant habitats. As microbiomes assemble over time, the strength of this habitat filtering increases, as does the diversity of the bacterial colonists. Because the dataset includes samples from two field sites over two years, I can identify a fraction of the microbiome that is repeatedly most successful in colonizing the root habitat patch. However, the associations between bacteria and their tissue habitats depend upon the taxonomic level at which I group the identifying DNA sequences.

In Chapter 2, I examine the role of microbe-microbe interactions in microbiome assembly by constructing species interaction networks from the dataset described in Chapter 1. The networks, representing bacteria as nodes and their inferred interactions as edges, are constructed for successive plant developmental stages and compared in order to learn how the structure of community interactions and the roles of species differ over time. The networks reveal interactions and indicators of species influence are transient during the lifetime of the host plant, regardless of the method used to infer interactions from marker gene abundance data.

In Chapter 3, I discuss when the microbiome phenotypes presented in Chapters 1 and 2 can be employed in genome-wide association studies (GWAS) to identify the host genes driving

microbiome assembly. The number and identity of plant loci associated with the microbiome will depend upon the choice to consider specific pairwise or diffuse community-level interactions, the choice to characterize microbes by taxonomic grouping or by traits, and on the extent to which the environmental plasticity of host-microbe interactions is exposed in the sampling scheme.

Chapter 4 narrows in scope from the *A. thaliana* microbiome to pathogenic microbes. One plant system perturbed by pathogens during infection is primary metabolism, or the set of biochemical reactions critical for cell growth and maintenance. I construct a network representing primary metabolites (nodes) and genes assigned to enzymatic functions (edges) from the *A. thaliana* genome. Gene-level expression data from infected and control plants is then integrated with the network to identify the biochemical reactions responsive to pathogen exposure. When applied in parallel to expression data from crops like *Solanum lycopersicum* (tomato), this approach reveals a common set of reactions responsive to *P. syringae* infection in diverse plant hosts.

In Chapter 5, I use a unique set of Col-0 lineages with different infection histories to test whether prior pathogen exposures alter host primary metabolism. The lineages have different histories of exposure to two *Pseudomonas* pathogens isolated from wild *A. thaliana.* The plants in the study include the infected offspring of naïve parents, the naïve offspring of infected parents, and a completely uninfected control lineage. I integrate gene expression data from plants with different infection histories with the *A. thaliana* metabolic model from Chapter 4 in order to generate predictions of biochemical activity in *Arabidopsis* primary metabolism. The predictions of reaction flux generated with naïve and pathogen-exposed plant transcripts reveal overlapping

metabolic responses to two different *Pseudomonas* pathogens in isolation and during co-infection. Notably, similar flux changes predicted for the naïve progeny of infected parents suggest that metabolic responses to pathogens are maintained across plant generations. This result raises the question of whether infection-induced metabolic changes influence microbiome assembly over multiple plant generations.

IV. How to reproduce the analysis in this dissertation

The nucleotide sequencing data downloaded and generated for this work are available in the National Center for Biotechnology Information's Sequence Read Archive (NCBI SRA). DNA sequences from 16S rRNA used in Chapters 1 and 2 are stored under NCBI SRA BioProject ID PRJNA607544. In Chapter 4, genome-wide mRNA datasets in the literature are downloaded from BioProject IDs PRJNA156671, PRJNA214335, PRJNA315516, and PRJEB21223. Genome-wide mRNA sequences generated for Chapter 5 are stored under BioProject ID PRJNA613833.

My annotated analysis scripts for Chapters 1 and 2, including lists of all the publicly available R (version 3.4.4) packages used therein, are available from my GitHub account in the folders listed below. My analysis for Chapters 4 and 5 was performed in Jupyter-based Narratives published in the Department of Energy Systems Biology Knowledgebase (KBase). KBase accounts are free for any user and the software used in KBase Narratives is publicly available through GitHub at the links below.

Chapter 1: github.com/krbeilsmith/KBMP2020_Microbes

Chapter 2: github.com/krbeilsmith/KBMP2020_Networks

KBase: www.kbase.us/, github.com/kbase and github.com/kbaseapps

Chapter 4: narrative.kbase.us/narrative/73012

Chapter 5: narrative.kbase.us/narrative/73121

CHAPTER 1

Natural bacterial assemblages in *Arabidopsis thaliana* tissues become more distinguishable and diverse during host development

CO-AUTHORS & CONTRIBUTIONS

Matthew Perisin and Joy Bergelson designed the field study described in this chapter. Matt planted, collected, processed, and sequenced the samples with help from other members of the Bergelson lab. I wrote the analysis scripts, produced the figures, and wrote the text with input from Matt and Joy. This work was supported by NIH grant R01 GM083068, the University of Chicago Biological Sciences Division, and NIH T32 GM07197. Computing resources were provided by The Center for Research Informatics at the University of Chicago.

We thank Dave Francis at the Michigan State Southwest Michigan Research and Extension Center for providing and prepping field plots. We thank Carlos Sahagun for assisting with planting and Timothy Morton, Benjamin Brachi, Talia Karasov, Manfred Ruddat, and Roderick Woolley for assistance with sampling. The DNA extraction and 16S amplicon sequencing protocols were developed in collaboration with Benjamin Brachi. Alison Anastasio provided additional experimental design advice. Members of the Bergelson lab and the Department of Ecology and Evolution at the University of Chicago, in particular Caroline Oldstone-Jackson and Brooke Weigel, provided valuable feedback during the data processing and analysis.

ABSTRACT

Developing synthetic microbial communities that can increase plant yield or deter pathogens requires basic research on several fronts, including the efficiency with which microbes colonize plant tissues, how plant genes shape the microbiome, and the microbe-microbe interactions involved in community assembly. Findings on each of these fronts depend upon the spatial and temporal scales at which plant microbiomes are surveyed. To study the spatial and temporal dynamics of bacterial colonization under field conditions, we planted and sampled *Arabidopsis thaliana* during two years at two Michigan sites and surveyed colonists by sequencing 16S rRNA gene amplicons. In our study, phyllosphere tissues housed increasingly distinct microbial assemblages as plants aged, indicating that plants can be considered as collections of tissue habitats in which microbial colonists – natural or synthetic – establish with differing success. Although assemblages primarily varied between roots and shoots, amplicon sequence variants (ASVs) also differentiated phyllosphere tissues. Increasing assemblage diversity indicated that variants dispersed more widely over time, decreasing the importance of stochastic variation in early colonization relative to tissue differences. Relationships between host genes and community diversity might vary depending on when samples are collected, given that assemblages grew more diverse as plants aged. Both spatial and temporal trends weakened when colonists were grouped by family, suggesting that functional rather than taxonomic profiling will be necessary to understand the basis for differences in colonization success.

INTRODUCTION

As plant tissues emerge and grow, new habitats are created for microbial colonists (Hardoim et al. 2015, Khare et al. 2018). While they represent only a fraction of colonist diversity (Kowalchuck et al. 2010, Bergelson et al. 2019), bacterial endophytes are important to host plant fitness because of their potential to affect nutrient uptake (Rodríguez and Fraga 1999, Franche et al. 2009, Marulanda et al. 2010), stress responses (Fernandez et al. 2012), and defenses against pathogens (Mercado-Blanco et al. 2004, Kloepper and Ryu 2006). Given these activities, natural and engineered bacterial communities have been proposed as tools for sustainably enhancing plant growth and stress resistance (Berg et al. 2017). When lineages in these communities are pathogenic or beneficial to the host, the efficiency with which they enter plant tissue is of particular interest (Laine and Hanski 2006, Kuiper et al. 2001, Silva et al. 2003). Despite stochasticity in colonization (Obadia et al. 2017, Lazzaro et al. 2017, Maignien et al. 2014), there is evidence that plants selectively filter the bacteria colonizing the intercellular space in their tissues. Characterizing spatial and temporal variation in this filtering is key to understanding how natural and cultivated communities assemble in the endosphere.

The idea that plant tissue filters bacterial colonists is supported by the observation that endophytic communities display only a fraction of the diversity found in soil. For example, the diversity of taxa found in the root endosphere is lower than in rhizosphere soil (Fitzpatrick et al. 2018). Furthermore, a subset of bacterial families is found at higher relative abundance in roots than in soil (Bulgarelli et al. 2015). Filtering is likely due in part to differences between soil and

root cell walls as substrates for colonization, as indicated by similarities between communities in live roots and wood slivers exposed to the same field-collected soil inocula (Bulgarelli et al. 2012). Filtering by living tissue may also involve selection for or against specific bacterial lineages, as indicated by community members enriched over soil levels in roots but not in wood sliver samples and vice versa (Bulgarelli et al. 2012).

Although plant tissue appears to filter bacterial colonists, we do not yet fully understand how spatial and temporal variation influences this process in natural environments. Variation in abiotic factors and the pool of soil colonists at planting sites can influence the efficiency with which bacterial lineages enter plants, leading to associations between geographic location and the composition of bacterial assemblages in plant tissues (Mighell et al. 2019, Lundberg et al. 2012, Wagner et al. 2016). Within plants, recent evidence suggests that communities in different tissues are composed of a common pool of systemic colonists (Massoni et al. 2019, Grady et al. 2019). However, individual lineages can display differences in colonization efficiency between roots and stems (Germaine et al. 2004) and several studies report variation between the bacterial assemblages found in different tissues (Coleman-Derr et al. 2016, Amend et al. 2019, Bodenhausen et al. 2013), supporting the idea that some bacteria are more successful than others in colonizing a given habitat within the plant. Variation in assemblage composition is also observed when plant tissues are sampled at different developmental stages (Lundberg et al. 2012, Chaparro et al. 2014, Leff et al. 2015). These temporal trends could be related to the time available for bacterial colonization of plant tissues before sampling, changes in how hosts filter colonists with age, or interactions arising as more bacteria cooperate or compete within the host.

Since bacteria alter plant tissues upon arrival, the host response to established colonizers could also change the efficacy of colonization later in development (Beattie et al. 1999, Brandl and Lidow 1998).

Since most surveys of plant colonists have either focused on a single tissue or taken only a snapshot of community composition in time, it is difficult to compare the extent and interaction of geographic, tissue-level, and temporal effects on plant endophyte filtering. Adding to the body of work exploring host plant control of colonization, we compared the relative influence of plant tissue type, age, harvest site, and year on the bacteria that naturally colonized a common haplotype of an annual plant, *Arabidopsis thaliana*. To understand the distribution of the bacterial traits driving these patterns, we examined whether relationships between variables and assemblage composition depended on the taxonomic level at which the surveyed colonists were grouped. To ascertain the consistency of spatial and temporal colonization patterns, we compared the tissues and ages at which bacterial lineages reached maximum prevalence or abundance between sites and years. In addition, we characterized how the diversity and evenness of colonists changed across plant tissues throughout development.

MATERIALS AND METHODS

Planting

Field experiments were replicated over two years (2012-14) at two locations: Michigan State Southwest Michigan Research and Extension Center (ME) and University of Chicago Warren

Woods Ecological Field Station (WW). Prior to planting in October, fields were tilled and grids were created with bottomless plastic pots (6-12 cm across) placed 2-5 cm into the ground and 10-30 cm apart. Within each grid, seeds for seven midwestern *A. thaliana* ecotypes were sown randomly and a fraction of pots were left empty for soil sampling. Seeds were surface-sterilized with ethanol and seedlings were thinned after germination with sterilized tweezers.

Sample collection

Plant sampling order was randomized and all tools were flame-sterilized with ethanol between samples. Root and above-ground tissues were separated into sterile plastic tubes. For soil samples, sterile tubes were pushed 2-5cm into the ground. Tubes were stored at -80 C until processing. To remove loosely associated microbes, each plant sample was washed twice by vortexing with surfactant buffer (Lundberg et al. 2012). Plant samples were then transferred to Matrix tubes (Thermo Scientific, Waltham, MA, USA). Above-ground tissue was first separated into compartments with a scalpel and tweezers. For large tissues, only enough material was added to allow for bead homogenization. For soil, samples were put through a 2 mm sieve and ~100 mg was transferred to a Matrix tube. The tubes were randomized in 96-well racks with respect to sampling site, year, and timepoint. To dry the material, tubes were frozen to -80 C and lyophilized overnight. To powder the tissue, sterile silica beads were sealed into each tube with a SepraSeal cap (Thermo Scientific) and tubes were shaken on the 2010 Genogrinder (SPEX, Metuchen, NJ, USA) (1750 RPM, 2 min). Dry mass was recorded and up to 36 mg of material was retained per tube. All tubes were then randomized in Nunc 96-deepwell plates (Thermo Scientific) for DNA extraction.

DNA extraction

Ground material was resuspended in TES (10 mM Tris-Cl, 1 mM EDTA, 100 mM NaCl) to a concentration of 0.04 mg/µL. Material was homogenized with the Genogrinder (1750 RPM) and homogenates (240 µL) were incubated (30 min) in new plates with lysozyme solution (Epicentre, Madison, WI, USA) at a final concentration of 50 U/µL. Proteinase-K (EMD Millipore, Billerica, MA, USA) and SDS were added to final concentrations of 0.5 mg/mL and 1%, respectively. Plates were incubated at 55 C for 4 h. An equal volume of 24:1 chloroform:isoamyl alcohol was mixed by pipette in each well. Plates were centrifuged at 6600xg with the Beckman Coulter Avanti J-25 (Beckman Instruments, Munich, Germany) for 15 min at 4 C. The top aqueous layer (350 µL) was removed and added to new plates with 500 µL 100% isopropanol. Plates were inverted to mix and incubated 1h at -20 C. After centrifugation for 15 min at 4 C, isopropanol was removed and DNA pellets were washed with 500 µL 70% ethanol. Pellets were dried in a chemical hood and resuspended in TE (100 µL, 10 mM Tris-Cl, 1 mM EDTA) by shaking. After incubation on ice for 5 min, plates were centrifuged for 12 min at 4 C and supernatants diluted 10X in TE were added to new 0.5 mL plates for PCR amplification.

16S rRNA gene amplification

The V5, V6, and V7 regions of the 16S rRNA gene were amplified from each sample using the 799F and 1193R primers with Illumina MiSeq adapters, and custom pads, linkers, and barcode sequences (Kozich et al. 2013). The PCR volume was 25 µL: 1 µL of 10X diluted DNA template, 0.2 µM of each primer, 1X 5PRIME HotMasterMix (5PRIME, Gaithersburg, MD, USA), and

0.8X SBT-PAR additive (Samarakoon et al. 2013). PCR amplification consisted of initial denaturation at 94 C for 2min, followed by 35 cycles of denaturation at 94 C for 30 s, annealing at 54.3 C for 40 s, and elongation at 68 C for 40 s, followed by a final elongation at 68 C for 7min. Each PCR was completed in triplicate, pooled, and purified with an equal volume of Axygen AxyPrep Mag PCR Clean-Up bead solution (Corning, Tewksbury, MA, USA). Amplicon concentrations were quantified by fluorimetry (QUANT-iT PicoGreen dsDNA Assay Kit, Life Technologies, Carlsbad, CA, USA) and 30 ng or a maximum of 30 μL per sample were pooled for six plates per sequencing run. Primer dimers and mitochondrial amplicons were removed by concentrating each amplicon pool 20X (Savant SPD121P SpeedVac Concentrator, Thermo Scientific) and purifying 300-700bp product with BluePippin (Sage Science, Beverly, MA, USA).

Sequence data

Amplicon pools were sequenced using the Illumina MiSeq platform and MiSeq V2 Reagent Kits (Illumina, San Diego, CA, USA) to produce paired-end 250 bp reads (MiSeq Control Software v2.5.0.5). MiSeq Reporter v2.5.1.3 demultiplexed samples and removed reads without an index or matching PhiX. Within QIIME2, cutadapt removed primers from the paired reads and DADA2 identified ASVs. Primers 799F and 1193R were used to extract reads in silico from the QIIME-SILVA 16S database. These reads were used to build a classifier using QIIME2's naive-bayes method and the sklearn algorithm was used to generate taxonomy assignments for the sequence variants. These assignments were used to filter any remaining mitochondrial and chloroplast sequences. Sequence variants with a frequency lower than 2 counts, samples with

fewer than 10 reads, and samples with notes on irregularities during collection were also removed. To generate a phylogeny for the sequence variants, QIIME2 was used to align the sequences with MAFFT and to infer and root a phylogenetic tree. The tree was imported along with the DADA2-generated ASV count table, the taxonomy, and the metadata into a phyloseq (McMurdie and Holmes 2013) class in R (version 3.4.4) (R Core Team 2018) for analysis. Count table transformations, pruning, and rarefaction were performed with phyloseq and distance matrix calculation, ordination, and PERMANOVA tests were performed with the vegan package (Oksanen et al. 2017). Phylogenetic analysis was performed with ape (Paradis and Schliep 2018) and picante (Kembel et al. 2010). Figures and supplemental figures were produced with ggplot2 (Wickham 2016) and ggpubr (Kassambara 2018).

Statistics

Three dissimilarity metrics were used to capture different aspects of microbiome variation. Presence-absence variation was represented by the Raup-Crick dissimilarity index, a probability of samples differing in composition based on ASV frequencies in the dataset. Alternatively, the Bray-Curtis dissimilarity quantified the abundance differences between ASV counts in each sample. The UniFrac distance incorporated presence-absence variation as well as phylogenetic relatedness between the ASVs present in samples based on the 16S gene tree.

ASVs associated with specific tissues or developmental stages were identified using the signassoc function of the indicspecies package (Cáceres and Legendre 2009, Cáceres 2020). This function calculated an indicator value index (IndVal) based on the product of two probabilities:

(1) the probability that a sample belonged to a habitat given ASV presence and (2) the probability that an ASV was present if a sample was taken from a habitat. For the habitats defined by each variable (six tissues, six developmental stages, two sites, and two years), indices were calculated independently for each ASV. The null hypothesis that no relationship existed between ASVs and conditions was tested by comparing the empirical index with a distribution generated by randomly permuting the ASV presence-absence count table. A two-tail p-value was used to select ASVs that are significantly more or less frequently observed in sampled belonging to a given condition ($\alpha = 0.01$).

Additional details on the methods are presented in Appendix A.

RESULTS

We planted surface-sterilized seeds of *A. thaliana* accessions from a single North American haplotype (Table 1.1) at two southwest Michigan sites in two consecutive years. These accessions germinate in the fall, overwinter as small rosettes, flower in the spring, and senesce in the early summer. We harvested roots and rosette leaves throughout vegetative growth and also stems, cauline leaves, flowers, and siliques as they became available during flowering and senescence (Table 1.2). To enrich for endophytes, bacteria were washed from the surface of plant tissues by repeated vortexing in surfactant buffer (Perisin et al. 2016). Topsoil was also collected from field sites at each timepoint during the second year of study. Bacterial lineages in the soil and plant tissue samples were quantified by amplification and sequencing of the V5, V6, and V7

regions of the 16S rRNA gene (16S) (Bodenhausen et al. 2013). The 16S sequences were grouped into amplicon sequence variants (ASVs) with DADA2 (Callahan et al. 2016) in QIIME2 (Bolyen et al. 2018). After filtering out singleton 16S variants and those from plant organelles, 10,803 ASVs were tallied for 1,272 samples (Table 1.3). A phylogenetic tree for the variants was inferred with FastTree using MAFFT-aligned 16S sequences (Price et al. 2010, Katoh et al. 2013). The ASVs were classified at seven taxonomic ranks based on the SILVA 16S database (Quast et al. 2012).

TABLE 1.1 *Arabidopsis thaliana* ecotypes planted in the study (from the HPG-1 haplogroup)

Accession	Ecotype ID	Collection Location	Coordinates
BRR4	470	Watseka, IL, USA	(40.8313, - 87.735)
LI-WP-041	546	Nissequogue, NY, USA	(40.9076, - 73.2089)
L-R-10	1797	Union Pier, MI, USA	(41.847, - 86.67)
MNF-Che-47	1942	Manistee National Forest, MI, USA	(43.5251, - 86.1843)
Pent-7	2191	Pentwater, MI, USA	(43.7623, - 86.3929)
PT1.85	8057	Hanna, IN, USA	(41.3432, - 86.7368)
SLSP-69	2285	Silver Lake, MI, USA	(43.665, - 86.496)

TABLE 1.2 Sample collections for the study

Year	Date	Site	Stage	Sample type(s)
1	10/12/12	ME	No Plant	Soil
1	10/15/12	WW	No Plant	Soil
1	10/26/12	ME	Two Leaf	Roots, Rosette Leaves
1	10/29/12	WW	Two Leaf	Roots, Rosette Leaves
1	11/29/12	ME,WW	Four Leaf	Roots, Rosette Leaves
1	02/13/13	ME	Six Leaf	Roots, Rosette Leaves
1	03/29/13	ME, WW	Eight Leaf	Roots, Rosette Leaves
1	05/01/13	ME, WW	Flowering	Roots, Rosette Leaves, Stems, Cauline Leaves, Flowers, Siliques
1	06/14/13	ME, WW	Senescent	Roots, Stems, Siliques
2	10/28/13	ME, WW	No Plant	Soil
2	12/04/13	ME, WW	Two Leaf	Soil, Roots, Rosette Leaves
2	04/16/14	ME, WW	Six Leaf	Soil, Roots, Rosette Leaves
2	05/15/14	ME, WW	Flowering	Soil, Roots, Rosette Leaves, Stems, Cauline Leaves, Flowers, Siliques
2	07/03/14	ME, WW	Senescent	Soil, Roots, Stems, Siliques

TABLE 1.3 Samples in the study after quality control (at site ME, at site WW)

	Soil	Roots	Rosettes	Stems	C. Leaves	Flowers	Siliques	Total
No Plant	15, 19							34
Two-Leaf Plant	6, 4	3, 4	8, 10					35
Four-Leaf Plant		29, 8	27, 29					93
Six-Leaf Plant	8, 4	32, 18	45, 26					133
Eight-Leaf Plant		37, 24	41, 35					137
Flowering Plant	4,3	61, 38	64, 49	41, 38	24, 22	42, 28	14, 25	453
Senescent Plant	6,8	60, 64		54, 75			49, 71	387
Total	77	378	334	208	46	70	159	1272

Bacterial assemblage composition was associated with plant tissue type and developmental stage. Plants shaped the bacterial assemblages they hosted, making them distinct from those in the surrounding soil (Appendix B). Rather than a single host environment, the plant appeared to be a collection of microbe habitats defined by tissue type and age. Samples from the same tissue or stage clearly shared a higher proportion of members than randomly compared samples (Figure 1.1). When samples from multiple tissues of the same individual plant were available, comparisons showed that they did not share a significantly higher proportion of members than randomly paired samples. Bacterial assemblages therefore appeared to be more similar between samples of the same tissue type from different plants than between samples from different tissue habitats in the same plant. Common environments also influenced assemblages, as evidenced by increased membership overlap in samples from the same site or year compared to random samples. Despite these patterns, the low proportion of members shared within groups conditioned on any study variable (< 15%) underscored the high variability of colonization.

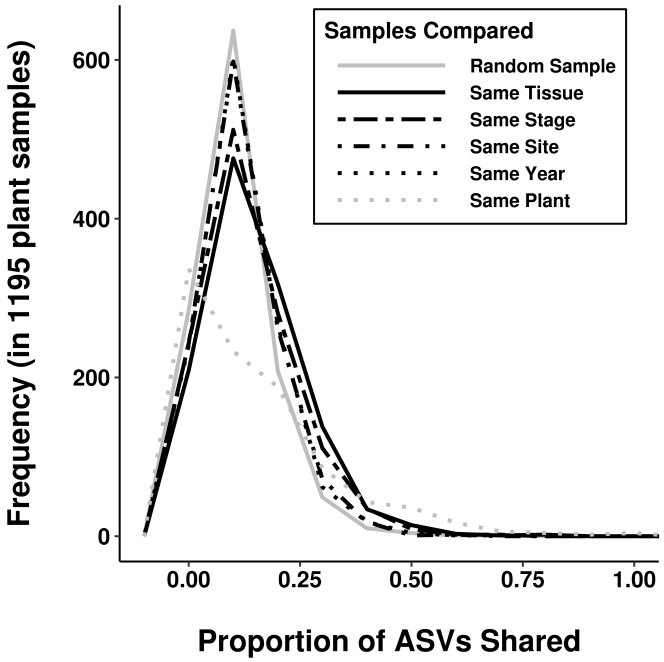

Figure 1.1 Bacterial assemblage composition was driven by the tissue sampled and the plant's developmental stage at harvest. The frequency distribution for the proportion of ASVs shared by plant samples selected randomly (solid gray, median 0.096) differed from those for plant samples selected with respect to tissue type, developmental stage, site, or year (patterned black lines). The proportion of shared members was significantly higher when sample selection was conditioned on each variable, but the strongest shifts were observed for tissue type and developmental stage (same tissue median: 0.133, $p < 2 \times 10^{-16}$; same stage median: 0.117, $p = 4 \times 10^{-11}$; same site median: 0.106, $p = 8 \times 10^{-5}$; same year median: 0.100, $p = 9 \times 10^{-4}$). The frequency distribution for the proportion of ASVs shared with any samples taken from different tissues of the same individual plant is also shown (dotted gray line, p=0.038).

The influences of host and environment on assemblage composition were further supported by analysis of variance. After randomly subsampling to 1000 counts, composition variation between samples was quantified with respect to ASV presence by Raup-Crick dissimilarity (Raup and Crick 1979), with respect to ASV abundance by Bray-Curtis dissimilarity (Bray and Curtis 1957), and with respect to ASV presence and phylogenetic relatedness by the unweighted UniFrac distance (Lozupone and Knight 2005). Analysis of variance with permutation was

performed on each dissimilarity matrix for each study variable ($\alpha = 0.001$) (Table 1.4a, 1.4b). Variance between ecotypes was not greater than the variance within them, which is unsurprising given the genetic similarity of the haplogroup to which they belonged (Exposito-Alonso et al. 2018). Variance among plant individuals was significant only when compared by Bray-Curtis dissimilarity. Sample preparation plates and sequencing runs differed significantly in composition only when compared by Bray-Curtis dissimilarity and UniFrac distance. Tissues, developmental stages, planting sites, and years differed significantly in composition regardless of the dissimilarity or distance used to quantify differences.

TABLE 1.4a Variables tested for association with dissimilarity index matrices by PERMANOVA

Variable	Description
Tissue	The part of the plant washed, ground, and sampled for DNA extraction
Stage	The developmental stage of the plant when the tissue was harvested
Site	The location of the field where the plant grew
Year	The year in which the plant was harvested
Ecotype	The *A. thaliana* genotype of the plant sampled
Sample Plate	The 96-well microplate in which sample extraction and 16S amplification were performed
MiSeq Run	The batch in which samples were sequenced on the Illumina MiSeq platform
Plant ID	The plant individual from which tissues were harvested

TABLE 1.4b Variables tested for association with dissimilarity index matrices by PERMANOVA

		Raup-Crick index			Bray-Curtis dissimilarity			UniFrac distance		
Variable	DOF	F	R^2	Pr (>F)	F	R^2	Pr (>F)	F	R^2	Pr (>F)
Tissue	5	577.184	0.803	< 0.001	23.909	0.145	< 0.001	21.600	0.133	< 0.001
Stage	5	120.622	0.461	< 0.001	14.606	0.094	< 0.001	9.091	0.060	< 0.001
Site	1	202.328	0.222	< 0.001	38.890	0.052	< 0.001	22.897	0.031	< 0.001
Year	1	93.356	0.116	< 0.001	21.224	0.029	< 0.001	11.904	0.016	< 0.001
Ecotype	6	1.717	0.014	0.342	0.931	0.008	0.748	0.802	0.007	0.958
Sample Plate	23	0	0	0.914	1.311	0.042	< 0.001	1.640	0.052	< 0.001
MiSeq Run	3	0	0	0.710	2.296	0.010	< 0.001	4.200	0.017	< 0.001
Plant ID	375	1.253	0.448	0.012	1.172	0.431	< 0.001	1.067	0.408	0.003

Assemblages were primarily influenced by the type of host tissue sampled and its age. Study variables were nested according to the experimental design to create a multivariate model for

permutational analysis of variance (PERMANOVA) (Anderson 2014). When PERMANOVA was performed with this model, tissue type and host developmental stage consistently explained the most sample variation regardless of the dissimilarity metric employed (Table 1.5). Tissue type was assigned between 13% and 38% of the total variance and developmental stage was assigned between 6% and 30% of variance. Since the residual sum-of-squares was markedly lowest in PERMANOVA on the Raup-Crick matrix, we focused on presence-absence variation in community composition when identifying the ASVs associated with specific tissues and developmental stages. An additional advantage of focusing on presence-absence variation is that 16S copy number varies among bacterial lineages.

TABLE 1.5 PERMANOVA results for rarefied data dissimilarity matrix ~ Year / Stage / Tissue + MiSeq Run / Sample Plate + Site

		Raup-Crick index			Bray-Curtis Dissimilarity			UniFrac Distance		
	DOF	F	R^2	Pr (>F)	F	R^2	Pr (>F)	F	R^2	Pr (>F)
Year : Stage : **Tissue**	17	256.795	0.376	< 0.001	7.330	0.127	< 0.001	7.949	0.144	< 0.001
Year : **Stage**	7	492.776	0.297	< 0.001	12.233	0.087	< 0.001	8.048	0.060	< 0.001
Site	1	2275.626	0.196	< 0.001	46.162	0.047	< 0.001	26.216	0.028	< 0.001
Year	1	1348.268	0.116	< 0.001	28.543	0.029	< 0.001	15.496	0.016	< 0.001
MiSeq Run : **Sample Plate**	20	0	0	1	1.346	0.027	< 0.001	1.389	0.030	< 0.001
MiSeq Run	3	0	0	1	3.18	0.01	< 0.001	5.59	0.018	< 0.001
Residuals			0.015			0.673			0.704	

Assemblages in phyllosphere tissues became more distinguishable from those in roots as plants matured. Assemblages varied more between root and shoot tissues than within the phyllosphere. In principal coordinate analysis (PCoA) based on their dissimilarities (Figure 1.2 A-C), samples from the stem and siliques clustered separately from rosette leaf samples and from root samples. This finding was robust to differences in rarefaction depth, filtering, and normalization of the count data (Weiss et al. 2017) (Figure 1.3). However, segregation along the first two principal coordinates was not clear when phyllosphere samples from flowering plants

were ordinated alone, suggesting that most of the association with tissue type was driven by differences between root and shoot (Figure 1.2 D-F). Supporting this interpretation, PERMANOVA on Raup-Crick dissimilarities of phyllosphere samples at flowering yielded a p-value below the significance threshold ($\alpha = 0.001$) (Table 1.6).

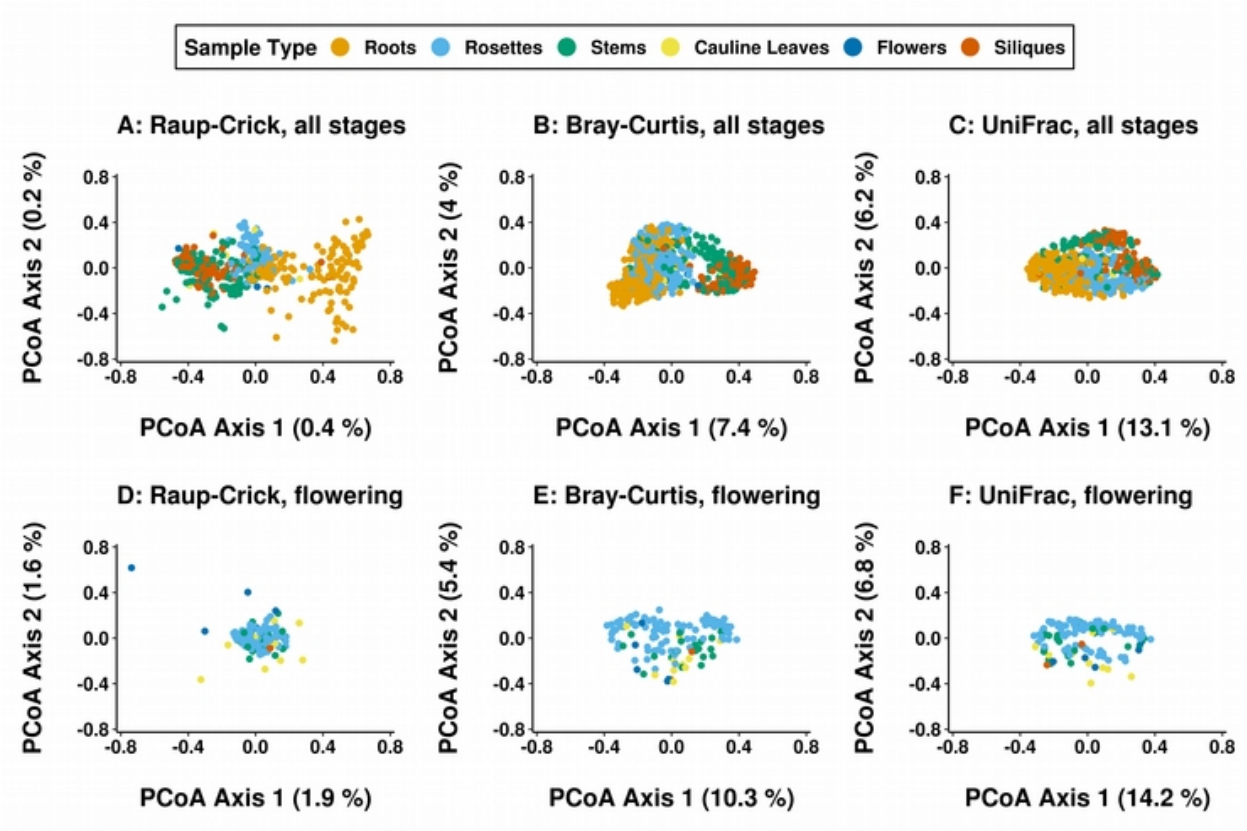

Figure 1.2 Roots and phyllosphere tissues housed distinct assemblages while tissues within the phyllosphere segregated between but not within stages. Plant samples segregated by tissue type in principal coordinate analysis (PCoA) based on their dissimilarities. The percentage of sample variance captured by the first two principal coordinates are listed on the x and y axis. (A,D) Raup-Crick dissimilarities are based on presence-absence differences between samples. (B,E) Bray-Curtis dissimilarities are based on quantitative differences in ASV counts between samples. (C,F) UniFrac distances incorporate phylogenetic relatedness of the ASVs present in samples based on the 16S gene tree. (A-C) For all dissimilarities, samples from roots (orange), rosette leaves (blue), and stems and siliques (green and red) clustered along the first coordinate. (D-F) In phyllosphere samples at flowering, rosette leaves (blue) overlapped with other phyllosphere tissues.

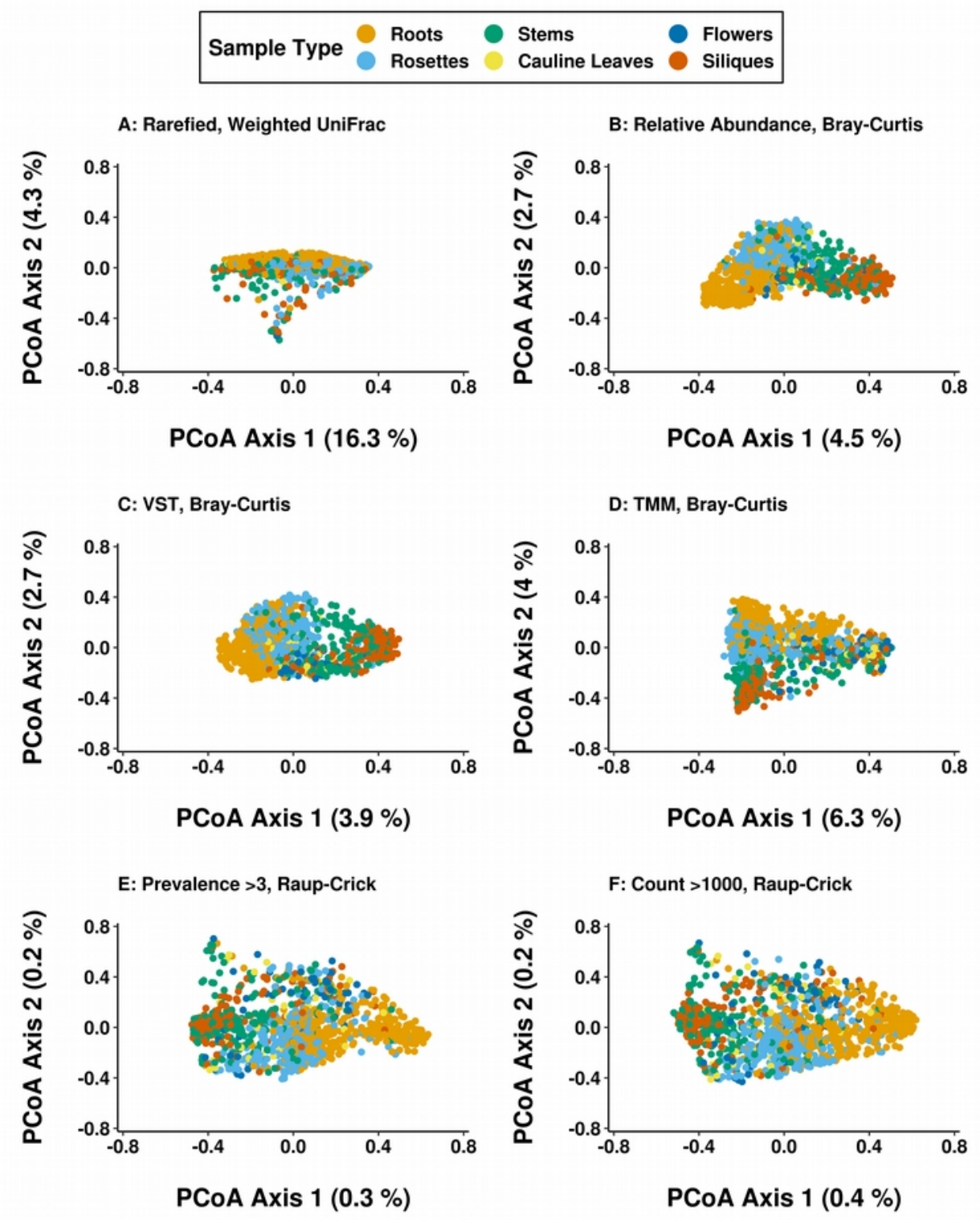

Figure 1.3 Plant tissue was the variable most strongly associated with sample composition, regardless of how composition variation was quantified or how the count tables for samples were transformed or thresholded.

Figure 1.3 Plant tissue was the variable most strongly associated with sample composition, regardless of how composition variation was quantified or how the count tables for samples were transformed or thresholded (continued). The relationships of host tissue type and stage to community composition were visualized by principal coordinate (PCoA) ordination of the samples based on dissimilarities or distances (Figure S1). Samples clustered primarily by the material sampled regardless of the dissimilarity or distance used. Observed tissue differences were robust to common methods for minimizing the effects of different count totals among samples. (A) Weighted UniFrac is influenced by the phylogenetic relationships of ASVs present in samples as well as quantitative differences in their abundances. This distance was calculated for a rarefied dataset. (B) Bray-Curtis dissimilarity is based on quantitative differences in ASV abundance across samples. This dissimilarity was calculated for relative abundances in all samples, without rarefaction. A variance-stabilizing transformation (C) or a log transformation (D) was performed on the count table before the Bray-Curtis dissimilarity was taken. Finally, the importance of host variables relative to site and year was robust to thresholding the ASV count table based on the prevalence or abundance. (E) ASVs were filtered for prevalence to include only those presence in more than three samples. (F) ASVs were filtered for abundance to include only those with a count total exceeding one thousand for all samples.

TABLE 1.6 PERMANOVA results for flowering phyllosphere samples

		Raup-Crick index		
	DOF	F	R^2	Pr (>F)
Year : **Tissue**	7	11.385	0.380	0.01
Site	1	42.098	0.201	< 0.001
Year	1	1.843	0.009	0.421
MiSeq Run : **Sample Plate**	20	0	0	0.917
MiSeq Run	3	0	0	0.773
Residuals			0.015	

To disentangle the roles of tissue and age in defining habitats within the plant, we compared root and shoot tissues with respect to the ASVs present both before and after developmental transitions. Roots and rosettes were compared between late vegetative and flowering stages while roots and stems were compared between flowering and senescence. Tissue assemblages grew more distinguishable later in development, with the proportion of variance explained by tissue relative to other host variables increasing at later stages (Table 1.7). The differentiation of tissue habitats over time was further examined by quantifying their β diversity at each stage. Pairwise dissimilarities of samples within and between tissue types were calculated and the

distributions of these distances were compared for both rosette leaves and roots (Figure 1.4). As development progressed, leaf assemblages simultaneously became more similar to each other and more distinct from those in the roots, perhaps due to unique selective pressures or a more restricted pool of potential colonists in the phyllosphere.

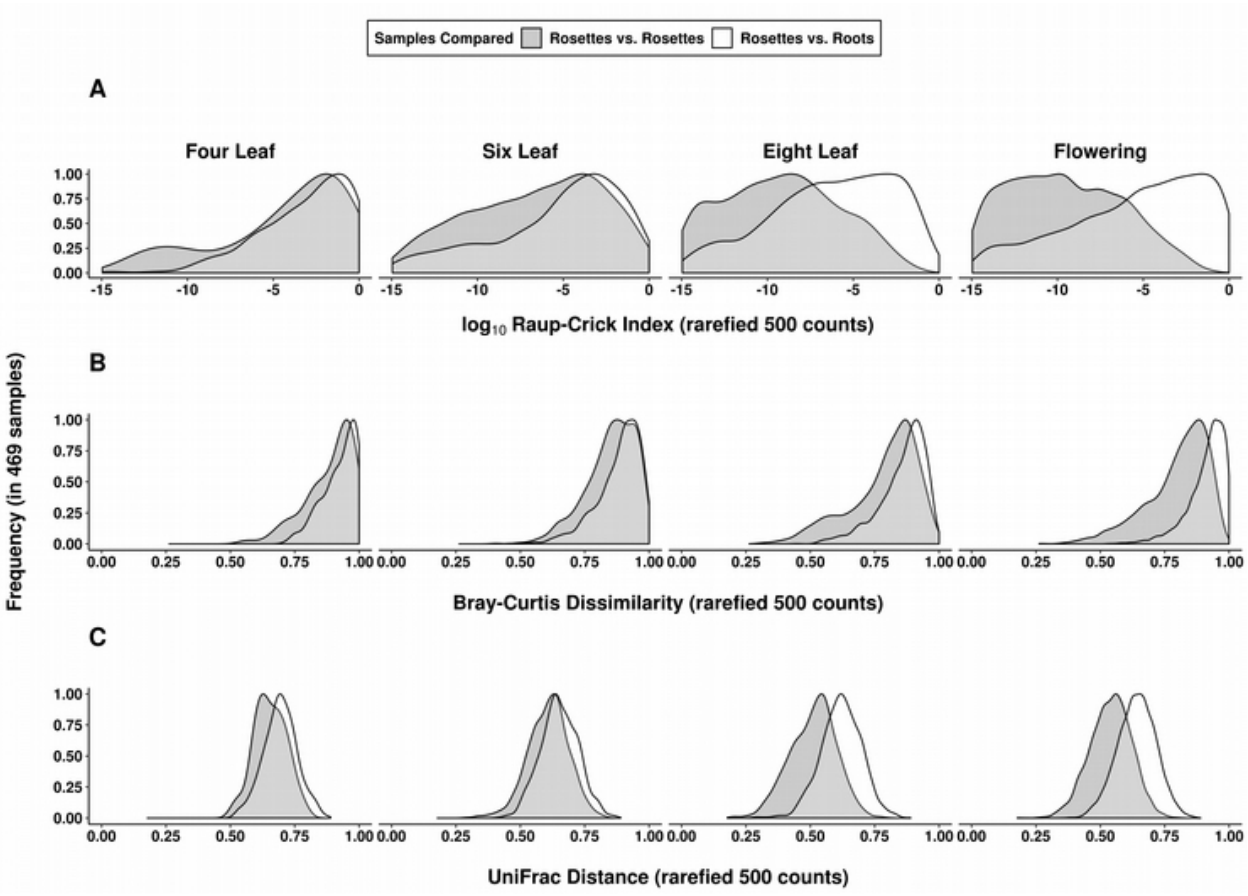

Figure 1.4 Phyllosphere assemblages became more distinguishable from those in roots as plants matured. The dataset was pruned to samples with at least 100 counts and 20 ASVs to calculate sample dissimilarities. The distribution of dissimilarities between rosette leaf samples (shaded distribution) was compared to the distribution of dissimilarities between rosette leaf and root samples (unshaded distributions). This procedure was repeated for (A) Raup-Crick dissimilarities, (B) Bray-Curtis dissimilarities, and (C) UniFrac distances. As plants matured (horizontal panels left to right), phyllosphere samples were increasingly similar to each other and distinguishable from the root samples.

TABLE 1.7 PERMANOVA results for tissue comparisons at different stages

	Samples	Tissue R^2	Tissue Pr (>F)
Vegetative Roots and Rosettes	69	0.187	0.013
Flowering Roots and Rosettes	163	0.621	< 0.001
Flowering Roots and Stems	94	0.391	< 0.001
Senescent Roots and Stems	216	0.881	< 0.001

Assemblage associations with tissue and stage depended on sub-genus variation in bacteria. In the study of microbiome assembly, patterns can depend upon the taxonomic resolution with which bacteria are surveyed. ASVs represent the finest resolution of bacterial lineages possible with 16S data. If the effect of host tissue type was driven by differential filtering of closely related variants rather than larger bacterial clades, then it should weaken as taxonomic resolution decreases. To determine the dependence of the observed associations on taxonomic grouping, PERMANOVA was repeated with a Raup-Crick dissimilarity matrix produced with tables of ASV counts aggregated at the level of genus, family, order, class, and phylum (Table 1.8). Associations between assemblage composition and host tissue and age weakened with coarser taxonomic groupings of ASVs from genera to phylum. The weakening associations with coarser grouping suggested that the distribution of colonists within the host was not driven by widely shared bacterial traits but rather by variation within genera, whether acquired by horizontal transfer or evolved in vertically inherited traits. As a consequence, differences in the colonization patterns across tissues and stages are erased upon averaging the prevalence patterns of recently diverged lineages across taxonomic groups.

TABLE 1.8 PERMANOVA results for different taxonomic groupings

		Tissue		Stage	
Taxonomic Rank	% ASVs unassigned	R^2	Pr (>F)	R^2	Pr (>F)
ASV	0	0.803	< 0.001	0.461	< 0.001
Genus	43.94	0.769	< 0.001	0.209	0.002
Family	17.71	0.682	< 0.001	0.146	0.132
Order	8.17	0.202	0.003	0.155	0.004
Class	2.97	0.172	< 0.001	0.081	0.065
Phylum	0.57	0.248	< 0.001	0.153	0.001

Assemblage associations with tissue and stage were largely driven by the same ASVs, which were neither tissue-specific nor transient. By sampling different tissues at multiple timepoints, we were able to compare the bacterial lineages that distinguished assemblages across space and time. Spatial and temporal trends in the data were largely driven by the same colonists. To find ASVs significantly associated with specific tissues and developmental stages ($\alpha = 0.01$), we compared the indicator value index of each ASV to a distribution generated by randomly permuting its presence-absence table (Cáceres and Legendre 2009). We found 460 ASVs that were significantly associated with specific tissue types; 70% of these were associated with roots and the remainder were associated with different phyllosphere tissues. Of the 268 stage-discriminating ASVs, 76% were also among the 460 that distinguished tissue types.

Assemblage differences between tissues were not driven by specialists and assemblage differences over time were not driven by transient community members. The ASVs associated with a specific tissue generally appeared in multiple plant tissues taken from a site in a given year rather than being restricted to a single habitat within the plant (93%). The ASVs associated

with a specific stage generally recurred in a tissue throughout development rather than appearing at a single harvest stage (91%). Thus, assemblages differences did not result from the exclusive presence of ASVs in specific tissues or at specific stages in development, but rather from quantitative differences in prevalence over space and time.

About a fifth of ASVs had consistent prevalence patterns across field sites and years while the rest had inconsistent spatial and temporal distributions. If tissue-specific host traits created environments favorable to particular colonists, then the spatial distributions of those colonists within the plant should be repeated across the sites and years in which they were observed. To assess whether ASV spatial distributions were repeated, the tissues in which ASVs reached maximum prevalence, when present, were compared between sites and years. Based on these comparisons, 21% (98/460) of the ASVs distinguishing tissues displayed consistent spatial trends, always peaking in prevalence in the same tissues (shown for Proteobacteria in Figure 1.5). For this fraction of ASVs, colonization patterns might be linked to tissue-specific host traits that differentially filter bacterial colonists. Of these consistently distributed ASVs, 79% were always most prevalent within roots, 11% in rosettes, 5% in stems, and 5% in siliques. Notably, the genus *Massilia* includes two distinct sets of ASVs that consistently peaked in different tissues; one set consistently peaked in roots while the second consistently peaked in siliques, emphasizing that sub-genus variation between bacterial lineages influenced their distributions within plants. For the remaining ASVs, including those in notable pathogen genera (Figure 1.5), spatial prevalence patterns were inconsistent between sites and years.

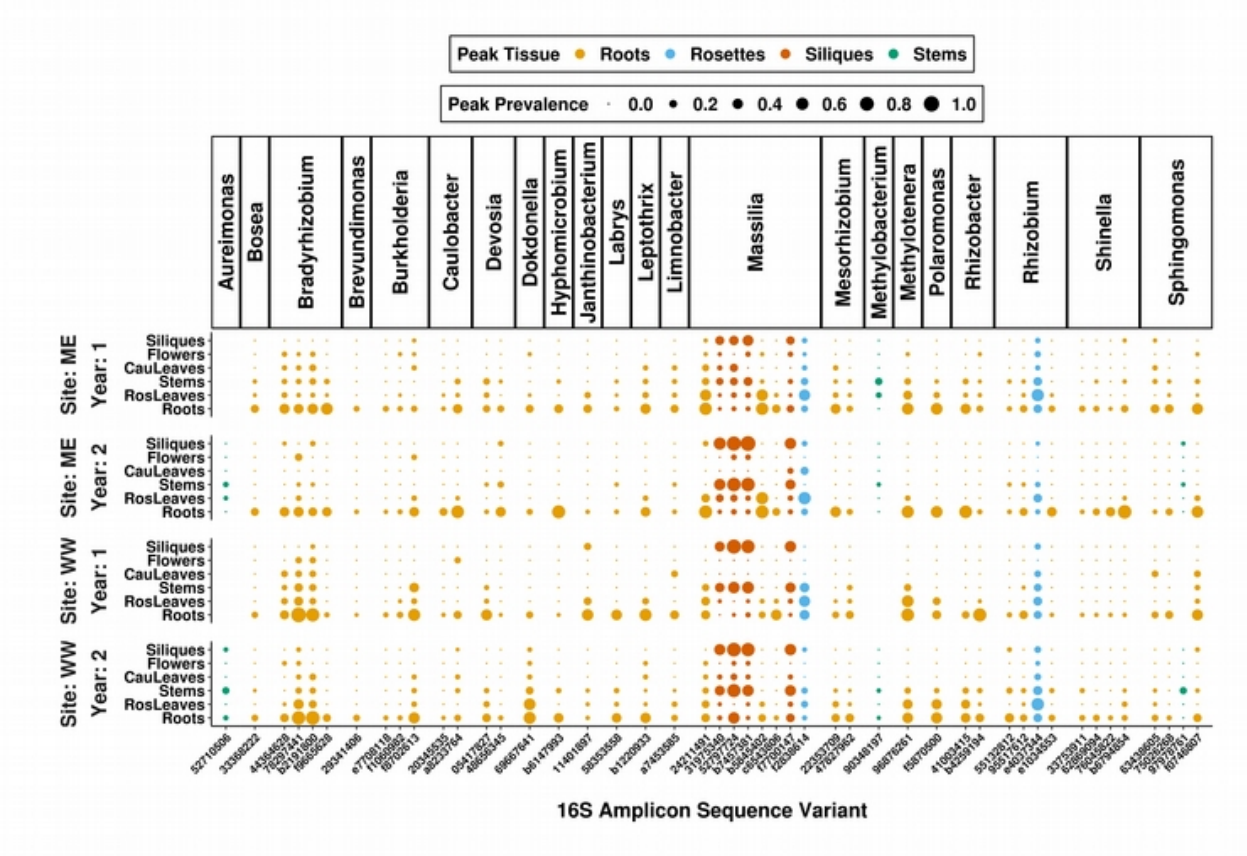

Figure 1.5 A fraction of ASVs behaved consistently across tissues at each site and in each year of the study. The ASVs distinguishing tissues, selected by indicator value indices, were filtered for those that reached maximum prevalence in the same tissue at each site and in each year when they were present. Proteobacteria ASVs with a total prevalence above five percent in all site and year combinations were selected for the plot. Panels on the y-axis separate sites and years and ASVs are grouped on the x-axis by genus. Dot sizes represent the maximum prevalence for each ASV in each tissue. Despite the significant association detected between assemblage composition and tissue type, only 21% of tissue-discriminating ASVs consistently reached peak prevalence in the same tissue. Of the ASVs that behaved consistently, 79% always reached peak prevalence in roots.

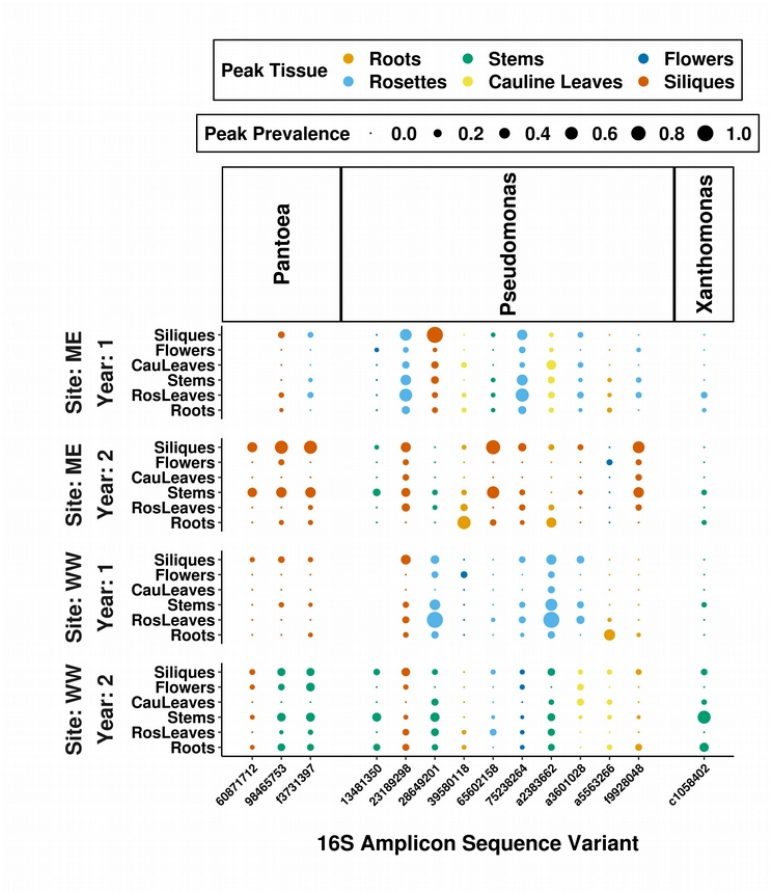

Figure 1.6 Most ASVs, including those in genera with pathogenic potential, do not have consistent spatial distributions across sites and years. The majority of ASVs distinguishing tissues did not reach maximum prevalence in the same tissue at each site and in each year when they were present. This pattern is exemplified by ASVs in the genera Pseudomonas, Xanthomonas, and Pantoea, which are notable for their pathogenic potential. Panels on the y-axis separate sites and years and ASVs are grouped on the x-axis by genus. Dot sizes represent the maximum prevalence for each ASV in each tissue.

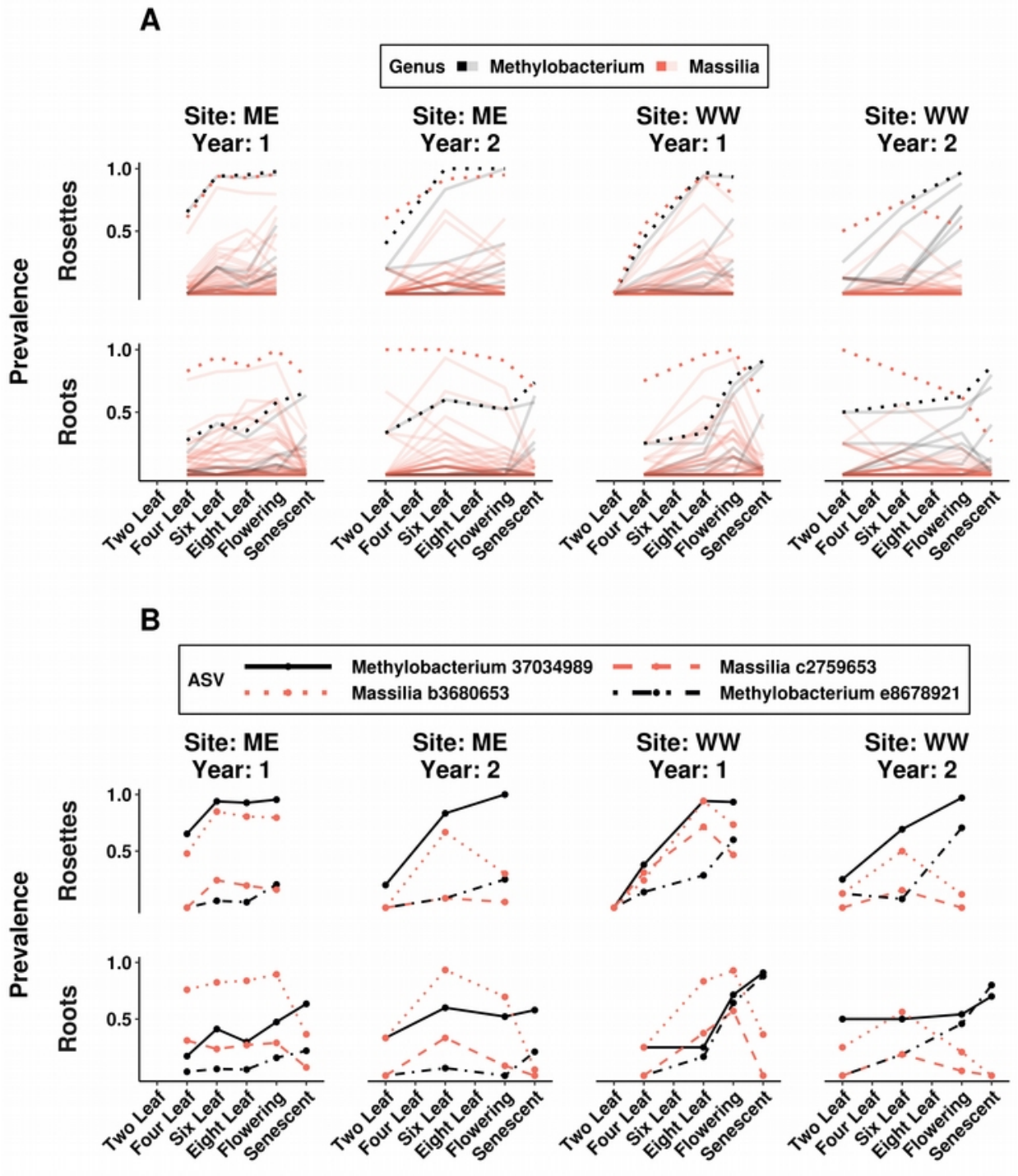

Figure 1.7 ASVs in the same genus have distinct temporal trends; a small number consistently reach high prevalence across sites and years despite variation in temporal trends.

Figure 1.7 ASVs in the same genus have distinct temporal trends; a small number consistently reach high prevalence across sites and years despite variation in temporal trends (continued). Plots feature Proteobacteria genera (*Massilia* in red and *Methylobacterium* in black) containing ASVs that distinguished developmental stages, based on indicator value indices, and reached over 70% prevalence in multiple sites and years. Horizontal panels separate the trends for roots and rosettes and vertical panels separate the trends for each site and year. (A) Bold trend lines show the temporal trends for counts grouped by genus, while transparent lines show the trends for individual ASVs in each genus. Genus-level trend lines subsume distinct ASV patterns. (B) Temporal colonization patterns were highly inconsistent despite the significant association between composition and stage, with only 23% of stage-discriminating ASVs consistently peaking at the same stage. Despite this variation, specific *Massilia* and *Methylobacterium* ASVs were among the most prevalent at both sites and in both years of study.

Temporal distributions, like spatial ones, were highly variable. Of the ASVs driving associations with stage, only 23% (62/268) reached peak prevalence at a consistent plant developmental stage across sites and years. Temporal distributions were also dependent on fine taxonomic variation because ASVs within each genera displayed a variety of dynamics (Figure 1.7). Among the Proteobacteria ASVs driving associations with age, the ones demonstrating the biggest changes in prevalence were found in both roots and rosettes across sites and years (Figure 1.7). Temporally dynamic *Massilia* ASVs peaked during vegetative growth or flowering and then declined. Temporally dynamic *Methylobacterium* ASVs consistently increased during plant growth and peaked at senescence.

Assemblages became more phylogenetically diverse and even over time. Temporal changes in assemblage α diversity can help explain the increased differentiation between assemblages inhabiting different tissues. The phylogenetic diversity of colonists was higher later in plant development (Figure 1.8A). Phylogenetic diversity was quantified by taking the tree of 16S variants present in each assemblage, weighting the branch lengths by variant abundance, and summing the branch lengths. This diversity trend was observed in each type of plant tissue sampled (Figure 1.9) but not in samples of the surrounding soil, indicating it was related to the colonization of plant tissue and not purely driven by the abiotic environment during sample collection. The increasing phylogenetic diversity suggested that bacteria from across the tree had dispersed more widely among plant assemblages later in development. With more opportunities to encounter plants over time, subtle differences in the ASV colonization success between tissues were more likely to be exposed.

Temporal trends in assemblage evenness suggested that tissue type was related to the size of the colonization bottleneck for specific ASVs rather than to their ability to dominate other colonists once established in the endosphere. Variants that reached high prevalence in a tissue did not overtake assemblages in terms of relative abundance. The average Shannon-Wiener index (Shannon 1948) of assemblages increased (Figure 1.8B) and the distribution of ASV relative abundances decreased (Figure 1.8C) during plant life. Together, these trends showed that instead of domination by a small number of successful variants, mature tissues on average housed a more evenly represented set of ASVs.

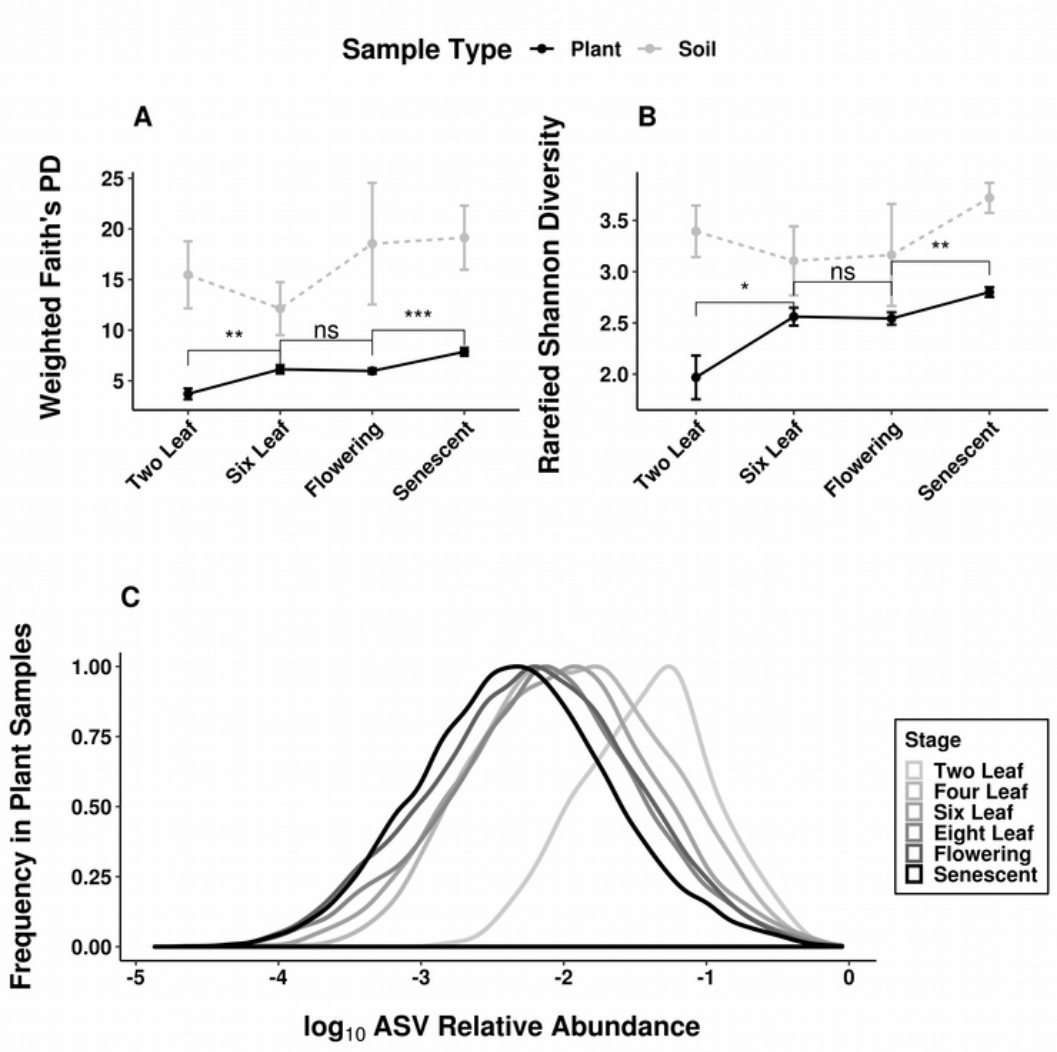

Figure1.8

Figure 1.8 Assemblages within plants became more phylogenetically diverse and even over time. (A-B) Plots show mean and standard error for diversity measures of the plant (black) and soil (gray) samples at each developmental stage in the second year of study. Significance is shown for pairwise Wilcoxon tests between stages as follows: ns (not significant), * ($p < 0.05$), ** ($p < 0.01$), *** ($p < 0.001$), **** ($p < 0.0001$). (A) Phylogenetic distance was measured as the branch length on the 16S phylogenetic tree between ASVs in the sample, weighted by ASV abundances. It increased on average with developmental stage. (B) Shannon-Wiener (H') indices from rarefied samples depended upon both the richness and evenness of samples and increased on average with developmental stage. (C) ASV relative abundances decrease during development as assemblages become more even. For plant samples with at least 100 counts, the relative abundance of each ASV present was calculated. The frequency distributions of these relative abundances are plotted for each developmental stage (shade), with relative abundance scaled by log10.

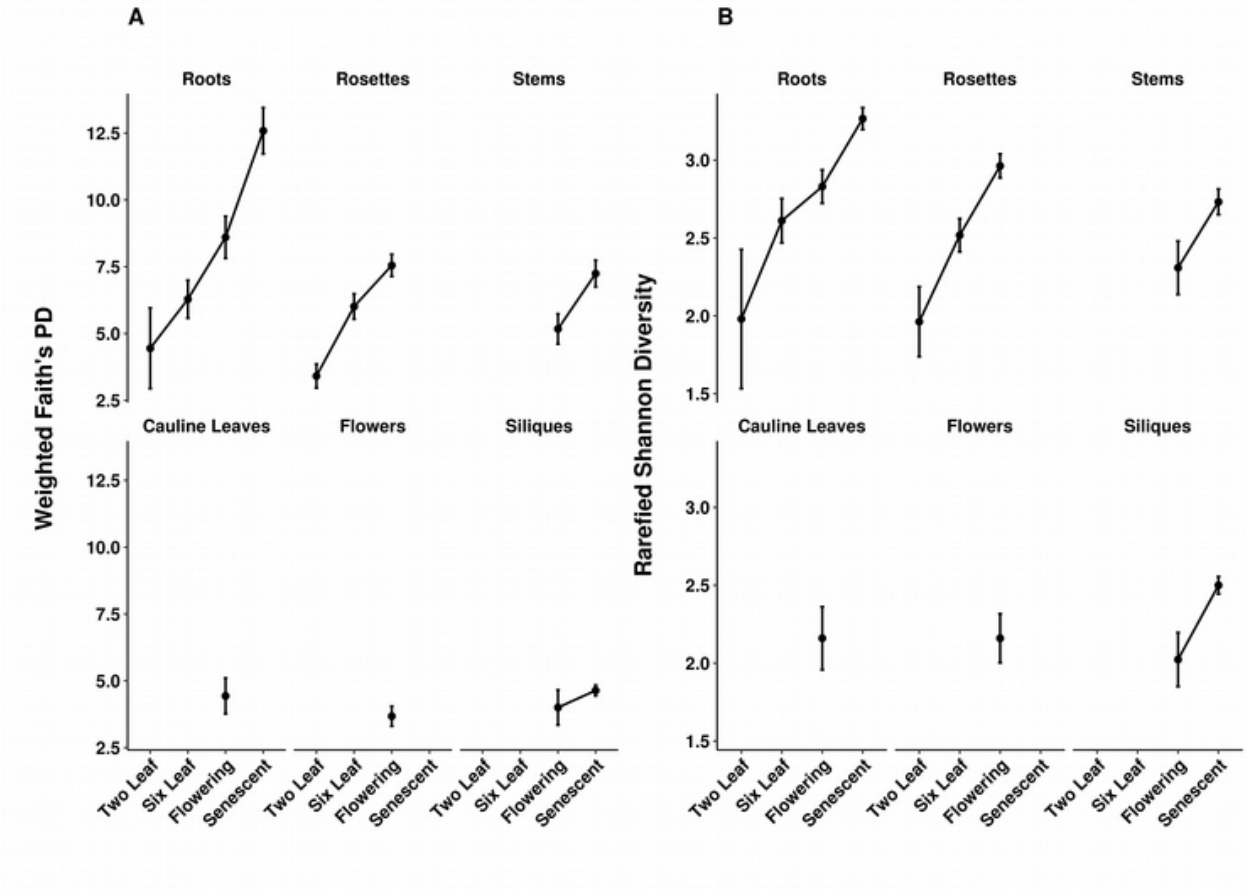

Figure 1.9 Diversity and evenness increased during development in all plant tissues. All plots show mean and standard error for diversity measures of the plant samples at each developmental stage in the second year of study. Flower and cauline leaves were only sampled at one developmental stage and thus do not have trend lines. (A) Phylogenetic distance was measured as the branch length on the 16S phylogenetic tree between ASVs in the sample, weighted by ASV abundances. It increased on average with developmental stage. (B) Shannon-Wiener (H') indices from rarefied samples depended upon both the richness and evenness of samples and increased on average with developmental stage.

DISCUSSION

We surveyed bacterial assemblages inside roots and phyllosphere tissues throughout the life cycle of the model annual plant *A. thaliana*. Consistent with previous results, we found that assemblage composition differed between tissues (Bodenhausen et al. 2013) and during the course of development (Bulgarelli et al. 2012, Chaparro et al. 2014). Because we examined multiple tissues and developmental stages in the same study, we were able to find three connections between these spatial and temporal trends in natural colonization that have not previously been identified.

First, the associations of host tissue type and developmental stage with assemblage composition were largely driven by the same colonists. These bacterial variants did not constitute omnipresent cores of tissue-specific inhabitants: those most strongly associated with tissue type typically did not reach even fifty percent prevalence within samples from a tissue type at a given time, underscoring the variability of community composition. Nor were these variants exclusive to specific habitat patch within the plant: most were observed at multiple sampling times and in multiple tissues. Instead, differential filtering over space and time quantitatively affected the prevalence of common endophytes, in agreement with a recent report of consistent organ occupancy in the plant microbiome (Massoni et al. 2019).

Second, endophytic assemblages filtered by root and shoot tissues became more distinguishable later in development. Specifically, leaf assemblages grew more differentiated from root

assemblages on average. ASVs with consistently higher prevalence in specific tissues suggested the existence of subtle differences between variants in the probability of colonizing different tissue niches. Increasing phylogenetic diversity indicated that bacterial lineages became more widely dispersed during plant development. As variants had more opportunities to colonize plant tissue, differences in the probability of colonizing particular tissues became more important in determining assemblage composition than the stochasticity of early colonization. Supporting this idea of the host plant gaining influence during assembly, a recent study of rice (*Oryza sativa*) roots found that microbiome composition was dynamic during vegetative growth and stabilized later in plant life (Edwards et al. 2018).

A third major pattern in our data is the variability of spatiotemporal distributions among lineages within the same bacterial family or even genus. Most ASVs associated with tissue type (70%) were significantly enriched or depleted in the roots. Indeed, many of these ASVs belonged to genera that were consistently detected as root endophytes in a study of 17 European sites over 3 years (Thiergart et al. 2020). These included members of *Massilia*, *Burkholderia*, and *Bradyrhizobium*. However, *Massilia* also contained ASVs that reached peak prevalence in rosette leaves or siliques. An important factor in detecting associations between host and colonists is the resolution with which bacterial lineages are grouped in count tables. When microbes derived from natural sources, including plant leaves, are passaged outside the host in minimal media, they produce communities that are similar at the family level despite being highly variable at the level of sequence variants (Goldford et al. 2018). In our study, grouping variants by family weakened associations with tissue type and erased the observed associations with harvest stage.

Unlike the carbon metabolism traits that determined community structure *ex situ* (Goldford et al. 2018), the functions selected by different tissue habitats may therefore not be shared broadly by lineages in a family.

In contrast to the roots, the phyllosphere presents microbes with a variety of challenges related to desiccation, toxins from other colonists, and motility (Vorholt et al. 2012). Adaptations known to mitigate these challenges, such as chemosensory and antimicrobial resistance genes, have occurred recently and vary at the sub-genus level in the lineages of known phyllosphere colonists (Clarke et al. 2016, Hwang et al. 2005). Traits like these could lead to greater success in colonizing the leaf niche, creating phyllosphere assemblages that are more similar to each other and distant from root assemblages. As a result, closely related strains of known pathogenic, plant-beneficial, or biocontrol taxa might not establish in the same way throughout their plant hosts. Functional profiling of the bacterial assemblages may therefore be more valuable than taxonomic profiling in understanding their spatial and temporal trends within plants.

Our experimental design, including replicate sites and study years, allowed us to characterize consistency in distributions over space and time for the variants associated with host plant features. Biotic factors that can differ between roots and the phyllosphere, like salicylic acid, have been manipulated and shown to influence community assembly in both lab and field conditions (Lebeis et al. 2015, Traw et al. 2007). If such deterministic factors drove the observed variation in composition, then we might expect to see bacterial endophytes with the same across-tissue spatial distributions in repeated surveys. However, only about 20% of variants

distinguishing tissues were consistently more prevalent in a specific tissue. The inconsistent behavior of most variants is perhaps not surprising given that samples typically did not share more than 15% of their colonists. This variability in the detection of endosphere colonists suggests that in nature, as in more controlled environments (Maignien et al. 2014), stochastic factors drive much of community assembly. Even if individual lineages interact consistently with host plants, chance entry of functionally redundant strains from a large pool of soil colonists may give rise to mosaic assemblages.

Despite the growing strength of tissue effects during development, assemblages did not collapse to a few successful inhabitants of each tissue. In contrast, assemblage evenness increased in all plant tissues, agreeing with previous reports that community diversity increases as host plant tissues age (Chaparro et al. 2014, Leff et al. 2015). While host tissues appear to create different colonization bottlenecks for bacterial lineages, they did not appear to favor the dominance of these ASVs over others within assemblages.

These findings have consequences for the study of another key determinant of plant microbial communities: host genotype. Host genotype effects on colonization efficiency and microbiome composition are found both among angiosperm species that diverged hundreds of millions of years ago and among crop accessions that diverged through domestication within ten thousand years (Bulgarelli et al. 2015, Germaine et al. 2004, Kembel et al. 2014, Laforest-Lapointe et al. 2016, Peiffer et al. 2013, Barak et al. 2011). Compositional differences in the field can even be related to polymorphisms within a plant species (Bergelson et al. 2019, Horton et al. 2014). Since

the variants associated with host variables are typically present throughout the plant and recurrent during development, filtering complex natural assemblages with these criteria can increase signal in the search for host polymorphisms linked to colonization success. Since tissue-associated assemblages become more distinct later in development, the host polymorphisms linked to variant prevalence or abundance might depend on which part of the plant is sampled, and when.

The effects of plant tissue type and developmental stage on assemblage composition were large compared to that of geographic site. Our results add to a growing body of evidence that the tissue sampled from a plant can explain more variation in microbial communities than geography. Studies of *A. thaliana* and *Boechera stricta* find that geographic sites and soil inocula play a substantial role in filtering plant microbial communities (Bulgarelli et al. 2012, Wagner et al. 2016). However, studies of cultivated *Agave* find that tissue explains more variation in community composition than species and site (Coleman-Derr et al. 2016). Tissue type also explains more variation in epiphytic community composition than sample site in the species *S. taccada* (Amend et al. 2019). Together, these results suggest that plants are best considered as collections of distinct habitat patches for bacterial colonization.

CHAPTER 2

Transient interactions and influence among bacteria in field-grown *Arabidopsis thaliana* tissues

CO-AUTHORS & CONTRIBUTIONS

I performed the analysis in this chapter based on Matt Perisin's dataset from Chapter 1. My work was supported by the University of Chicago Biological Sciences Division and NIH T32 GM 7197. I wrote the text with contributions from Matt, Joy Bergelson, and Greg Dwyer.

INTRODUCTION

During flowering, angiosperms are visited by a species-rich and diverse set of pollinators. Meanwhile, their tissues harbor species-rich and diverse assemblages of microbial life. Traditional studies of both plant-pollinator mutualisms and plant-microbe symbioses focus on just a few of the species sustained by plant resources. Ecological networks capture a broader view of the complex, multi-species interactions in these systems. With these networks, we can ask how the structure of community interactions (edges) and the roles of species (vertices) differ between environments and over time. Global properties of interaction networks help us predict how the indirect consequences of perturbations might spread in an ecosystem (Burkle and Alarcón 2011, Delmas et al. 2019).

Previous research on networks in plant ecology has focused on networks built from direct observations of pollinators in the field (Ings et al. 2009). Compared with pollinator assemblages, microbial assemblages offer a complimentary set of advantages and limitations for testing ecological hypotheses with species interaction networks. These advantages and limitations relate to the interpretation of edges, the identity of nodes, and the ease of capturing temporal variation in species interactions.

One advantage of plant-pollinator networks is that their edges represent the frequency of physical encounters between species in their natural environments. In contrast, a limitation of the microscopic scale is that it precludes direct observations of species interactions. This is especially true in field conditions where controlled competition assays and metabolic profiling are not feasible (Venturelli et al. 2018, Wei et al. 2015). Instead, DNA sequences are used to infer interactions (Layeghifard et al. 2017) that represent multiple ecological possibilities: cooperation by sharing secreted metabolites, direct competition via toxins, indirect competition for limited host resources, or entrainment by common plant or environmental factors (Figueiredo and Kramer 2020).

In plant-pollinator networks, confidence in the identity of nodes is limited because species must be classified accurately in real time and their identities cannot be confirmed by DNA sequencing. Microbial interaction networks, in contrast, are generated from nucleotide signatures recovered from frozen samples. Samples can therefore be collected quicker in large numbers and species identities and interactions resolved at a later time and even reassessed as phylogenies improve.

Because of limits to field site access or observation hours at a given site, plant-pollinator networks are not often replicated and typically provide only a snapshot of species interactions at one point in time (Olesen et al. 2010). Yet during the lifespan of host plants, species abundances in plant-associated assemblages vary (Beilsmith et al. 2020, Chaparro et al. 2014, Edwards et al. 2018, Hannula et al. 2019) between temporal niches defined by seasons (Basilio et al. 2006), years (Olesen et al. 2008), and decades (Olesen et al. 2011). Compared to observations of plant pollinator assemblages, samples of microbial life in plant tissues are relatively easy to collect (Meyer and Leveau 2012), making feasible large and replicated field surveys in multiple environments over time (Faust et al. 2015, Bohan et al. 2017).

Time-lapse networks can capture how ecological relationships in micorbial communities change with species composition and interaction context (Tylianakis and Morris 2017, Herrera 1988). For example, vertices, edges, and global network properties such as connectance vary during leaf litter decay (Krüger and Buscot 2016) and with succession in root-associated assemblages (Shi et al. 2016). Even when the same species occur together, edges can vary if interactions depend on the number and identity of other species in the system. In microbes, this context-dependence been demonstrated in principle by controlled studies in the *Drosophila* gut (Gould et al. 2018). The ecological interactions represented by edges can also vary across environmental gradients such as drought and soil pH (de Vries et al. 2018, Fan et al. 2018).

To date, most work adding this temporal variation in the microbial networks associated with

living plant hosts has been done in greenhouse microcosms (Shi et al. 2016, de Vries et al. 2018), limiting the degree of species turnover and environmental variation compared to studies in the open field. A major challenge is determining the spatial and temporal scales at which to sample to assess species turnover and interaction robustness in natural systems. In network analysis, observations from mutiple timepoints or study sites are often pooled to produce a matrix of robust species interactions (Barberán et al. 2012). While cumulative networks over a timespan are more likely to capture context-independent interactions and rare species, networks for consecutive time slices provide temporal resolution (Schoenly and Cohen 1991, Hagen et al. 2012, Poisot et al. 2015). In plant-pollinator networks, time-specific and cumulative networks differ in their structural properties and composition (Rasmussen et al. 2013). The same trade-off between robustness and resolution applies at the microbial scale; in addition, evolutionary dynamics can change interactions mapped between microbes over time, as observed for phage-bacteria infection networks (Gurney et al. 2017, Fortuna et al. 2019). Basic research on the turnover in network membership and interactions is required to determine how temporal niches should be defined and how data can be appropriately pooled when studying the assembly and resilience of plant-associated microbial co-occurrence networks. In this study, we compare interaction networks of bacteria inhabiting the tissues of *A. thaliana* between three major developmental stages of plant life in the field, replicating this time series over two years and at two independent field sites.

METHODS

Overview

We constructed networks of bacterial interactions in *A. thaliana* tissues from root samples taken at three successive stages during plant development. We used measures from network theory to characterize these networks and compared the measures across plant development to quantify the turnover of bacterial interactions and to examine changes in global structure over time. We also compared networks constructed with leaf samples from vegetative and flowering plants. We repeated all of these comparisons with independent time series in *A. thaliana* plants from two field sites and two different years. A full description of the dataset can be found in Beilsmith, Perisin, and Bergelson 2020.

Bacterial abundances in field samples

To characterize variation in bacterial interactions as plant hosts aged, we grew hundreds of *A. thaliana* in field conditions and harvested plants at each major developmental stage of their life cycle: vegetative growth, flowering, and senescence. After harvest, we separated and washed leaves and roots as well as the stems and seed-bearing siliques present at senescence. After separation, we immediately froze the tissues to capture the composition of the microbial communities they harbored. The tissues were homogenized before DNA extraction, eliminating spatial structure within them.

We inferred interactions between bacterial lineages in plant tissues based on their relative

abundances. To find these abundances, we amplified a region of DNA from the bacterial 16S rRNA gene with PCR to serve as a marker for distinct bacterial lineages in samples. The counts of each 16S amplicon sequence variant, or ASV, represented the abundances of distinct bacterial lineages. The counts were divided by the total number of sequences obtained from each sample to generate relative abundances. To avoid inferring spurious interactions with rare bacteria, we filtered for bacterial ASVs with a minimum prevalence (three samples) and abundance (two hundred and fifty) in the dataset using phyloseq (McMurdie and Holmes 2013) in R (R Core Team 2018).

Correlation and conditional independence of bacterial abundances

Since network structures depend on the methods used to construct them (Hirano and Takemoto 2019, Williams et al. 2014, Closs and Lake 1994), we inferred interactions with two different approaches to determine whether temporal trends were robust. While both of our approaches assumed sparse networks, they used different criteria to determine which pairs of bacteria likely interacted.

The first approach was based on Pearson correlations between log-transformed abundances for each pair of bacteria. The SparCC pipeline, used to implement the correlation approach, finds correlations that arise from interactions between bacteria while accounting for the fact that relative abundances in each sample are not independent (Friedman and Alm 2012). The pipeline assigns bootstrapped p-values to each pairwise correlation between bacterial ASVs. When the correlation coefficient's absolute value was above a minimum threshold (0.3) and the p-value fell

below a maximum threshold (0.001), an edge was drawn between the ASVs.

The second approach was based on conditional dependence between bacterial ASVs, or pairwise relationships between ASV abundances that held regardless of the abundances of other ASVs in the dataset. The SPIEC-EASI pipeline implements the conditional dependence approach in one of two ways (Kurtz et al. 2015, Liu et al. 2010). One option uses the inverse covariance matrix of log-transformed ASV counts. When the entry for two ASVs in this matrix is zero, then their abundances are conditionally independent and no edge is drawn between them. The other option uses regression-based neighborhood selection to find the smallest set of interaction partners for an ASV that leave its interactions conditionally independent of other ASVs in the dataset. We used SPIEC-EASI with neighborhood selection to create edges between conditionally dependent ASVs. We visualized and analyzed the resulting networks with igraph (Csardi and Nepusz 2006) in R. Our figures were produced with igraph, ggplot2 (Wickham 2016), and ggpubr (Kassambara 2018).

Properties of network edges

An open question is how consistently plant-associated bacterial networks share the global properties reported thus far in the literature, including widespread positive interactions. To address this question, we found pairwise Pearson correlations between ASV abundances in networks constructed using both correlation and conditional dependence approaches. To determine whether the strength of interactions was associated with the relatedness of ASVs, we used a phylogenetic tree based on the 16S gene sequences to find the total branch length between

each pair with the package vegan (Oksanen et al. 2017). We fit a linear model to the absolute values of edge correlation coefficients and branch lengths to assess whether the ecological and evolutionary relationships between bacteria might be related.

To examine how consistent interactions were over the plant's lifetime, we first found the bacteria that were present in networks before and after plant transitions to flowering and senescence. We then found the fraction of edges between these nodes that remained intact at the following developmental stage. To generate a null expectation for the conservation of interactions, we used the edge density in each network to find the probability of edges overlapping by chance.

Properties of networks nodes

We also compare the bacterial lineages identified as highly influential hubs in the network at each stage in plant development. We found hubs using two measures of a node's influence in the network, following the approach by Agler et al 2016 (Agler et al. 2016). Degree, or the number of edges connecting to a node, represented the number of ASV interactions. Betweenness centrality, the number of times a node fell on the shortest path between pairs of nodes in the network, favored ASVs at highly influential positions within the network structure.

Given that global network properties could change in datasets spanning plant development, hubs were detected with a dynamic threshold rather than a single cutoff for node degree. Nodes with degree and betweenness z-scores corresponding to the ninetieth percentile of a network were classified as hubs. The conservation of hubs across networks were compared using heatmaps and

the UpSetR package (Conway et al. 2017).

Global properties of networks

We wanted to investigate whether the interaction networks were consistent over time with respect to global properties. The property of scale-free structure (Agler et al. 2016, Shi et al. 2016, Barberán et al. 2012) is particularly important because the identification of hubs, or highly connected nodes, in scale-free networks is used to find putatively influential microbes (Barabási and Albert 1999, Röttjers and Faust 2018, Berry and Widder 2014, Banerjee et al. 2018, Röttjers and Faust 2019) relevant to plant health (Agler et al. 2016) or key to assembling communities (Xian et al. 2020). The degree distribution of each network was fit to a power law to assess whether it was consistent with that of a scale-free network (Barabási and Albert 1999). A null degree distribution was generated for each network with the Barabási-Albert model for growth with preferential attachment using the “psumtree” algorithm in igraph's sample_pa function and a number of nodes equivalent to that in the empirical network.

Another global property we examined over time was the modularity of the networks. The extent of structure in each interaction network was assessed by clustering it into communities based on edge betweenness (igraph's cluster_edge_betweenness) and calculating modularity with the resulting community structure (igraph's modularity). Since modularity depends on network size, this property was compared between networks with the z-score of their modularity in distributions based on one hundred randomly generated graphs of the same size (igraph's erdos.renyi.game). A local clustering coefficient was calculated for each node using igraph's

transitivity function. A linear model was fit to the log10-transformed clustering coefficient vs. log10 degree for each node to assess whether the network showed evidence of a hierarchical structure (Ravasz and Barabási 2003).

RESULTS & DISCUSSION

DNA-based networks of plant-associated microbes have been built to compare rhizosphere communities under different soil treatments (Schmidt et al. 2019) and to compare root and shoot communities in plants with different modes of nutrition (Fitzpatrick and Schneider 2020). In the model plant *A. thaliana*, networks have also been used to characterize inter-kingdom interactions in the root microbiome (Bergelson et al. 2019) and to find putative keystone taxa in the leaf-associated microbiome (Agler et al. 2016). These previous studies of plant microbial networks reported that most interactions between bacteria were positive. We too found that abundance correlations were overwhelmingly positive throughout development, regardless of the tissue sampled, the site of harvest, and the method used to construct the networks (Table 2.1). Interactions are visualized for root networks based on conditional dependence in Figure 2.1 and for the corresponding leaf and stem networks in Figure 2.2.

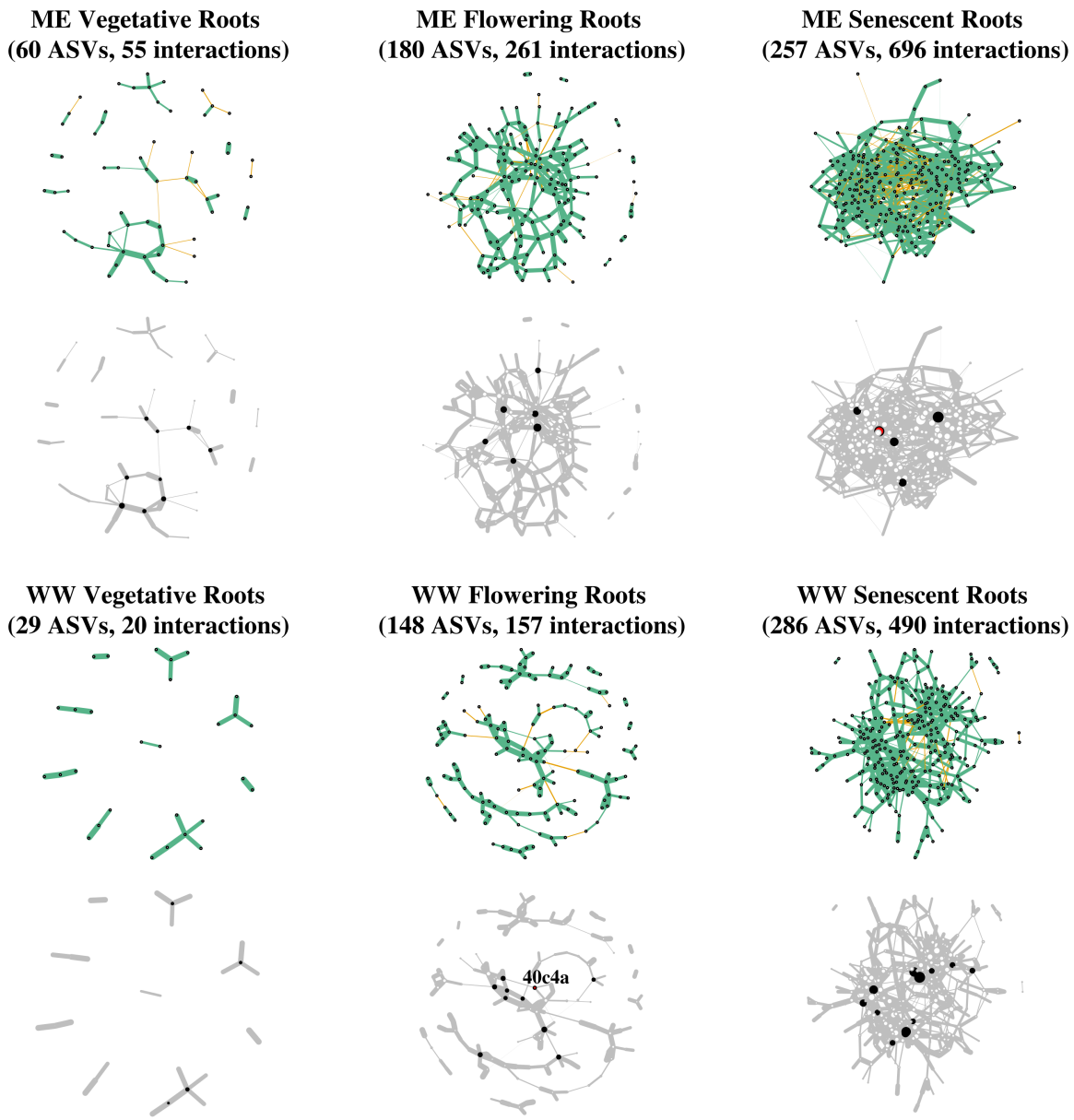

Figure 2.1 Bacterial interactions in *A. thaliana* remain overwhelmingly positive while hub identities and network structure change during development. The inferred interaction networks for root bacteria at each stage (column) and site (row) are displayed. In the top image, color indicates positive (green) and negative (orange) abundance correlations between 16S ASVs in the network. The thickness of the edges between ASVs represents the absolute value of the correlation coefficient. While networks inference from sequencing data cannot distinguish between direct microbe-microbe interactions and those mediated by, for example, the plant host, patterns can be observed in the nature of the interactions and the ASV involved in them. At all developmental stages, a majority of interactions are positive.

Figure 2.1 Bacterial interactions in *A. thaliana* remain overwhelmingly positive while hub identities and network structure change during development (continued). The bottom image highlights the network's hub bacteria, as determined by degree and betweenness centrality above the ninetieth percentile, in black. Hub identities are inconsistent: only one bacterial variant (40c4a, colored red) in the genus *Geodermatophilus* is classified as a hub in multiple root networks.

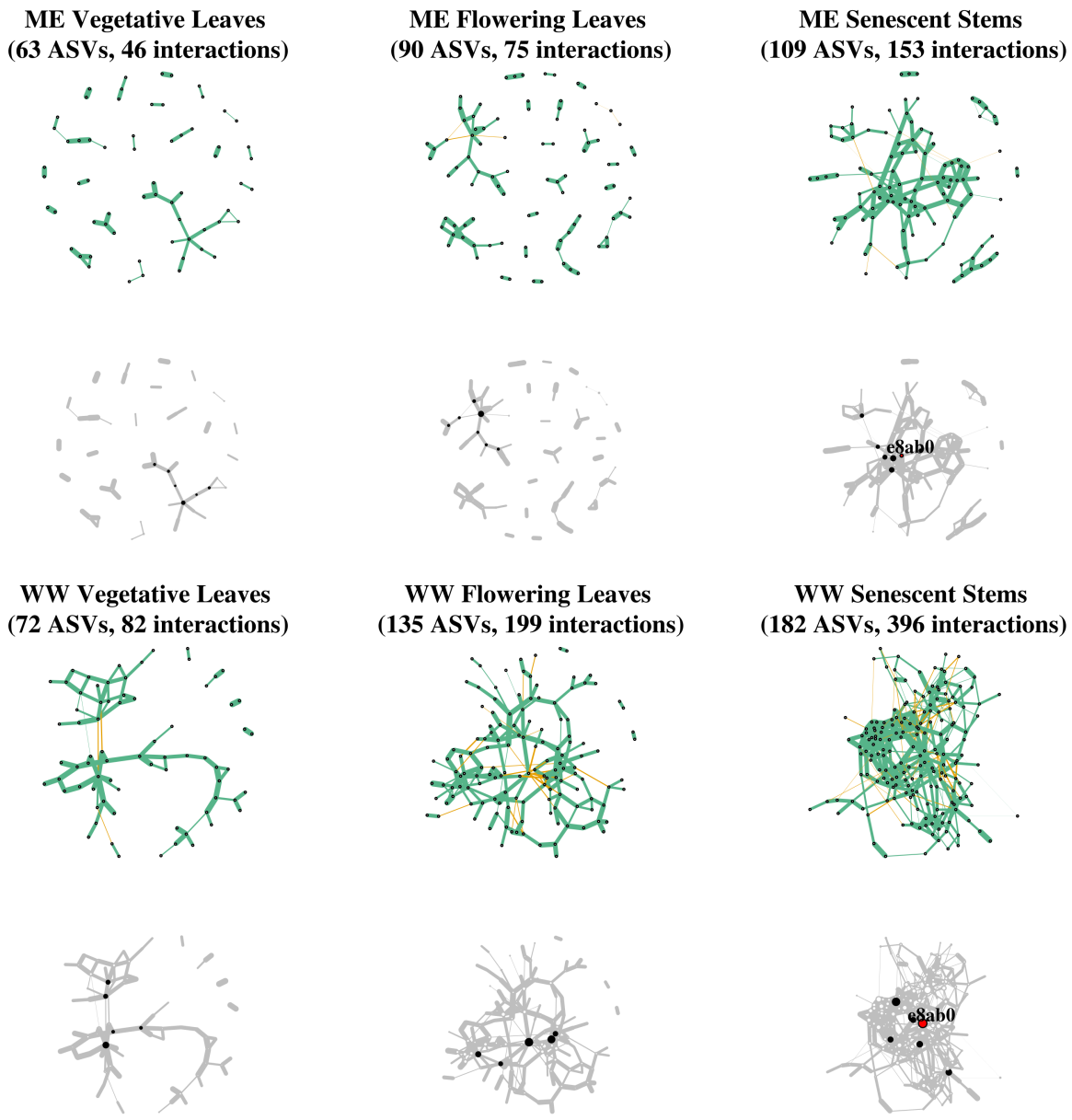

Figure 2.2 Phyllosphere bacterial interaction networks. These phyllosphere bacterial networks further support points made with the root networks in Figure 2.1. Color indicates positive (green) and negative (orange) abundance correlations and line thickness corresponds to the correlation coefficient. Hubs are highlighted in black and those conserved across networks are red and labeled with partial ASV identifiers. Only one ASV (e8ab0), in the genus *Nocardioides*, was conserved.

Table 2.1 Statistics for networks of bacterial interactions inferred by conditional dependence (top) and correlation (bottom) from 16S sequence counts in *A. thaliana* tissues.

Conditional Dependence Networks

					PD vs. PCC[a]		*Scale-free*[b]		*Modularity*[c]	*Hierarchical*[d]		
	Site	Nodes	Edges	% Pos	m	p-value	α	p-value	z score	R^2	m	p-value
Vegetative												
Roots	ME	60	55	0.78	0.0	0.5	3.0	0.96	3.7	0.9	-2.3	1.8×10^{-5}
	WW	29	20	1.00	-0.1	0.3	3.1	1.00	3.9	NA	NA	NA
Leaves	ME	63	46	1.00	0.0	0.5	4.0	1.00	3.8	1.0	-2.7	4.3×10^{-62}
	WW	72	82	0.96	-0.1	0.0	2.8	0.39	3.3	1.0	-2.4	2.2×10^{-13}
Flowering												
Roots	ME	180	261	0.90	0.0	0.3	2.3	6×10^{-7}	8.5	0.5	-1.2	8.2×10^{-10}
	WW	148	157	0.91	-0.1	0.1	3.2	0.81	7.6	1.0	-2.4	1.2×10^{-15}
Leaves	ME	90	75	0.93	0.0	0.5	3.4	0.74	4.3	0.9	-1.6	1×10^{-4}
	WW	135	199	0.83	-0.1	0.0	2.4	4.2×10^{-3}	6.8	0.5	-1.3	6.5×10^{-7}
Senescent												
Roots	ME	257	696	0.80	0.0	0.0	1.8	6.9×10^{-24}	8.8	0.4	-1.2	2.5×10^{-20}
	WW	286	490	0.96	0.0	0.0	2.3	7.8×10^{-9}	14.0	0.7	-1.6	8×10^{-35}
Stems	ME	109	153	0.95	0.0	0.4	2.5	8.5×10^{-3}	7.3	0.6	-1.7	2.5×10^{-9}
	WW	182	396	0.83	-0.1	0.0	2.0	2×10^{-10}	8.7	0.4	-1.1	1.4×10^{-9}

Correlation Networks

Vegetative												
Roots	ME	83	159	0.73	0.0	0.4	1.9	8.1×10^{-4}	2.8	0.2	-0.5	2×10^{-3}
	WW	97	448	0.60	0.0	0.5	2.0	1.9×10^{-9}	0.1	0.0	-0.1	0.49
Leaves	ME	27	17	0.94	0.0	0.1	3.2	1.00	3.2	1.0	-2.7	0.00
	WW	86	395	0.61	0.0	0.1	2.0	3.2×10^{-8}	0.1	0.0	0.1	0.35
Flowering												
Roots	ME	143	375	0.77	0.0	0.7	1.9	0.19	1.8	0.2	-0.3	1.6×10^{-4}
	WW	170	1294	0.62	0.0	0.0	2.0	0.00	13.0	0.1	0.1	2×10^{-3}
Leaves	ME	108	286	0.69	0.0	0.0	1.8	0.12	3.0	0.0	-0.1	0.51
	WW	144	1220	0.60	0.0	0.0	2.0	2.8×10^{-28}	6.3	0.1	0.2	1.7×10^{-4}
Senescent												
Roots	ME	228	1052	0.71	0.0	0.4	2.0	2.4×10^{-24}	5.4	0.0	-0.1	4.2×10^{-2}
	WW	259	1067	0.75	0.0	0.3	1.7	4.5×10^{-6}	0.5	0.0	-0.2	4×10^{-3}
Stems	ME	101	352	0.78	0.0	1.7×10^{-7}	1.7	1.1×10^{-4}	4.2	0.3	-0.4	2.8×10^{-7}
	WW	120	339	0.80	0.0	3.9×10^{-6}	1.8	5.8×10^{-3}	7.1	0.2	-0.4	2.6×10^{-5}

a. Phylogenetic distance (PD) vs. Pearson correlation coefficient (PCC) of edges: slope (m) and p-value for F statistic

b. Power law fit to degree distribution $Pr(k) \propto k^{-\alpha}$, α and Kolmogorov-Smirnov p-value

c. Z-score for empirical network modularity in distribution obtained by simulating random networks of the same size

d. Local clustering coefficient vs. node degree: coefficient of determination (R^2), slope (m), and p-value for F statistic

The high fraction of positive relationships suggests that competitive interactions do not stabilize plant bacterial communities as they do in models of the human gut microbiome (Coyte et al. 2015). A preponderance of positive interactions is perhaps not surprising given that we considered bacteria only; broader surveys of the plant microbiome indicate that negative interactions are more common between kingdoms than within them (Bergelson et al. 2019, Agler et al. 2016). Positive correlations between bacterial ASV abundances might arise from metabolic cooperation (Freilich et al. 2011), a confounding non-bacterial microbe (Bergelson et al. 2019) or a confounding abiotic factor (Naylor et al. 2017). Irrespective of the cause, we found that the strength of these inferred ecological relationships was unrelated to the phylogenetic distances between bacteria (estimated by Pearson correlation coefficients (PCC) and phylogenetic branch length (PD) respectively) (Table 2.1), although we note that 16S has poor resolution in distinguishing closely related species (Janda and Abbott 2007).

Bacterial interactions in the roots and the leaves were transient, at least with respect to our relatively coarse temporal sampling. We began by analyzing the networks constructed with conditional dependence. With each successive developmental transition in roots, 60 to 80% of bacteria present before the transition were retained (Table 2.2). These bacteria increased their number of interaction partners (Figure 2.3) over time, but less than 20% of their previous interactions remained intact (Table 2.2) at each successive stage. In leaves, a smaller fraction of the community was conserved across the transition from vegetative growth to flowering. Of bacteria in the leaf network during vegetative growth, 35% to 51% were present after flowering. Again, less than 20% of their relationships were conserved across stages (Table 2.2). Interactions

were also transient in correlation-based networks. Between 60 and 80% of nodes were retained across developmental transitions but less than 33% of their previous interactions remained intact, with one exception involving a small network for leaves early in development.

Table 2.2 Edges in bacterial networks are transient, even when nodes are retained in successive developmental stages.

		Conditional Dependence Networks					
		Stage 1	**Stage 2**	**Shared Nodes**	**% Shared Nodes**	**% Shared Edges**	**Expected % Shared Edges**
Roots	**ME**	Vegetative	Flowering	48	80	3.33	0.05
		Flowering	Senescent	126	70	6.67	0.04
	WW	Vegetative	Flowering	19	65.52	25	0.16
		Flowering	Senescent	91	61.49	11.59	0.04
Leaves	**ME**	Vegetative	Flowering	22	34.92	16.67	0.056
	WW	Vegetative	Flowering	37	51.39	11.54	0.158

		Correlation Networks					
		Stage 1	**Stage 2**	**Shared Nodes**	**% Shared Nodes**	**% Shared Edges**	**Expected % Shared Edges**
Roots	**ME**	Vegetative	Flowering	60	72.29	13.48	0.22
		Flowering	Senescent	97	67.83	16.27	0.3
	WW	Vegetative	Flowering	79	81.44	28.27	1.29
		Flowering	Senescent	103	60.59	17.37	0.855
Leaves	**ME**	Vegetative	Flowering	13	48.15	100	1.48
	WW	Vegetative	Flowering	53	61.63	32.31	2.82

Interactions were not better preserved when comparing networks from the same tissue or from the same field site or developmental stage. Generally, no more than 40% of interactions were conserved across any two networks, whether based on conditional dependence or correlations (Table 2.3, Table 2.4). The exceptions to this result involved the networks based on vegetative samples which had a small number of nodes and interactions to begin with.

The fraction of network edges expected to overlap across developmental transitions by chance was less than 2% in all comparisons and with both methods of edge inference (Table 2.2, Table 2.3, Table 2.4). With this context, the 3 to 17% of edges retained over each timestep in conditional dependence networks and the 13 to 33% retained over each timestep in correlation

networks suggest the existence of persistent relationships between bacterial abundances that shape the network structure throughout plant life. However, the majority of interactions inferred between plant-associated bacteria do not appear to endure over time even when both species involved are still present in the network. Thus, the influence of the interactions represented by those edges on microbial community structure is likely transient on the timescale of the host's life.

Table 2.3 Most edges are not shared between networks from different tissues at the same site and developmental stage.

				Conditional Dependence Networks			
		Tissue 1	**Tissue 2**	**Shared Nodes**	**% Shared Nodes**	**% Shared Edges**	**Expected % Shared Edges**
Vegetative	**ME**	Roots	Leaves	18	30	28.57	0.12
	WW	Roots	Leaves	15	51.72	0	0.16
Flowering	**ME**	Roots	Leaves	38	21.11	23.08	0.05
	WW	Roots	Leaves	66	44.5	16	0.08
Senescent	**ME**	Roots	Stems	62	24.12	13.11	0.13
	WW	Roots	Stems	92	32.17	15.12	0.06

				Correlation Networks			
		Tissue 1	**Tissue 2**	**Shared Nodes**	**% Shared Nodes**	**% Shared Edges**	**Expected % Shared Edges**
Vegetative	**ME**	Roots	Leaves	15	18.07	33.33	0.33
	WW	Roots	Leaves	51	52.58	27.04	1.5
Flowering	**ME**	Roots	Leaves	52	36.36	18.84	0.43
	WW	Roots	Leaves	87	51.18	25.65	1.91
Senescent	**ME**	Roots	Stems	52	22.81	22.45	0.74
	WW	Roots	Stems	73	28.19	21.8	0.6

Table 2.4 Most edges are not shared between networks from different sites taken from the same tissue

				Conditional Dependence Networks			
		Site 1	**Site 2**	**Shared Nodes**	**% Shared Nodes**	**% Shared Edges**	**Expected % Shared Edges**
Roots	**Vegetative**	ME	WW	6	10	50	0.89
	Flowering	ME	WW	54	30	13.04	0.02
	Senescent	ME	WW	103	40.08	5.15	0.04
Leaves	**Vegetative**	ME	WW	16	25.4	40	0.1
	Flowering	ME	WW	42	46.67	9.09	0.02
Stems	**Senescent**	ME	WW	65	59.63	16.92	0.1

				Correlation Networks			
		Site 1	**Site 2**	**Shared Nodes**	**% Shared Nodes**	**% Shared Edges**	**Expected % Shared Edges**
Roots	**Vegetative**	ME	WW	41	49.4	18.18	0.37
	Flowering	ME	WW	57	39.86	14.04	0.38
	Senescent	ME	WW	84	36.84	16.4	0.3
Leaves	**Vegetative**	ME	WW	9	33.33	100	0.46
	Flowering	ME	WW	56	51.85	26.09	0.9
Stems	**Senescent**	ME	WW	50	49.5	15.83	0.7

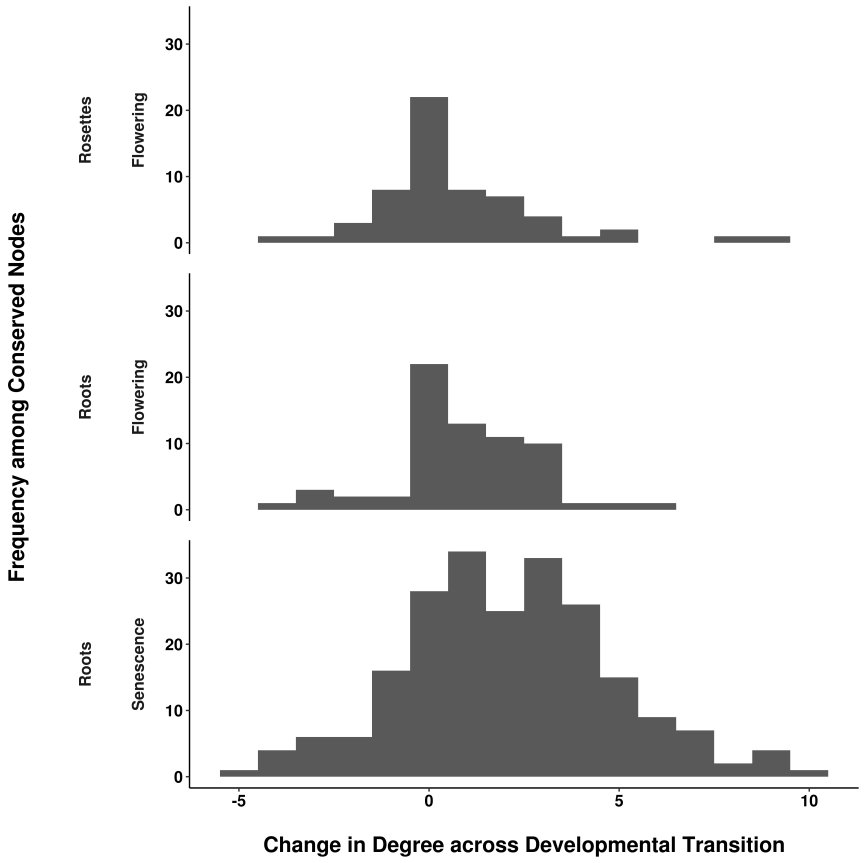

Figure 2.3 The number of ASV interaction partners tends to increase over major developmental transitions. For ASVs (nodes) remaining after a transition, the number of edges (degree) before and after the transition were compared. On the x-axis, positive numbers indicate net connections gained while negative numbers indicate net connections lost. The bar height on the y-axis indicates the frequency of conserved nodes with the corresponding number of gains or losses. Interactions gained and lost are shown for the transition to flowering in roots and rosette leaves and for senescence in roots (vertical panels).

The influence of bacteria was also inconsistent across tissues, sites, and developmental stages. Hub bacteria were identified by degree and betweenness centrality scores above the ninetieth percentile for nodes in a network. Neither of these metrics correlated with the prevalence or abundance of the bacteria in the dataset. In networks based on conditional dependence, no bacteria were consistently designated as hubs throughout plant life and only 11% persisted as hubs across any tissues, sites or more than one stage (Figure 2.4). For example, only one recurring hub was observed in root networks (Figure 2.1). The inconsistency of influential bacteria was robust to lowering the thresholds on degree and betweenness centrality that defined hubs (Table 2.5). The identities of influential bacteria were also transient across space and time in correlation-based networks, with less than 30% of hubs holding that status in more than one network (Figure 2.5). Hub microbes have been proposed to propagate the effects of the abiotic environment or host plant throughout the microbial community because changes in their abundance drive secondary effects for a large number of interacting species (Agler et al. 2016, Banerjee et al. 2018). Although hub status in computed networks is insufficient to establish keystone taxa (Röttjers and Faust 2018), the inconsistency we find in hub identities indicates that the microbes thought to play keystone roles in plant microbial communities shift over a host's lifetime.

Table 2.5 Conservation of influential taxa across networks with varying hub thresholds.

	Hubs conserved across at least two networks (%)	
Hub threshold (%)	***Conditional dependence***	***Correlation***
90	11	30
80	15	33
70	21	37

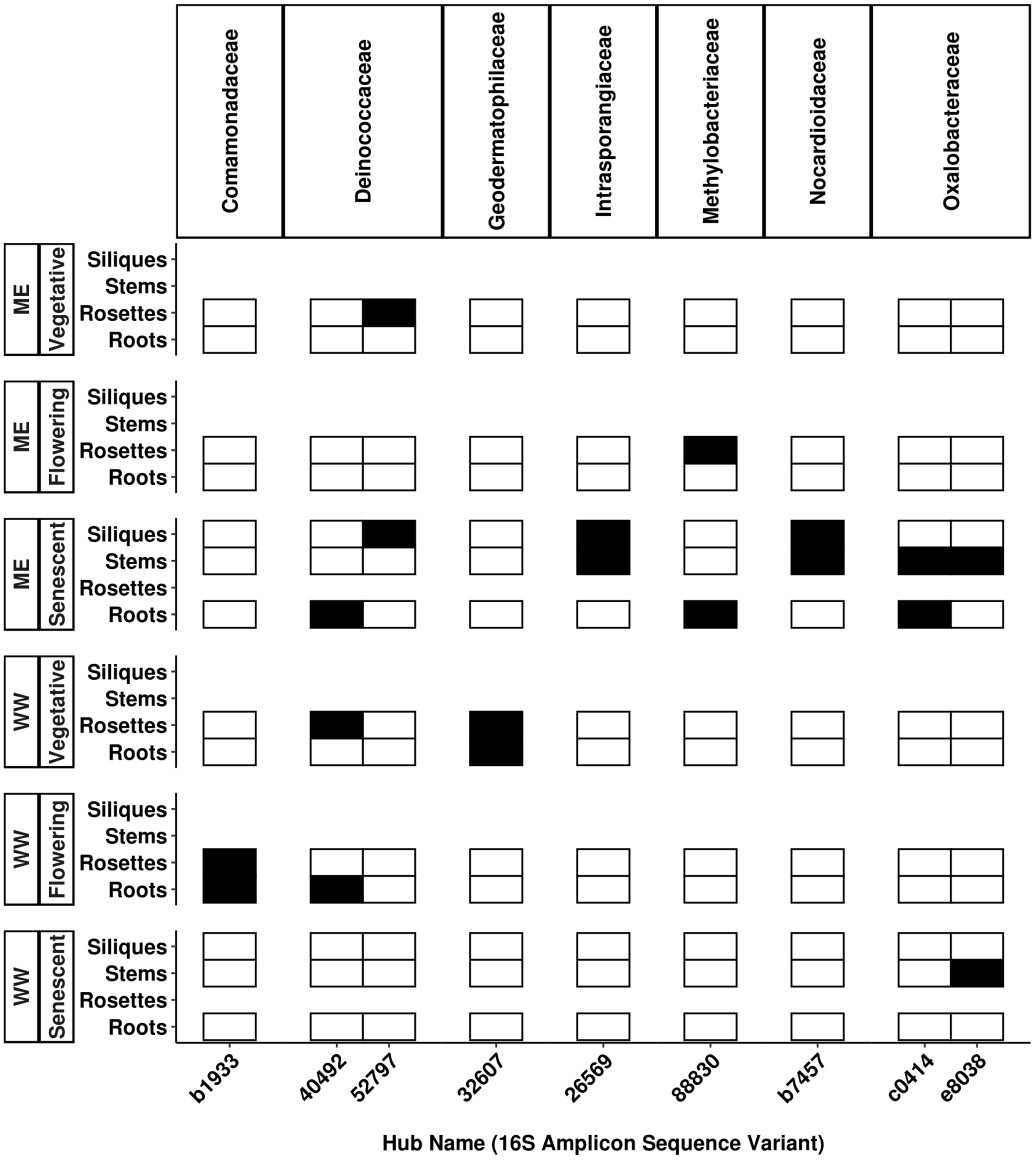

Figure 2.4 Distribution of recurring hubs (columns) in conditional dependence networks. Hubs are organized by family and compared across developmental stages, field sites (vertical panels) and tissues (rows). Black cells indicate the ASV is both present and meets the hub criteria; white cells indicate the ASV is either not present or not a hub.

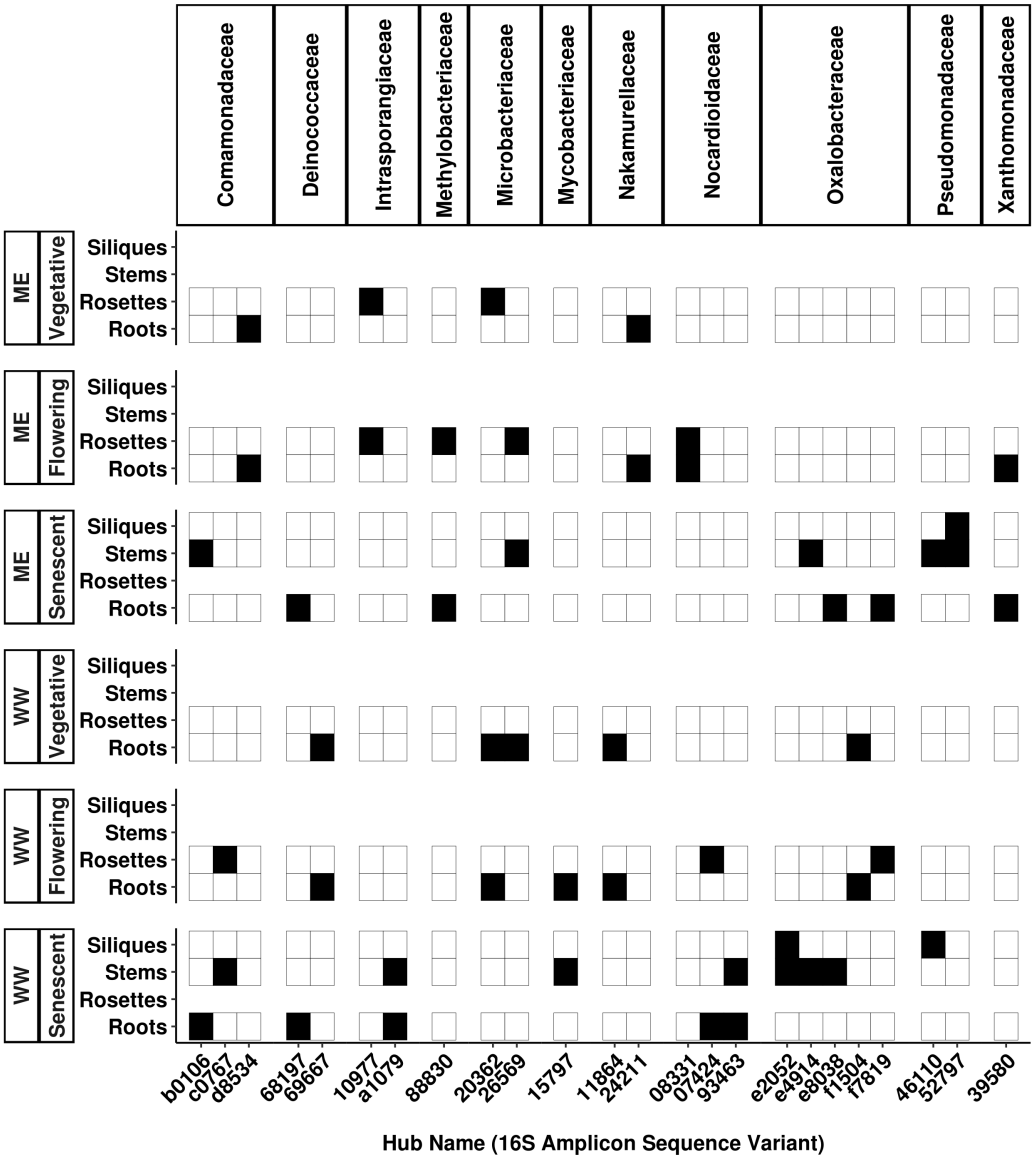

Figure 2.5 Distribution of recurring hubs (columns) in correlation-based networks. Hubs are organized by family and compared across developmental stages, field sites (vertical panels) and tissues (rows). Black cells indicate the ASV is both present and meets the hub criteria; white cells indicate the ASV is either not present or not a hub.

Networks generated from sequencing plant microbiota are used to explore the resilience of interactions at this scale to changing environments or pathogen invasions (Bohan et al. 2017, Lozupone et al. 2012, Derocles et al. 2018). When examining global network properties related to ecosystem resilience, it is important to consider how variable they are in natural environments and how they are influenced by the methods used to build networks. Notably, field studies of crop systems have come to opposite conclusions about how the same pathogen shapes interactions in the rhizosphere compared to bulk soil based on these properties (Qi et al. 2019, Choi et al. 2020). We examined how global properties of our bacterial networks like modularity, hierarchical structure, and scale-free structure changed during the course of host development.

Modularity is linked to the stability of ecological systems. High modularity destabilizes communities, especially when interactions are strongly positive (Grilli et al. 2016). As the host aged, networks based on conditional dependence became more modular relative to networks of the same size with randomized edges (Table 2.1). However, networks based on correlations lacked this temporal trend in modularity. Networks based on conditional dependence from all tissues and both sites displayed a power-law relationship between node clustering coefficient and degree throughout development, consistent with a hierarchical structure in which hubs connect modules of lower-degree nodes (Ravasz and Barabási 2003). Correlation-based networks, however, lacked this evidence of hierarchy.

We tested whether the degree distribution of each network was well fit by the power-law distribution characteristic of scale-free networks (Barabási and Albert 1999). Although the

smallest conditional dependence networks from early roots and leaves were not inconsistent with scale-free structure, some flowering networks and all senescent networks diverged from this structure (Table 2.1). Correlation-based networks also varied with respect to whether or not their degree distributions fit the expectation for scale-free structure. This finding adds to evidence that strict scale-free structure is uncommon in biological networks (Broido and Clauset 2019) and shows that systems can move in and out of this paradigm over time.

CHAPTER 3

Genome-wide association studies on the phyllosphere microbiome: Embracing complexity in host–microbe interactions

CO-AUTHORS & CONTRIBUTIONS

The text in this chapter is drawn from my contributions to a review of genome-wide association studies (GWAS) on the phyllosphere micorbiome. The review was led by Manus Thoen and written by myself, Manus, Benjamin Brachi, Andrew Gloss, Mohammad Khan, and Joy Bergelson. The full text is published as Beilsmith et al. 2019 in The Plant Journal 97(1):164-81. The work was supported by the Dropkin Foundation and the University of Chicago Biological Sciences Division. We thank members of the Ecology & Evolution Microbiome Journal club at the University of Chicago for lively discussions that inspired the review.

ABSTRACT

Environmental sequencing shows that plants harbor complex communities of microbes that vary across environments. However, many approaches for mapping plant genetic variation to microbe-related traits were developed in the relatively simple context of binary host–microbe interactions under controlled conditions. Recent advances in sequencing and statistics make genome-wide association studies (GWAS) an increasingly promising approach for identifying the

plant genetic variation associated with microbes in a community context. This review discusses early efforts on GWAS of the plant phyllosphere microbiome and the outlook for future studies based on human microbiome GWAS. We then argue that perspectives on the mechanisms, evolution, and environmental dependence of plant–microbe interactions will influence the microbiome traits that can be mapped to plant genes. We review how these choices shape the design, sampling, and phenotyping phases of GWAS pipelines.

INTRODUCTION

From mountain shrubs (Ruiz-Pérez et al. 2016) to seagrass (Fahimipour et al. 2017), and from subarctic grass (Uroz et al. 2016) to equatorial forest canopies (Lambais et al. 2006), the photosynthesizing tissues of plants are colonized by a massive and diverse set of microbial colonists, including bacteria, fungi and phages (Lindow and Brandl 2003, Rastogi et al. 2013; Morella et al. 2018, Sapp et al. 2018). Plant biologists and breeders are now confronted with the challenge of translating the wealth of knowledge gained from the study of individual plant–microbe interactions to an understanding of factors shaping the microbiome. This will require not only disentangling host–microbe from microbe–microbe interactions, but also integrating the role of environmental variation.

Genome-wide association studies (GWAS) are a powerful and flexible tool for tackling this challenge. GWAS refer to the application of association mapping, a technique that tests for statistical associations between genetic variation at a locus and organismal phenotypes, to

thousands or millions of loci throughout the genome (Brachi et al. 2011). Quantitative indices of community composition and diversity vary among plant genotypes grown in common environments (i.e. they are heritable; Peiffer et al. 2013, Horton et al. 2014), making them suitable phenotypes to interrogate through GWAS. A key strength of GWAS in this context is the flexibility of its underlying statistical models, which can compare and quantify how the effects of genotype on phenotype vary across environmental conditions, while also controlling for environmental noise that confounds associations in the field (Korte et al. 2012, Sasaki et al. 2015). Recent computational and statistical advances allow GWAS to gain power through modeling the relationships between traits (Hackinger and Zeggini 2017), offering an exciting avenue for leveraging the wealth of different microbial lineage-level and community-level phenotypes that can be extracted from phyllosphere DNA sequencing.

Most GWAS of plant–microbial interactions have focused on how plant genotype shapes pairwise interactions with a single microbial taxon. Representative examples can be found in the 35 GWAS of direct pathogen challenge collated by Bartoli and Roux (2017), or the 13 GWAS of disease resistance in maize collated by Xiao et al. (2017). GWAS have also been used to probe the genetic basis of plant mutualisms with microbes, such as the relationships between legumes and their nodule-forming bacteria (Stanton-Geddes et al. 2013, Curtin et al. 2017).

In contrast, only a single GWAS of the plant phyllosphere community has been published. Horton et al. (2014) grew a panel of 196 *Arabidopsis thaliana* genotypes in the field, and characterized the leaf microbial community by sequencing taxonomic marker genes for bacteria

and fungi. They found that additive genetic variation in the host influenced these leaf microbes at the community level. For the top eigenvectors produced by ordination analysis (principal component analysis; PCA), 9% and 11% of the variance in bacterial and fungal communities, respectively, were explained by plant SNPs. However, this result was only obtained after rare species represented by a small fraction of the total sequencing reads in each community were excluded. Without restricting the microbiome profile to the most-sequenced taxa, heritability was no longer observed for ordination outputs, and mapping SNPs to community-level traits proved impossible.

Using these restricted community datasets, Horton et al. (2014) successfully mapped both community-level traits and the presence/absence of the most abundant individual taxa in the leaf microbiome to host SNPs. Eigenvectors produced by PCA of the fungal community were associated with SNPs located in candidate genes for cell wall integrity. Species richness of the bacterial community was associated with SNPs in a set of genes enriched for trichome-related functions. But the strongest associations in the entire study were found with the presence/absence of one or a few taxa. The SNPs in these associations were located in genes involved in defense, signal transduction, kinase activity, cell walls and cell membranes. The strength of these associations, coupled with the necessity of removing rarely sequenced taxa to observe any heritability, suggests that plant influence may be best observed in the most frequent phyllosphere colonists rather than comprehensive community phenotypes.

Because a larger number of GWAS have been performed on the human and mouse microbiome,

these models offer a glimpse of the advantages and challenges that await plant biologists as the library of microbiome-associated variants grows. GWAS has successfully identified over 200 SNPs associated with variation in human microbiome composition (Knights et al. 2014, Blekhman et al. 2015, Davenport et al. 2015, Bonder et al. 2016, Goodrich et al. 2016, Turpin et al. 2016, Wang et al. 2016, Demmitt et al. 2017, Igartua et al. 2017, Kolde et al. 2018, Rothschild et al. 2018), and a few genomic regions associated with microbiome composition in mice (Org et al. 2015). The majority of these associations are with the relative abundances of individual bacterial families or genera, although some studies report associations with overall community composition or species richness (Wang et al. 2016). Genes involved in metabolism and immunity are heavily represented in regions harboring significant associations.

The majority of associations found in human and mice GWAS to date are unique to single studies (Wang et al. 2016). The failure to consistently detect the same associations across studies may reflect differences in the genetic ancestry of the mapping populations used in each study, the environments inhabited by those populations, or methods used to sequence and quantify microbial communities. Alternatively, the paucity of shared associations may arise because their effects on microbiome composition are too small to permit consistent detection with the sample sizes employed, which ranged roughly from 100 to 1000 individuals. Some of these associations may also fail to replicate because they are false positives. Thus, even in humans, our understanding of the relationship between the host genome and the microbiome remains limited.

ARGUMENT

Microbiome GWAS results will not always paint a similar picture of how host genomes influence microbial colonists, as demonstrated by the challenges of synthesizing results across mammalian studies of this kind. We argue that perspectives on how plant–microbe interactions evolve define the choice of phenotypes to be measured, and thus may be the most important determinant of the single nucleotide polymorphism (SNP) variants that GWAS identifies as mediators of host genotypic effects on the microbiome. Experimental design and sampling technique will determine the extent to which genotype by environment effects are captured and the relevance of results across different spatial and temporal scales. In addition, the number and identity of host loci associated with the microbiome will depend upon the choice to consider pairwise or community-level interactions, on whether the interacting microbes are characterized by taxonomic grouping or by traits, and on the extent to which the environmental plasticity of interactions is exposed in the sampling scheme.

Study Design

The extent to which features of the phyllosphere microbiome are shaped by host genotype or the environment is still an open question. Depending on whether they are interested in the effects of plant genotype on microbiome composition in a single environment, average effects that are consistent across environments, or changes in effects across environments (i.e. genotype-by-environment, or G × E, interactions), investigators can design studies to better control environmental effects in the lab or better quantify them in the greenhouse or field.

Environmental variation can mask observable relationships between host genes and features of the phyllosphere microbiome by creating differences in host gene expression and microbial growth across replicates of the same host genotype. Associations that might be quenched by such variation could be recovered in studies that use synthetic plant microbial communities to gather phenotypes on a diverse panel of host plants in tightly controlled environmental conditions (Bodenhausen et al. 2014, Bai et al. 2015). A study in the model plant *A. thaliana* demonstrated that host genetic effects on the leaf microbiome can be measured in synthetic communities. Differences in community composition were observed among four *A. thaliana* accessions 2 weeks after gnotobiotic leaves were sprayed with a synthetic community composed of seven microbes from the most abundant phyla sampled in natural leaf communities (Bodenhausen et al. 2014).

When it occurs across sites in the field, environmental variation can cause leaf microbial community phenotypes to map differently to plant genotypes. This principle was demonstrated in a recent field study of a perennial mustard (*Boechera stricta*) grown in a variety of environments differing in elevation, moisture, temperature, soil nutrients and plant diversity. In *B. stricta* leaves of 48 lineages, both the overall community diversity and the relative abundance of specific taxonomic groups were better predicted by site-specific rather than site-averaged host genetic effects (Wagner et al. 2016). Thus, environmental variation is a key variable that must be considered in order to fully understand the plasticity of the map between host plant genotype and leaf microbiome phenotype. Environmental variables have been manipulated in GWAS of other

environmentally sensitive plant traits (Li et al. 2010, Wu et al. 2018). For example, one GWAS of flowering time in *A. thaliana* was replicated in growth chambers simulating four distinct natural climates. Some polymorphisms associated with flowering time overlapped between climates while others were climate-specific (Li et al. 2010). Studies like this one may be similarly useful in establishing which features of the environment most influence the host variants affecting leaf communities. Studies using environment as a variable can be used to find polymorphisms that have robust effects on the microbiome across environments. Further, GWEIS (gene-by-environment genome-wide interaction study) following on such a design can be used to find the polymorphisms that cause plasticity (Hamza et al. 2011).

Unlike flowering time, plant microbiomes vary not only with geography but also with abiotic conditions within sampling sites. Abiotic conditions should be measured across samples in case they affect the chosen phenotype for microbiome GWAS. For example, a study of soybean and wheat root microbiomes across many different field sites found that pH and nitrate content correlated with community diversity (Rascovan et al. 2016).

Sampling

The microbes of the phyllosphere must be sampled, identified and enumerated in order to generate phenotypes for use in GWAS. While imaging of hybridized fluorescent probes or colony isolation and identification with Sanger sequencing can produce these phenotypes for bacteria, next-generation sequencing approaches are the most high-throughput and comprehensive options for characterizing leaf microbiomes. The most common form of

microbiome data is sequences for variable regions of a conserved marker gene like bacterial 16S rRNA or fungal ITS. After DNA is extracted from lysed leaf tissue, a portion of the marker gene is amplified from the sample of environmental DNA by polymerase chain reaction (PCR) and sequenced on a next-generation platform such as the Illumina MiSeq or HiSeq. An alternative to marker gene sequencing is the collection of metagenomic data. In this approach, a library of environmental DNA is prepared from leaf lysate and sequenced without targeting specific types of microbes or particular genes using shotgun sequencing. Many methodological considerations for these approaches are not specific to phyllosphere sampling of microbial communities and have been reviewed elsewhere (Knight et al. 2018). Below we elaborate on phyllosphere-specific sampling concerns.

A leaf is a heterogeneous environment composed of microhabitats defined by abiotic conditions and proximity to physical features like stomata. Fluorescence imaging indicates that populations of some phyllosphere bacterial species grow to different densities in these microhabitats (Burch et al. 2013, Peredo and Simmons 2018). These findings support the idea that leaf microhabitats could harbor different microbial communities in nature (Edwards et al. 2015). Because the leaf is heterogeneous, sampling designs should ensure that within-leaf variation will not confound the relationship between SNPs and microbe communities. Microbiome samples should be collected either from consistent positions on leaves or by homogenizing the leaf tissue before extraction to ensure that any within-leaf variation is well sampled so that it does not confound associations between plant genotype and microbial community phenotype.

Leaf tissue can be homogenized by grinding with a mortar and pestle or by bead beating. However, first it is useful to separate epiphytic and endophytic microbes because transcriptome data suggest that host plants and microbial colonists have different interactions at these locations. For example, gene expression patterns of the phyllosphere bacterium *Pseudomonas syringae* are distinct when it is grown on leaf surfaces, where genes involved in motility are significantly more transcribed, relative to the apoplast, where genes related to iron and nitrogen uptake are more transcribed (Yu et al. 2013). To preserve separate inner and outer communities, vortex washes, sonication in buffer and surface sterilization have each been used to first collect 'epiphytes,' although the steps employed are highly variable between studies, and their efficacy at removing tightly adhered surface microbes is not well characterized.

One fundamental limitation of phyllosphere sampling is its destructive nature. Unlike many microbiome samples from human hosts, which can be collected repeatedly from the same individuals via stool samples or cheek swabs, the leaf microbiome, with its lower bacterial load, requires the harvest of whole tissue, especially in small species or young seedlings. Even when large plants are sampled repeatedly, the possible effect of tissue wounds on microbiome assembly must also be considered. Tissue wounds can trigger immune responses by releasing damage-associated molecules in the plant and can also induce phytohormone synthesis (Savatin et al. 2014). *A. thaliana* mutants lacking phytohormone synthesis or signaling-related genes have distinct root communities in natural soils (Lebeis et al. 2015), suggesting that altered levels of phytohormones could affect community assembly in nature. In addition, the age and maturity of plants and leaves at the time of sampling can influence the leaf microbial community (Meaden et

al. 2016, Wagner et al. 2016).

In metagenome and marker gene datasets, DNA extraction and sequencing methods can bias the measured relative abundances of microbial taxa. This bias has been quantified by comparing relative abundances from sequencing data to expected abundances given samples with a defined mix of bacterial cells from diverse taxa (Morgan et al. 2010, Fouhy et al. 2016). In marker gene data, the bias created by the choice of target sequences for amplification is an even more important concern. Primer choice affects how much chloroplast DNA is amplified, the number of reads that can be classified, and the relative abundances of different taxa in phyllosphere sequence samples. If amplification and sequencing methods severely skew the abundances of taxa influenced by host genotype in the community profile, then differences at these steps could limit our ability to synthesize GWAS findings across work in the model system *A. thaliana* and across plant systems in general.

Phenotyping

The most important aspect of any GWAS pipeline is defining which traits will be mapped to the host genome. This in turn determines the type of raw phenotype data to collect. To help determine the type of microbiome sequencing to perform and the way in which data will be analyzed, we suggest focusing on three main questions.

- Is the research question better addressed by associating host genes with microbial taxa or with the biological functions encoded in microbial genomes?

- If the research focuses on taxa, what mode of host–microbe interactions is more likely: diffuse interactions between hosts and microbial communities or targeted interactions between hosts and specific microbes in the community?

- How important is the environment in shaping host–microbe interactions, and do the hypotheses under consideration require explicit investigation of the impact of host genotype-by-environment interactions on microbial communities?

In the following three sections, we discuss how the answers to these three questions impact the selection of phenotypes for microbiome GWAS, as well as how diverse approaches can be used to build a richer picture of how host plants shape the microbial communities of their leaves.

Phenotyping Microbial Taxa or Microbial Functions

Instead of focusing on how host genotype influences a single microbial taxon, microbiome GWAS focuses on how host genotype affects the overall community. The most intuitive way of characterizing microbial communities is to identify microbial species or strains, and quantify their abundances. In practice, the most common strategy is to sequence PCR-amplified marker genes bearing taxonomically relevant information, and then group these sequence reads into operational taxonomic units (OTUs) or amplicon sequence variants (ASVs). OTUs are generated by clustering sequence reads based on similarity thresholds (Westcott and Schloss 2015), while ASVs are generated by modeling and correcting sequencing errors so that even single-nucleotide

differences can be used to distinguish biological variants (Callahan et al. 2016).

In order for the abundances of OTUs or ASVs to be meaningful community phenotypes, two conditions must be fulfilled. First, the genes targeted must provide sufficient resolution of taxa. Although regions of 16S ribosomal RNA (rRNA) have been used most often, they afford poor resolution at the level of genera or finer clades. This limitation can be circumvented by using faster-evolving genes to gain deeper taxonomic resolution (Bartoli and Roux 2017). For example, gyraseB sequences allow pathogenic and commensal OTUs within a microbial genus to be distinguished (Barret et al. 2015). The tradeoff in using less conserved genes is that multiple or degenerate primers may be necessary to sample all genera of interest. Second, OTU- or ASV-based taxonomic units must harbor phenotypic differences relevant to plant–microbe interactions. In particular, GWAS based on these phenotypes are likely to identify associations driven by microbial motifs conserved within families or orders.

Obtaining sufficient taxonomic resolution with universal primers is a key limitation of marker gene microbiome characterization. Fortunately, untargeted whole metagenome sequencing has become a viable alternative. Untargeted data from whole communities can be used to find known genes or taxa in the sample or give sufficient coverage to reconstruct whole genomes of community members. Untargeted data can offer finer taxonomic resolution by covering more of each microbe's genome. These read-based and assembly-based approaches are described by Knight et al. (2018), and reviewed by McIntyre et al. (2017) and Vollmers et al. (2017), respectively. Subsequently, genomes assembled de novo from metagenomic reads, or obtained

from a collection of publicly available genomes for plant-associated microbes (Levy et al. 2018), can be used as references to delineate and quantify the abundance of up to hundreds of unique strains per bacterial species in each phyllosphere DNA sample (Albanese and Donati 2017, Smillie et al. 2018). A challenge for these techniques, however, is separation of host and microbe DNA, as the typically larger size of host genomes limits the coverage of small microbial genomes when sequencing mixtures of host and microbe DNA. In addition, the completeness of reference databases limits the ability to assign reads to genes or to assemble fragmented microbial genomes by comparative methods.

Irrespective of whether amplified gene fragments or untargeted shotgun sequencing is used to identify taxonomic units, these methods enable a calculation of the relative abundances of members of the microbial communities. However, absolute abundances provide greater information content for understanding the interactions between hosts and microbes (Vandeputte et al. 2017). One way to overcome this limitation is to estimate microbial cell density in samples using flow cytometry, for example (Props et al. 2017). Taxa counts can then be converted to absolute abundances and used to compare the population size of a particular microbe between plant genotypes.

A more inherent limitation arises from the possibility that lateral gene transfer among microbes may obliterate the taxonomic signal associated with ecologically important microbial genes. Comparative studies have revealed that bacterial genomes contain both a shared core and a divergent accessory component. Core genes can encode essential cell functions or clade-specific

adaptive traits (Lassalle et al. 2011). Accessory genes are acquired through lateral transmission by conjugation, transformation and transduction (Soucy et al. 2015), and can include genomic islands with virulence factors (Araki et al. 2006). While the frequency of these transfers is difficult to quantify, over evolutionary time they can contribute thousands of genes to microbial genomes that are not conserved among closely related taxa (Philippe and Douady 2003). The movement of these genes across microbial phylogenies creates incongruities between the history of the transferred genes and the rest of the genome. Because laterally transferred genes are phylogenetically incongruent, taxonomic groupings based on the similarity of a conserved marker gene will not correspond to the presence or absence of these genes. As a result, microbiome GWAS may be less effective than a disease case-control GWAS for detecting resistance genes that target effector proteins, which often show a highly variable presence/absence pattern within species groups indicative of horizontal transfer.

An alternative approach to phenotyping a microbial community is to focus on functions present in the microbial community rather than taxa (Bonder et al. 2016; Levy et al. 2018). This approach could identify associations driven by microbial traits irrespective of whether they are found in the core or accessory genome, and regardless of whether they are phylogenetically congruent within lineages. Whole-genome sequencing of leaf microbial communities will be a superior phenotyping method if metabolic, competitive and virulence traits are more important than phylogenetic identity for defining the effect of a microbe within a leaf community. Furthermore, if bacterial genomes are well-annotated, the association of genetic variants with specific microbial genes rather than taxonomic assignments would give more insight into the

possible biological mechanisms underlying the associations.

If microbial functions are to be used as a phenotype, they can be inferred either from the microbial genomes or from plant responses to the microbial community. In the first approach, unassembled reads from whole-genome sequencing are annotated (Keegan et al. 2016). The annotated genes are grouped into functional systems like biosynthetic or signaling pathways based on databases like KEGG (Kanehisa et al. 2017). Then, the frequencies of these systems can be compared across plant genotypes. This approach was used to find microbial functions that were more frequent in the rhizosphere of a fungus-resistant bean cultivar than in the rhizosphere of a related, susceptible cultivar (Mendez et al. 2018). In particular, samples from the resistant cultivar had a higher abundance of microbial sequences associated with the production of rhamnolipids and phenazines. In the second approach, plant physiological responses that are linked to microbial community function are measured in the presence of defined synthetic communities. An example of a plant phenotype linked to microbial community function appears in Paredes et al. (2018). In this study, bacterial strains were grouped into broad functional categories based on how they affected the response of *A. thaliana* to phosphate starvation (as measured by phosphate accumulation in shoots). Groups of eight or nine strains from each functional category were then combined into synthetic communities that were used to infect host plants; plant phosphate accumulation was then measured. Community membership explained over half of the variation in plant phosphate accumulation, suggesting that this phenotype is closely tied to microbial functions in the plant. If applied to GWAS, plant phenotypes tied to community functions in planta could help identify host genes responsible for structuring the leaf

community to favor or limit specific functions instead of taxa.

Targeted Versus Diffuse Plant–Microbe Interactions

When investigating the host genes shaping the relative abundance of microbial taxonomic units (OTUs or ASVs), the associations detected will depend upon which aspects of the microbial community are mapped. This choice will be driven by one's perspective on whether plants primarily influence their microbiomes through targeted interactions with key taxonomic groups or diffuse interactions with entire communities. Two emerging paradigms for plant relationships with leaf microbial communities can be used to illustrate the difference between these perspectives: hub taxa versus ecosystems on a leash (Foster et al. 2017).

Hubs are identified as highly connected nodes in networks constructed to represent interactions between taxa in the microbiome. Because the interactions between microbes cannot be directly observed in nature, they are inferred from relationships among taxa relative abundances as measured by sequencing reads. The edges of the network represent these relationships in the data, and the nodes they connect each represent a clade of bacteria grouped by a threshold of sequence similarity. The degree, defined as the number of edges connecting to a node, and measures of its centrality in a network can be used to classify a subset of microbial taxa as hubs; these hubs are believed to be involved in some of the most important putative microbe–microbe interactions in the plant community. Inferring networks based on microbiome data from amplicon sequencing is an active field of research; the main challenge being to control for compositional effects (i.e. biases induced by the finite number of reads per sample). Current

methods rely on computing sparse correlation or partial correlation matrices among transformed OTU relative abundances (Friedman and Alm 2012, Kurtz et al. 2015).

Agler et al. (2016) sequenced bacterial, fungal and oomycete marker genes in leaves of wild *A. thaliana* from five sites, as well as the leaves of three *A. thaliana* accessions in common garden plots. They constructed a network of the community members based on correlations among relative abundances, and then used measures of closeness centrality, betweenness centrality, and degree to select nodes as putative hubs. One isolate of a fungal hub taxon, *Albugo*, was used to infect plants and test the effects of hub presence on other phyllosphere biota as measured by changes in community diversity and variability relative to communities on control plants. Communities containing the potential hub taxon were found to have lower richness and to be less variable among replicates than control communities. The authors hypothesized that *Albugo*, and hub taxa in general, stabilize plant microbial communities through strong negative interactions with other microbes. In this framework, plant interactions targeting a hub taxon can produce propagating effects that influence the entire microbial community.

In order to find genes responsible for plant interactions with microbes of ecological importance, investigators may choose to apply GWAS to the relative or absolute abundance of candidate hubs identified through network analysis of whole-community data. If plants shape their microbiome through pairwise interactions with hub species that in turn shape the community through microbe–microbe interactions, phenotypes focused on these hub taxa may be better explained by host genotype and facilitate detection of the underlying genetic variants through GWAS.

An alternative scenario is that leaf microbial communities are shaped by a large number of diffuse interactions. Just as the cumulative action of weakly repressive miRNAs can stabilize the transcriptome and canalize gene expression (Ma et al. 2018), small plant–microbe effects in aggregate may have a significant effect on microbial load and diversity in the phyllosphere. Plants may be selected to act diffusely to cultivate and stabilize microbial communities if these communities provide a benefit, either directly or indirectly by preventing colonization of pathogens. This idea has been described as the plant holding the microbial ecosystem on a leash (Foster et al. 2017). While it is not yet clear which measure(s) of a microbiome makes it most resistant to invasion, or provides robust direct benefits, it is clear that comprehensive community phenotypes, rather than individual taxa abundances, would be the most appropriate phenotypes for characterizing the genetic architecture of the ‘leash’ through GWAS. Measures of community diversity or evenness could capture some of these effects, just as species richness captured the many microbe–microbe interactions of the *Albugo* hub within the leaf community. Temporal sampling schemes or the construction of interaction networks can also allow inference of community stability in terms of species turnover or persistence.

The goal of many microbiome studies is to identify manipulations of the microbiota that may enhance plant traits like yield or pathogen resistance (Chaparro et al. 2012). GWAS can be used to identify plant alleles that support the assembly of ‘good’ microbiomes that provide these benefits to plants. Selecting appropriate phenotypes for these GWAS will depend on whether targeted taxa or diffusely selected community properties are thought to be the means by which

microbiomes provide benefits to plants. For example, one way in which specific microbes can improve pathogen resistance is by producing elicitors of defense responses in plants. Members of the bacterial genus *Bacillus* elicit plant responses that reduce the severity of subsequent pathogen infections (Kloepper et al. 2004). There is even evidence that *Bacillus subtilis* can be chemoattracted to root exudates, suggesting that the plant recruits bacteria with protective effects (Rudrappa et al. 2008). For cases like this, when targeted recruitment of specific taxa for plant benefits seems likely, GWAS phenotypes based on the abundances of just a few taxa would best capture the plant variation relevant to the interaction. However, there is a long record of ecological studies linking increased species diversity to decreased temporal variability of biomass and species composition (McCann 2000, Allesina and Tang 2012). Community evenness, or a lack of dominance by a small number of taxa, has been linked to lower invisibility (Hillebrand et al. 2008). For cases where a diverse and even community is thought to prevent invasion or limit the growth of pathogenic microbes, phenotypes like Faith's phylogenetic distance or the Shannon evenness index would be a better choice to capture relevant host variation in a GWAS.

Plant–Microbe Interactions in Natural Environments

Although it is well established that the environment affects genotype–phenotype mapping in plants, it is unclear how this effect influences the associations found with microbiome traits. Environmental effects may be mediated by modifiers of the causal loci. In these cases, genome-wide tests of interactions that establish the contributions of both a SNP and its interaction with an environmental variable could help to identify the modifier loci (Hamza et al. 2011). Another

possibility is that overlapping sets of genes map to microbiome phenotypes in different environments. While some of these genes will only show associations in some environments, others will show robust signal across many environments. A comparison of significant associations identified in independent GWAS that have been performed in multiple environments will allow the identification of environment-independent and environment-dependent host effects on leaf communities.

The scenario of both environmentally sensitive and environmentally robust variation in host influence originally appeared to apply for the trait of flowering time in *A. thaliana* (Li et al., 2010). But further study (Brachi et al. 2010) revealed that very different environments could produce association results with very little overlap for this trait. Brachi et al. (2010) measured flowering time over two years in the field for tens of thousands of plants from natural lines, RILs and NILs in order to perform QTL mapping and GWAS. The phenotypes gathered from plants in the field mapped to different loci than those gathered from greenhouse plants, with limited overlap in the genes where candidate SNPs fell. The GWAS on field samples detected many candidate genes that were not associated with flowering time variation among accessions in the lab, including a number of circadian-clock-related genes. The unpredictable and variable environment in the field may cause certain regulatory variation to be more important in complex traits than is revealed in the relatively stable conditions of a greenhouse or incubator. If this is the case for microbiome-related host variation, then conducting phenotyping in the field will be of particular importance to finding candidate alleles with ecological relevance or utility field conditions.

CONCLUSION

There are precious few studies mapping the effects of host genetic variants on the plant microbiome. From work on the human microbiome, however, we know that there are few consistent association peaks across studies. At this point, the extent to which this lack of consistency between studies is biological versus methodological is unclear. Inconsistencies could arise from biological differences in the diversity of host genotypes included in the mapping panel or the interactions of uncontrolled environmental factors with host and microbe genotypes. They could also arise from methodological differences in how microbiome sequence data are reduced to phenotypes or in power due to the sample sizes and statistical approaches used. If broad patterns in host–microbe interactions are to be discovered, it is important that care is taken in designing future plant GWAS studies to ensure that the detected associations can be meaningfully compared across studies. We highlight below a few areas of basic research that we expect will help inform successful GWAS design for leaf microbiome traits.

Elucidating the genetic bases of leaf microbial community variation will require that we first establish how much of the variation in taxon abundances and microbial community properties are heritable. Without this knowledge, it will be difficult to focus phenotyping on community traits or members that are likely to be influenced by host genetic variation. We must also understand the environmental scales at which microbiomes show consistent patterns in composition to select appropriate mapping populations. Efforts characterizing environmental variability in the phyllosphere microbiome will prove invaluable in improving the power and

success of mapping efforts.

That said, the most important choice that an investigator must make is the trait to be mapped. We still have only limited information on which aspects of microbial communities influence their function. Are community measures such as diversity or richness meaningful predictors of the behavior of communities, or should one focus on particular genes, functional systems, strains or genera? Establishing whether the plant fitness benefits conferred by microbes arise from community-level properties or specific taxa will be particularly important for determining which phenotypes can be used in GWAS to identify loci for use in breeding or engineering for properties like yield or pathogen resistance.

If it is found that strain-specific interactions provide most of the selective pressure on plants and their interactions with microbial partners, techniques that move beyond traditional GWAS to target these specific interactions will likely prove most fruitful. For example, Wang et al. (2018) used a two-way mixed effects model to map simultaneously both host and pathogen genomic variation shaping their interaction. They found non-overlapping regions of plant genomes conferring resistance to *Xanthomonas arboricola* or *Xanthomonas campestris*. In addition, the regions conferring resistance to the latter had different levels of specificity for strains within the pathogen species. The same approaches could be applied to provide an initial picture of how genes shape the interaction between host plants and their beneficial or commensal microbial inhabitants.

CHAPTER 4

Exploring plant-microbe interactions by integrating transcript abundance data with a plant primary metabolic network

CO-AUTHORS

The work in this chapter was a collaboration with Samuel Seaver at Argonne National Laboratory. The tools used to analyze RNA sequencing data in the Department of Energy Systems Knowledgebase (KBase) were created by Sunita Kumari, Vivek Kumar, Doreen Ware, Christopher Henry, Robert Cottingham, and Adam Arkin. My work was supported by the University of Chicago Biological Sciences Division and NIH T32 GM 7197.

ABSTRACT

Metabolism is a complex interface between plants and the microbes they host. Primary metabolism, the biochemical reactions involved in cell growth and maintenance, is altered in response to pathogen invasion as well as during the establishment of symbioses with microbes. These metabolic changes have consequences for ecosystem processes in natural systems and yield in agricultural ones. Here, we integrate a network of primary metabolism with gene expression data from inoculated plants in order to learn how microbial partners and pathogens shape the flow of energy in model and crop systems. Case studies with data from model and crop

plants suggest common responses of lipid and lignin biosynthesis to the pathogen *Pseudomonas syringae* as well as common responses of one-carbon metabolism and thiamine biosynthesis to eukaryotic microbes.

INTRODUCTION

Metabolism is one of the most complex interfaces between plants and their antagonists in systemic responses to invasion. In a variety of plant species, metabolites are differentially abundant in infested or infected plants compared to healthy ones (Tan et al. 2004, Marti et al. 2013, Alamgir et al. 2016, López-Gresa et al. 2010). These changes in metabolic profile composition indicate that plant metabolic systems are reconfigured in response to biotic stress. Secondary metabolites such as phenylpropanoids, terpenoids, and glucosinolates accumulate in the tissues of plants under attack to deter antagonists (Kliebenstein 2004). Primary metabolism, consisting of the biochemical reactions required for cell maintenance and proliferation, is also remodeled during biotic stress, with consequences for the production and allocation of plant energy (Rojas et al. 2014, Bolton 2009). For example, photosynthesis in tomato leaves, as measured by chlorophyll fluorometry, is reduced at the sites of bacterial (*Psuedomonas syringae* DC3000) or fungal (*Botrytis cinerea*) infections but increased in the surrounding tissue (Berger et al. 2007). Sugar storage in tomato leaves, measured by chromatography, decreases after fungal (*B. cinerea*) infection (Berger et al. 2004). In the model plant *Arabidopsis thaliana*, foliar sucrose decreases and aromatic amino acid content increases after bacterial (DC3000) infection (Ward et al. 2010).

Plant primary metabolism is also responsive to the presence of microbial partners, such as symbiotic microbes housed in the roots. Primary metabolites diffuse from the roots into the soil and there is evidence that this process is sensed and regulated by the plant (Canarini et al. 2019). These exudates provide resources for rhizosphere microbiota, which in turn have consequences for plant health (Pascale et al. 2020). Metabolism within root tissues is reconfigured in response to colonization by arbuscular mycorrhizal fungi (Rivero et al. 2015), shaping the host niche for the microbes in return for nitrogen fixation.

Changes in primary metabolism in response to microbes are difficult to study given the breadth of carbon skeletons and functional groups present in primary plant metabolites. Polar and nonpolar, volatile and nonvolatile, and high and low molecular weight compounds are best extracted, ionized, and detected by different mass spectrometry methods (Piasecka et al. 2019, Rico et al. 2011). This complexity renders the plant primary metabolome a difficult phenotype to capture. A broader range of compounds can be analyzed with nuclear magnetic resonance spectroscopy, but biologically relevant compounds at low concentrations may go undetected with this approach (Marshall and Powers 2017). Even with a combination of these methods, the plasticity and natural diversity of plant metabolomes still complicates comparative studies of stress-induced metabolomic profiles across genotypes and cultivation conditions (Abdel-Farid et al. 2009).

In a parallel approach, the reconfiguration of primary metabolism is investigated through the

enzymes that are differentially regulated in response to the presence of microbes. Genes with significantly different transcript counts between invasion and control conditions are annotated according to databases like the GEO (Barrett et al. 2012, Li et al. 2018). These annotations are used to assign genes to areas of metabolism and make inferences about their differential regulation between control plants and those treated with microbial cultures (Schwachtje et al. 2018). A limitation of this approach is that metabolic pathways function not as discrete units but as part of a system. Increases in output from one branch of primary metabolism can have different biological implications depending on the behavior of other branches. Furthermore, variable annotations in the plant phylogeny can make it difficult to translate basic research from model systems to crops (Nelissen et al. 2014).

Microbe-induced metabolic reconfiguration can instead be studied at the system level by integrating transcript abundance data with a metabolic network (Brown et al. 2004). The integration approach allows the effects of biotic stress on metabolic systems to be compared across divergent plant genomes. The functional, homology-based, curated metabolic annotations in PlantSEED can be used to reconstruct models of primary metabolism from multiple plant genomes in a consistent manner (Seaver et al. 2014, Seaver et al. 2015, Seaver et al. 2018). We can then identify and compare the metabolic subsystems and reactions responsible for stress response in each species.

This integration can identify findings that translate between model and crop species. For example, stress responses may differ between the model *Arabidopsis* and crop species due to

intense artificial selection on metabolic pathways (Miao et al. 2017). In this paper, we examine the system-level metabolic responses of *A. thaliana* and tomato (*Solanum lycopersicum*) leaves to bacterial (*P. syringae* DC3000) and fungal (*B. cinerea*) infection by integrating metabolic models with expression matrices produced from public mRNA-seq data.

The integration approach also allows the effects of different microbes to be compared within plant species, facilitating the identification of enzymes subject to synergistic or antagonistic selective pressures. In this paper, we compare the system-level metabolic responses of *S. lycopersicum* leaves to pathogenic and nonpathogenic *Pseudomonas* strains and of *S. lycopersicum* to three eukaryotic pathogens. The goal of these case studies is to integrate transcript abundance data with a biochemical network to highlight specific reactions in primary metabolism associated with to the presence of microbes in both model and crop plant species.

METHODS

Public Data and Analysis Notebooks

The data processing and analysis for this project are available in the form of persistent electtronic notebooks from the Department of Energy Systems Biology Knowledgebase (KBase). The analysis notebooks include:

A tutorial on generating expression matrices from Illumina RNA-seq data:

https://narrative.kbase.us/narrative/67235

A tutorial on integrating expression data with a network of plant primary metabolism:

https://narrative.kbase.us/narrative/67339

Analysis for the case studies on public data presented in this chapter:

https://narrative.kbase.us/narrative/73012

Selecting Plant-Microbe Case Studies

Plant metabolism is reconfigured in response to interactions with microbial partners and pathogens. Metabolic reconfiguration mediated by transcriptional changes can be detected by comparing genome-wide mRNA sequence (mRNA-seq) profiles between infected and control plants. From NCBI, we imported public Illumina mRNA-seq data from *Arabidopsis thaliana* and *Solanum lycopersicum* (tomato) samples exposed to microbes or to sterile buffer as a control.

We selected four datasets to demonstrate the integration of transcript abundance data with a network of plant primary metabolism. The datasets were originally collected to find *A. thaliana* and *S. lycopersicum* genes that were differentially expressed in response to biotic interactions with bacterial or eukaryotic microbes. Each set included sequencing data from replicate plants subjected to a microbe treatment or an appropriate control treatment. Each replicate, post quality control, included tens of millions of genome-wide mRNA reads obtained on Illumina sequencing platforms.

Since the datasets were collected for different investigations, they were not collected with perfectly uniform procedures or from plants grown and infected in completely consistent conditions (Table 4.1). However, the invasion and control treatments within each study are

directly comparable. When comparing data between studies, we processed the Illumina reads to obtain consistent length, depth, and quality across datasets (Table 4.2).

	Howard et al. 2013	**Rosli et al. 2013**	**Coolen et al. 2016**	**Rezzonico et al. 2017**
Paper	10.1371/journal.pone.0074183	10.1186/gb-2013-14-12-r139	10.1111/tpj.13167	10.1038/s41598-017-04792-5
NCBI SRA accession	SRP010938	SRP028553	PRJNA315516	PRJEB21223
Plant genotype	*Arabidopsis thaliana* Col-0	*Solanum lycopersicum* L, cv. Ailsa Craig	*Arabidopsis thaliana* Col-0	*Solanum lycopersicum*, cv. Heinz 1706
Growth substrate	soil	peat and vermiculite 2:1	river sand and soil 1:1	soil
Growth conditions	greenhouse	greenhouse, 25/19°C (day/night), photoperiod 12h, photon flux density 200 µmol m^{-2} s^{-1}, relative humidity 70%	growth chamber, 21°C , photoperiod 8h, photon flux density 100 µmol m^{-2} s^{-1}, relative humidity 70%	greenhouse with open windows, 20 – 26°C with highs up to 40°C, photoperiod 16h, minimum 80 kW m^{-2}
RNA extraction	Qiagen Plant RNeasy	unspecified	Qiagen Plant RNeasy	NucleoSpin RNA Plant
Developmental Time	4-6 weeks	4 leaves	5 weeks	2 week cuttings
Invader species	Pseudomonas syringae pv tomato DC3000	Pseudomonas syringae pv tomato DC3000	*Botrytis cinerea* strain B05.10	*Botrytis cinerea* strain T4; *Oidium neolycopersici; Phytophthora infestans* strain K5276
Invader density	10^7 cfu/ml	OD600 = 0.1 (usually 10^6 – 10^8)	10^5 spores/ml	10^6 spores/ml; 10^4 spores/ml; 10^5 sporangia/ml
Inoculation method	infiltration	dip	4 × 5ul spore droplets	8 × 10µl drops applied to 15 leaflets
Tissue	leaf	leaf	leaf	leaf
Time of harvest (hours post inoculation)	12	19	12	24
Plants pooled for samples	20	3	4	8
Biological replicates	2	3	3	3
Sequencing	Illumina paired-end 75bp	Illumina single-end 50bp	Illumina single-end 100bp	Illumina paired-end 50bp

Table 4.1. Details of the public datasets used to compare transcriptional responses to infection between pathogens and plant species.

To compare the responses of plant primary metabolism to a common invader, *P. syringae* strain DC3000, we used datasets from experiments by Rosli et al. 2013 and Howard et al. 2013 in *A.*

thaliana and *S. lycopersicum*, respectively. For the same two plant systems, the Coolen et al. 2016 and Rezzonico et al. 2017 datasets were used to compare the responses of primary metabolism to infection by the fungus *B. cinerea.* In *S. lycopersicum*, Rosli et al. 2013 data were used to compare responses to pathogenic and nonpathogenic bacterial strains while Rezzonico et al. 2017 data were used to compare responses to three different pathogenic eukaryotic microbes.

Quality control

Read libraries for the selected datasets were downloaded from the NCBI Sequencing Read Archive using the accessions listed in Table 4.1. Initial quality control statistics for the libraries were generated with FastQC (version 0.11.5). Since the FastQC results indicated that regions of low quality and Illumina primers were present in the reads, Trimmomatic (version 0.36) was used to remove sequencing adaptors, crop reads, and filter reads based on quality scores and length. Adapter clipping was performed for Illumina TruSeq primers. For adaptor clipping, the seed mismatch was set to 2, the simple clip threshold was set to 10, and the palindrome clip threshold for paired reads was set to 30. Parameters for filtering were determined by examining summary statistics along the length of the sequences. Based on quality scores, 15bp were removed from the beginning of all reads. Reads were filtered for a minimum quality score of 30 using a sliding window size of 4 bp. Finally, reads shorter than 30 bp were excluded. FastQC was run again after these read-processing steps to verify that the low-quality regions and primers were removed from the reads.

	Condition	Library	Reads /Pairs	μ Length (bp)	μ Quality (σ)	% Aligned
H13 ***Arabidopsis thaliana***	control	SRR446039	10,480,497	35	37.39 (1.44)	97.01
		SRR446040	31,079,386	35	38.81 (2.14)	97.32
	infection	SRR446043	12,088,491	35	37.35 (1.46)	97.81
		SRR446044	24,659,952	35	39.02 (1.96)	97.33
R13 ***Solanum lycopersicum***	control	SRR948467	7,197,759	30	38.90 (2.25)	89.81
		SRR948468	10,386,980	30	38.94 (2.28)	88.59
		SRR948469	10,821,437	30	39.28 (1.80)	90.00
	infection	SRR948470	13,773,453	30	39.22 (1.90)	89.44
		SRR948471	11,324,684	30	38.95 (2.21)	89.59
		SRR948472	11,957,956	30	39.31 (1.77)	89.24

Table 4.2. Statistics of quality-controlled read libraries used to compare transcriptional responses to bacterial infection in *A. thaliana* and tomato. Rosli et al. 2013 and Howard et al. 2013 published RNA-seq data on responses to the bacterial pathogen *P. syringae* pv. tomato strain DC3000 in *S. lycopersicum* and *A. thaliana*, respectively. The control treatment in these datasets was mock infection with sterile buffer. Before using these data to find responses to a common pathogen across plant systems, we first compared the read counts, lengths, quality scores, and alignment rate.

Expression Matrix Construction

Libraries of trimmed sequencing reads were aligned to annotated reference genomes using HISAT2 (version 2.1.0). Reads were aligned to either the *A. thaliana* annotated TAIR10 reference genome assembly or the *S. lycopersicum* annotated reference genome. Both annotated genomes were obtained from the JGI Phytozome database (version 12.1). For both alignments, the minimum and maximum intron length were set to 20 bp and 500,000 bp, the penalty for ambiguous bases was 1, and the default 'fr' orientation was used. Alignment quality was assessed with QualiMap reports (version 2.2.1) and was over 90% for all the datasets (Table 4.2).

After alignment, reads were assembled into transcripts using StringTie (version 1.3.3b). The minimum isoform abundance was set to 0.1, the minimum transcript length was set to 200bp, the minimum read coverage was 2.5, and the minimum locus gap was 50 bp. A minimum of 1 spliced read with 10 bp on either side of a junction was required for an alignment to span it. The transcript counts for each gene were normalized to transcripts per million (TPM) and averaged across the biological replicates for each experimental condition to generate a matrix with one expression value for each condition. The log2 fold change (log2 FC) of these transcript abundances were compared between infection treatments and control samples.

Primary Metabolic Gene Annotation and Metabolic Network Construction

Using the PlantSEED resource in KBase, we annotated the genomes of the two species with enzymatic functions and reconstruct primary metabolic networks from plant genomes. The PlantSEED-curated *A. thaliana* genome is already annotated with the metabolic functions necessary to reconstruct a network of reactions in primary metabolism. To propagate these annotations to the *S. lycopersicum* genome, the app “Annotate Plant Enzymes with OrthoFinder” was run in KBase to find genes deemed orthologous to the curated *Arabidopsis* enzymes at a threshold of 55% amino acid sequence identity. The App assigns tomato genes the functional metabolic annotations of the *Arabidopsis* enzymes with which they are deemed orthologous. This consistent annotation across the genomes of plant species permits the same biochemical reactions to be evaluated for their responses to infection in *Arabidopsis* and crop transcriptomes.

A network of primary metabolic reactions was reconstructed based on the curated PlantSEED template of biochemical reactions. The PlantSEED project curates *Arabidopsis* enzymes involved in plant primary metabolism, the distinct mass-balanced biochemical reactions that they catalyze, and their localization within plant cells. These data are consolidated into a template of primary metabolism. Using the "Reconstruct Plant Metabolism" App, we reconstructed the network of plant primary metabolism comprised by over 1,000 of these curated enzymes and biochemical reactions spanning several organelles, including the plastid and mitochondrion.

Integrating Transcript Abundances with the Metabolic Network

We integrated the expression matrix with enzymes in the metabolic reconstructions. Using the "Integrate Abundances with Metabolism App," we assigned transcripts to their associated biochemical reactions in the network and obtained scaled expression scores at the reaction level. The Mahalanobis distance was used to quantify deviations between the percentile ranks of reaction scores based on transcript counts from control and microbe-exposed plants. A p-value cutoff of ≤ 0.05 was used to find the reactions influenced by each type of microbe exposure in the datasets.

RESULTS & DISCUSSION

The results in the tables below present three key quantities for each biochemical reaction in each plant system. First, the log2 fold change (log2 FC) in transcripts per million (TPM) for the infected samples compared to the control samples of a plant. This quantity indicates the direction

of the inferred change in gene expression as well as its magnitude. Second, the Mahalanobis quantifies deviations between a reaction's expression based on transcript counts from control and microbe-exposed plants. Larger distances indicate reactions farther from a 1:1 correspondence that are likely influenced by infections. Third, a p-value cutoff of ≤ 0.05 was used to select the candidate reactions influenced by each type of microbe exposure in the datasets.

Common responses of tomato to a pathogen (DC3000) and a partner (Pf55)

Several reactions in *S. lycopersicum* primary metabolism were associated with the presence of bacteria, whether pathogens or partners. When gene abundances were mapped to the metabolic network, reactions in amino acid metabolism and lipid biosynthesis appeared influenced by treatments with bacterial cultures of both pathogenic *P. syringae* DC3000 and nonpathogenic *P. fluorescens* strain 55 (Table 4.3). Several reactions in energy metabolism also contained enzymes that were differentially expressed in the presence of both *Pseudomonas* strains: glycolysis, the TCA cycle, and the synthesis of phylloquinone for photosynthesis. These results suggest that at least some of the responses of primary metabolism to the treatments were related to bacterial colonization broadly rather than specifically to antagonism by pathogens.

Common responses of model and crop plants to a pathogen

Model and crop species did overlap in the metabolic pathways that appeared responsive to pathogens. Susceptible *A. thaliana* and *S. lycopersicum* showed transcriptional changes in pathways involved in reducing and hydroxylating fatty acids when they were confronted with pathogenic *P. syringae* DC3000 (Table 4.4). The transcript counts associated with these reactions

increased in pathogen-exposed plants compared to control plants of both species. The fatty acids produced by these pathways provide are used in both lipids for the cell membrane and cutin for the cell wall. In addition, a pathway involved in lignin biosynthesis for the cell wall has increased transcript abundances in both the model and crop species upon infection with DC3000. Together, these results suggest common metabolic reconfigurations that enable reinforcement of plant cell barriers upon invasion by a pathogen.

A reaction in the pentose phosphate pathway was also influenced by gene expression changes in both *A. thaliana* and tomato, albeit in opposite directions. The same reaction was responsive to both DC3000 and *P. fluorescens* 55 in *S. lycopersicum*, indicating that this reaction may be influenced by the presence of bacteria regardless of the nature of their relationship to the plant host.

Model and crop plant species also appeared to share at least one metabolic response to fungal pathogens linked to S-adenosylmethionine (SAM) (Table 4.5). The increased expression for genes in this pathway could indicate increased methionine biosynthesis. However, processes involving folate could also be driving the observed changes in this reaction via effects on SAM. Reactions involving folate, SAM, and methionine are broadly involved in one-carbon metabolism, supplying methyl groups for a number of processes.

Common responses of a model plant to a bacterial and fungal pathogen

Within a plant species, common biochemical reactions were implicated in the responses to

pathogens from different domains. By comparing the responses of *A. thaliana* to the bacterial pathogen *P. syringae* and fungal pathogen *B. cinerea*, we found several biochemical reactions influenced by both microbes, including the same reaction in the pentose phosphate pathway that responded to DC3000 (Table 4.6). These reactions can be investigated further to determine if they are part of a general response to biotic stress or if they represent sites of antagonism in the evolution of responses to pathogens in different domains.

Common responses of a crop to three eukaryotic pathogens

Within a plant species, common reactions were implicated in the responses to diverse eukaryotic pathogens with different lifestyles. By comparing the responses of tomato to three eukaryotic pathogens, we find a common influence on thiamine biosynthesis (Table 4.7). Pathogens classified as necrotrophic (*B. cinerea*), biotrophic (*O. neolycopersici*), and hemibiotrophic (*P. infestans*) all decreased expression of a thiamine biosynthesis enzyme in plants leaves compared to control treatments. A key cofactor for core metabolic enzymes such as pyvurate dehydrogenase, thiamine has already been linked to plant defense responses and crop yields by controlled studies with exogenous treatments (Fitzpatrick et al. 2020, Ahn et al. 2005).

Table 4.3 Common responses of *Solanum lycopersicum* primary metabolism to *Pseudomonas* exposures

			DC3000			Pf55		
Compartments	**Reactions**	**Subsystems**	log2 FC	Mahalanobis	p value	log2 FC	Mahalanobis	p value
Plastid	rxn00069	Glutamine, glutamate, aspartate, asparagine metabolism	2.32	23.8	1.070E-006	1.51	11.69	0.000628
Plastid	rxn15435	Glycolysis and Gluconeogenesis, Sucrose metabolism	3.09	14.1	0.000173	3.35	15.24	0.0000946
Cytosol	rxn01182, rxn01427, rxn01571	Lipid biosynthesis	1.86	11.62	0.000652	1.02	4.16	0.04
Cytosol	rxn00330	TCA cycle	-1.25	11.44	0.00072	-1	6.14	0.01
Plastid	rxn19240, rxn19241, rxn19242	Branched-chain amino acid metabolism	-0.9	5.12	0.02	-0.99	4.86	0.03
Cytosol	rxn01975	Pentose phosphate pathway	0.94	4.48	0.03	1.74	11.78	0.000599
Plastid, Peroxisome	rxn24330	Phylloquinone biosynthesis	-0.92	3.75	0.05	-1.41	3.84	0.05

Table 4.4 Common responses of plant primary metabolism to *Pseudomonas syringae* DC3000

			Solanum lycopersicum			***Arabidopsis thaliana***		
Compartments	**Reactions**	**Subsystems**	log2 FC	Mahalanobis	p value	log2 FC	Mahalanobis	p value
Cytosol	rxn16289, rxn16291, rxn16300	Lipid biosynthesis	1.78	14.29	0.000157	1.98	19.08	0.0000125
Plastid, Cytosol	rxn01975	Pentose phosphate pathway	0.94	4.48	0.03	-1.28	7.86	0.01
Cytosol	rxn00495	Lignin biosynthesis, Phenylpropanoids general biosynthe	2.09	7.81	0.01	1.1	6.37	0.01

Table 4.5 Common responses of plant primary metabolism to *Botrytis cinerea* strains

			Solanum lycopersicum			***Arabidopsis thaliana***		
Compartments	**Reactions**	**Subsystems**	log2 FC	Mahalanobis	p value	log2 FC	Mahalanobis	p value
Cytosol	rxn00128	Methionine and cysteine metabolism	1.93	49.65	1.83E-012	1.48	4.11	0.04

Table 4.6 Common responses of *Arabidopsis thaliana* primary metabolism to bacterial and fungal pathogen exposures

			DC3000			***B. cinerea***		
Compartments	**Reactions**	**Subsystems**	log2 FC	Mahalanobis	p value	log2 FC	Mahalanobis	p value
Cytosol	rxn15069	Sucrose metabolism	2.43	18.36	0.0000183	0.41	5.01	0.03
Plastid	rxn00018, rxn02251	Calvin-Benson-Bassham, Photorespiration, Rubisco shun	1.55	12.24	0.000468	-0.43	7.14	0.01
Plastid	rxn01975	Pentose phosphate pathway	-1.28	7.86	0.01	0.94	16.49	0.0000489
Endoplasmic reticulum	rxn02554, rxn04426	Lipid biosynthesis, Phenylpropanoids general biosynthesi	-1.27	7.4	0.01	0.71	22.08	2.610E-006

Table 4.7 Common responses of *Solanum lycopersicum* primary metabolism to eukaryotic pathogen exposures

			B.cinerea			***O. neolycopersici***			***P. infestans***		
Compartments	**Reactions**	**Subsystems**	log2 FC	Mahalanobis	p value	log2 FC	Mahalanobis	p value	log2 FC	Mahalanobis	p value
Plastid	rxn27029, rxn27030	Thiamin biosynthesis	-1.28	5.37	0.02	-0.69	5.52	0.02	-1.76	3.7	0.05
Plastid	rxn01000, rxn19253	Tyrosine and phenylalanine metabolism	2.1	10.19	0.00141	ns	ns	ns	0.88	6.59	0.01
Cytosol,Plastid, Mitochondria	rxn00187	Glutamine, glutamate, aspartate, asparagine metabolism	-1.18	6.13	0.01	ns	ns	ns	1.17	6.08	0.01
Cytosol,Plastid, Mitochondria	rxn00191	Alanine, serine, glycine metabolism	1.21	6	0.01	ns	ns	ns	0.72	4.91	0.03
Peroxisome	rxn00248	Glycolysis and Gluconeogenesis, TCA cycle	-1.27	5.52	0.02	ns	ns	ns	-1.14	3.83	0.05
Plastid, Peroxisome, Mitochondria	rxn00272, rxn00275, rxn00424	Alanine, serine, glycine metabolism, Photorespiration (oxidative C2 cycle)	-1.18	4.51	0.03	ns	ns	ns	-1.03	4.15	0.04
Cytosol, Mitochondria	rxn01213, rxn01466	Geranylgeranyldiphosphate biosynthesis	2.56	7.51	0.01	0.24	4.76	0.03	ns	ns	ns
Cytosol	rxn01454	Isoprenoid biosynthesis, mevalonate branch	2.35	4.86	0.03	-0.6	4.76	0.03	ns	ns	ns
Plastid	rxn14094, rxn14142, rxn14261, rxn24358, rxn24360, rxn24362	Phylloquinone biosynthesis	ns	ns	ns	-0.77	15.12	0.000101	-1.59	4.02	0.04
Cytosol	rxn03970, rxn03971, rxn05143	Phenylpropanoids general biosynthesis	ns	ns	ns	0.72	11.76	0.000605	1.22	8.55	0.00346
Cytosol	rxn00412	Pyrimidine de novo biosynthesis	ns	ns	ns	0.85	6.63	0.01	1.35	4.42	0.04
Cytosol, Plastid	rxn15494	Calvin-Benson-Bassham cycle, Glycolysis and Gluconeo Sucrose metabolism	ns	ns	ns	0.36	4.51	0.03	-1.09	6.26	0.01
Cytosol, Nucleus	rxn00097	Purine de novo biosynthesis	ns	ns	ns	0.81	4.93	0.03	-1.94	11.63	0.00065
Mitochondria	rxn08094	TCA cycle	ns	ns	ns	0.44	4.42	0.04	0.74	4.98	0.03
Cytosol, Plastid	rxn15493	Glycolysis and Gluconeogenesis, Sucrose metabolism	ns	ns	ns	0.24	3.99	0.05	0.67	4.57	0.03

ID	Reaction	Subsystems	Compartment
rxn16289	NADPH + O_2 + H^+ + Palmitate => H_2O + NADP + 16-Hydroxypalmitate	Lipid biosynthesis	Cytosol
rxn16291	NADPH + O_2 + H^+ + cis-9,10-Epoxystearic acid => H_2O + NADP + 9,10-Epoxy-18-hydroxystearate	Lipid biosynthesis	Cytosol
rxn16300	NADPH + O_2 + H^+ + 9,10-Dihydroxystearic acid => H_2O + NADP + 9,10,18- Trihydroxystearate	Lipid biosynthesis	Cytosol
rxn01975	NADP + beta-D-Glucose 6-phosphate <=> NADPH + H^+ + 6-phospho-D-glucono-1-5-lactone	Pentose phosphate pathway	Plastid, Cytosol
rxn00495	L-Phenylalanine => NH_3 + (E)-Cinnamate	Lignin biosynthesis, Phenylpropanoids general biosynthesis	Cytosol
rxn00128	S-Adenosyl-L-methionine <=> H^+ + 5-Methylthioadenosine + 1-Aminocyclopropane-1-carboxylate	Methionine and cysteine metabolism	Cytosol

Table 4.8 Reactions predicted to respond to infections in both model and crop plant hosts.

CONCLUSIONS

Integration of expression data with a metabolic network thus yielded a set of candidate reactions that can focus study of plant-microbe relationships on those metabolic responses that translate between model and crop plant hosts, summarized in Table 3.8. In particular, the case studies considered here suggest that reinforcement of the cell membrane and wall through lipid and lignin biosynthesis are important responses of both model and crop systems to the bacterial pathogen *P. syringae* DC3000. Results in tomato also suggest that further study of one-carbon metabolism and thiamine biosynthesis may be useful in elucidating the effects of eukaryotic microbes with diverse modes of life on plant hosts.

CHAPTER 5

Transgenerational responses of *Arabidopsis thaliana* primary metabolism to infection with *Pseudomonas* pathogens

CO-AUTHORS & CONTRIBUTIONS

Jacob Herman and Joy Bergelson designed the infection history scheme that produced the plant lineages used in this work. I extracted RNA from these plants and analyzed the transcript data. Together with Sam Seaver, I integrated this transcript data with a model of plant primary metabolism. My work was supported by the University of Chicago Biological Sciences Division and NIH T32 GM 7197.

ABSTRACT

Research on plant immune responses to bacterial pathogens has been complicated by two revelations of the high-throughput sequencing era of plant biology. Firstly, microbiome sequencing has revealed that plants are frequently infected by multiple pathogens simultaneously. Controlled lab studies show that the defense systems activated by different pathogens may even antagonize each other via hormonal crosstalk. Secondly, epigenetic marks on plant genomes are altered by pathogen infection and these changes can be transmitted to offspring, meaning that offspring immune responses can depend upon parental infection histories. One element of the plant's response to pathogens is the reconfiguration of primary

metabolism, however it is unclear if the effects of biotic stress on metabolism in particular are transmitted across generations. Although lab studies typically study infections by single pathogens within a plant's lifetime, an understanding of how infection-induced phenotypes affect subsequent generations and how these phenotypes differ between single and co-infections is essential for translating knowledge of plant immunity to field settings.

In this chapter, we processed RNA-seq data in order to find changes in gene expression induced by single or co-infections of *Arabidopsis thaliana* Col-0 with two bacterial pathogens isolated from natural *A. thaliana* populations in the midwest United States: *Pseudomonas syringae* strain NP29.1A and *Pseudomonas viridiflava* strain RMX3.1b. We characterized the responses of primary metabolism to infection at the system level by using the transcriptome data to constrain flux balance analysis on models of plant primary metabolism (tFBA). We then compared the metabolic changes predicted in response to different infection histories both within a generation and across generations.

INTRODUCTION

Co-infections with *Pseudomonas* pathogens

Interactions between bacterial pathogens and plant leaves provide a system in which to study the evolution of virulence and immunity. Furthermore, they provide an opportunity to connect findings about host-pathogen evolution to global carbon cycles since pathogens destroy leaf tissues and limit their photosynthetic production and carbon sequestration. Strains of the genus

Pseudomonas are among the best-studied microbes with pathogenic potential which colonize plant leaves (Xin et al. 2018). These gram-negative γ-Proteobacteria can interact with a wide-range of plant hosts, either living as foliar epiphytes or invading the apoplast (Hirano and Upper 2000). In pathogenic interactions, the presence of virulence-associated *Pseudomonas* effector proteins can initiate a hypersensitive response in host leaves and lead to tissue necrosis (Katagiri et al. 2002).

Pseudomonas syringae and *Pseudomonas viridiflava* are natural pathogens of *Arabidopsis thaliana* in the midwest (Jakob et al. 2002). The strains *P. syringae* NP29.1a and *P. viridiflava* RMX3.1b were isolated from field-collected *A. thaliana* leaves and show evidence of distinct life histories: *P. syringae* NP29.1a is thought to interact with the plant as a biotroph while *P. viridiflava* RMX3.1b is thought to live as a necrotroph. Biotrophs exploit host cells without killing them whereas necrotrophs cause tissue death. Perhaps because of their different modes of life and the accompanying consequences for host tissue, these pathogens induce distinct signaling responses in *A. thaliana* (Jakob et al. 2007). Specifically, *P. viridiflava* RMX3.1b is distinguished by its induction of jasmonic acid (JA) signaling pathways. JA pathways are thought to operate in immune responses to necrotrophic pathogens, whereas salicylic acid (SA) pathways are critical for responses to biotrophic *P. syringae*. SA and JA signaling are required for the expression of distinct plant peptides with antimicrobial activity, such as SA-dependent pathogenesis-related protein 1 (PR-1) and JA-dependent plant defensin 1.2a (PDF1.2a) (Sels et al. 2008). Ultimately, the induction of different pathways can have consequences for subsequent plant-microbe interactions. Demonstrating this, diversity differs between bacterial assemblages

surveyed in plants with mutations in the SA and JA pathways (Kniskern et al. 2007).

Because *P. syringae* and *P. viridiflava* induce distinct responses in the host plant, there is a potential for these responses to conflict. This conflict becomes relevant in field conditions where multiple strains of pathogens colonize the same populations of host plants (Bartoli et al. 2018, Karasov et al. 2018). For example, different metabolic reconfigurations may be advantageous to the host plant in response to the presence of bacteria activating SA and JA pathways. Evasion strategies, leading to programmed death of compromised tissue, favor necrotrophic growth while endurance strategies, reallocating resources to challenged cells, can sustain biotrophs (Seifi et al. 2013). Defense responses may differ when they are induced by a pathogen in isolation as opposed to a community context in which microbes compete, cooperate, and cheat during co-infections of a host.

To assess infection-induced phenotypes, we assayed gene expression in plants with histories of infection by two different bacterial pathogens isolated from natural *A. thaliana* populations in Michigan: *P. syringae* NP29.1A and *P. viridiflava* RMX3.1B (Figure 5.1A). We also examined gene expression in response to a co-infection with both *P. syringae* and *P. viridiflava* to determine if the responses in transcription varied between single and co-infection conditions.

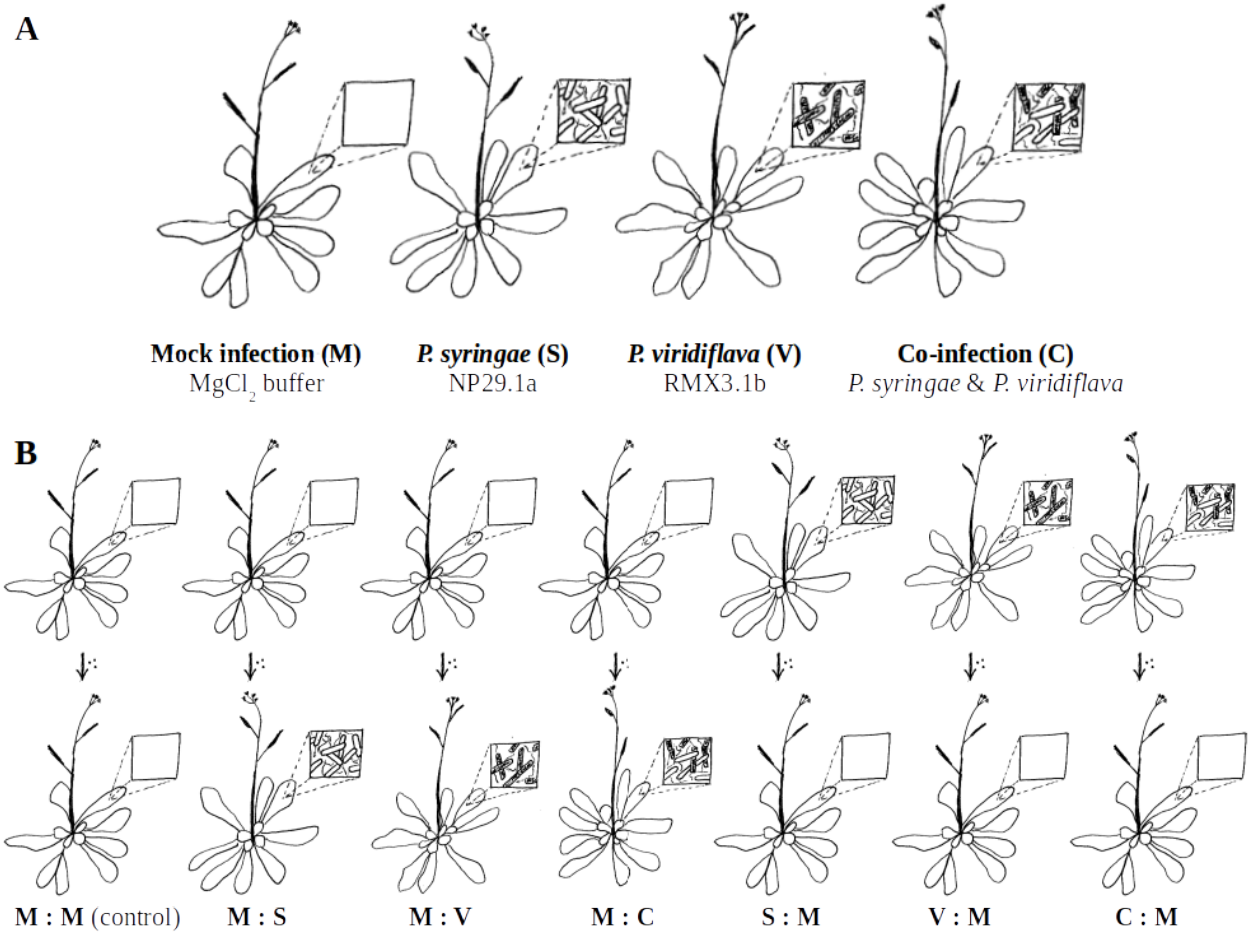

Figure 5.1 Experimental design for studying transgenerational and co-infection responses. (A) An outline of the pathogen infection and control treatments applied to *A. thaliana* leaves by spray inoculation. (B) An outline of the infection histories established in *A. thaliana* lineages for this work.

Transgenerational responses to infection

Plant phenotypes are shaped by prior exposures to biotic stress both within and across generations (Holeski et al. 2012). Within generations, plant immune responses are primed by prior infections, thereby affecting the outcome of subsequent infections (Mauch-Mani et al. 2017). In addition to within-generation priming, plants exhibit infection-induced phenotypic changes across generations. Without any pathogen exposure, a plant can exhibit physiological

differences depending on whether its parent experienced biotic stress (Slaughter et al. 2012, Rasmann et al. 2012). As a result, the progeny of infected parents may respond differently to infections than those of unexposed parents. These phenotypic responses to biotic stress across generations can be caused by maternal effects like seed provisioning (Ganguly et al. 2017) or epigenetic mechanisms like DNA methylation (Luna et al. 2012). Transgenerational defense induction appears to be widespread in plant systems, having been documented both in the model species *Arabidopsis thaliana* and in crops like tomato (*Solanum lycopersicum*) (Holeski et al. 2012).

These findings raise the practical question of whether immune priming and transgenerational defense induction can be manipulated to protect crops from pathogens (Ramírez-Carrasco et al. 2017). At the same time, they raise the ecological question of whether defense induction shapes microbiome assembly and growth-defense tradeoffs in subsequent host generations (Karasov et al. 2017). Most plant-microbe studies sow and harvest plants within a single generation, limiting our understanding of how plants interact with microbes in natural settings when successive generations are exposed to the same biotic environment. A prerequisite to answering both the applied and ecological question is to develop and characterize a system in which the genetically identical offspring of naïve and infected plants are assayed in both control and infection conditions. Such a system allows the identification of phenotypic changes within and across generations that are associated with infection history rather than genotype.

To ask how pathogen exposures in a previous generation shape the habitat for microbes within

plant leaves, we performed comparisons between the uninfected offspring of uninfected parents (control), the infected offspring of uninfected parents (within-generation infection), and the uninfected offspring of infected parents (previous generation infection) (Figure 5.1B).

The metabolic interface of host-pathogen interactions

Plant tissues create distinct habitats that filter the pool of bacterial colonists in the surrounding soil, shaping the composition of microbial communities. In these habitats, the nutrients available for bacteria are determined by the host's photosynthesis and carbon assimilation. These metabolic systems are reconfigured at the transcription level in response to biotic stress, thereby changing the environment for bacterial endophytes (Ward et al. 2010, Schwachtje et al. 2018). This raises the question of whether the metabolic reconfiguration induced by infection might affect plant tissues in subsequent generations.

For immune responses mediated by individual peptides or R genes, transcript abundances for single loci may provide a good indicator of changes in resistance phenotypes. Yet in metabolic pathways, increased expression at a discrete locus might be insufficient to increase the production of a metabolite if the required enzyme has multiple subunits. Even when one gene codes one enzyme, the the effects of increased expression at a discrete locus could be dampened by a compensating pathway.

To study infection-induced metabolic reconfiguration at the system level, we generated a metabolic model of *A. thaliana* from its annotated genome (Henry et al. 2010, Seaver et al. 2014,

Seaver et al. 2018). We then used our genomic expression datasets to constrain flux balance analysis on the primary metabolic model. Flux balance analysis (FBA) uses linear programming to solve steady-state equations generated from a stoichiometric matrix of compounds and reactions in the model (Orth et al. 2010, Yuan et al. 2016). FBA can predict relative flux through each reaction in the model and cumulative fluxes through reactions grouped into metabolic subsystems or classes. To limit the activity of biochemical reactions in FBA, enzyme-coding genes can be subjected to an expression threshold determined by transcript counts (Beckers et al. 2016). These transcriptome-constrained FBAs (tFBAs) represent the predicted responses of primary metabolism to the environments from which the transcriptome profiles are collected, in our case infected and uninfected leaf tissues.

When tFBAs constrained with data from stressed and control plants are compared, metabolic flux differences related to invasion can be identified. Changes in the flux predicted under constraints from infected plants, as opposed to control plants, indicate candidate reactions shaping the response of host metabolism to prior infections. Changes in the flux predicted under constraints from plants with a history of parental infection, as opposed to control plants, indicate candidate reactions that maintain metabolic changes induced by infection in subsequent generations.

MATERIALS & METHODS

Experimental Design

In order investigate the effects of single- vs. co-infection on transgenerational priming, we generated plant lineages with different histories of bacterial infection over two generations (Fig.1). We grew two *A. thaliana* accessions in uniform, favorable greenhouse conditions in order to control for effects of parental environment and seed age. Offspring of these plants were the ‘parental’ plants in this growth chamber study.

Parent plants were spray-inoculated with the natural *A. thaliana* pathogens *Pseudomonas syringae* (Michigan strain NP29.1a) and *P. viridiflava* (Michigan strain RMX3.1b), alone and in combination (co-infection). These strains naturally infect A. thaliana and the species they represent have been reported to induce distinct defensive pathways that antagonize each other through hormonal crosstalk (*P. syringae* induces SA-pathway defenses; *P. viridiflava* induces JA-pathway defenses). Mock inoculations were performed as controls, consisting of spray application of sterile buffer.

Offspring of inoculated plants were subjected to the same treatments as their parents, yielding all permutations of mock inoculation and bacterial infection treatments across two generations for these accessions. Leaves from two replicate offspring plants in each condition were collected for RNA-seq to obtain transcription profiles for each condition and compare these profiles across conditions.

This framework enables discovery of phenotypic changes that are induced by (i) two pathogens with different life histories and by (ii) single- vs. co-infection. Furthermore, it will enable the discovery of (iii) transgenerational induced changes, or phenotypic changes in infected plants that are maintained in their uninfected offspring.

Plant growth conditions

All plant genotypes were grown in uniform, favorable greenhouse conditions in order to control for effects of parental environment and seed age. Seed stocks from these plants were used to found the experimental lineages. The first (Parental) and second (Offspring) generations of plants in each lineage were grown in the same conditions. Plants were grown in 36-cell plastic flats on peat and perlite substrate. A growth chamber provided 8h of artificial light and plants were watered daily. The growth chamber's temperature was 20ºC.

Pathogen infections

Whole plants were spray-inoculated with ~$2x10^8$ cfu/ml (or sterile 10mM MgSO4 buffer) at 4.5 weeks post germination. Plants were maintained at 100% RH for 4 days post inoculation. Disease symptoms were scored at 6 days post infection.

Sample collection and RNA extraction

Two leaves were harvested from each plant in the offspring generation and immediately frozen. The biological cabinet and all tools used in the extraction were scrubbed with RN-Zap spray and

rinsed with RNA-grade water. All disposable supplies used in the extraction were RNAse-free. Reagents were prepared as instructed in the Qiagen RNeasy Plant Mini Kit. Sample tubes were removed from the -80C freezer and placed in liquid nitrogen. One at a time, sample tubes were removed from the liquid nitrogen and weighed to ensure the leaf tissue did not exceed 100mg. An RNAse-free pellet pestle was used to grind each sample for ~30 seconds in its tube. Samples were returned to liquid nitrogen until all samples were ground. At this point, samples were suspended in buffer RLT from the RNeasy kit, vortexed for 20s, and incubated for three minutes in a 54C heat block to lyse the cells. Even though the lysis was sufficient to extract enough RNA for sequencing, it was not complete. The debris remaining behind after lysis were stored at -80C. The RNeasy kit was used to homogenize samples, bind the RNA to a column (150ul EtOH added to ~250-300ul lysate), wash the column (30s centrifuge runs, new collection tubes were used after the first wash), and elute RNA (50ul RNA-grade water). The RNA yield was between 0.85ug and 80ug total (measured by Agilent 2100). The 260/280 was between 1.9 and 2.16 and the 260/230 was between 0.5 and 2.2 in water (measured by NanoDrop).

Transcriptome sequencing

The following library preparation and sequencing were performed by Novogene. Total RNA samples were enriched for mRNA with oligo(dT) beads. The mRNA was randomly fragmented and used as template to synthesize cDNA using random hexamer primer, a custom second-strand synthesis buffer (Illumina), dNTPs, RNase H and DNA polymerase I. The library was subjected to terminal repair, A ligation and sequencing adaptor ligation. The cDNA library was then size selected for 250~300 bp inserts, PCR amplified, and tested for insert size and concentration by

Agilent 2100 and qPCR. After quality control, the libraries were pooled based on concentration and sequenced on an Illumina platform to produce paired end 150bp reads.

Analysis in KBase

The sections below describe an analysis workflow in the Department of Energy's Systems Biology Knowledgebase, or KBase (Arkin et al 2018). In KBase, users can create a permanent electronic analysis notebook in the form of a "Narrative." Narratives are composed of Markdown cells with information about the project, "Apps" which run publicly available software and save a record of the parameters and terminal output, and code cells supporting custom commands in Python. The Narrative for this project can be found at the following link: https://narrative.kbase.us/narrative/73121

Quality Control

Quality control was performed with Apps running FASTQC to assess quality scores for the mRNA sequencing reads and Trimmomatic to remove Illumina primers, low quality reads, and short reads. The average initial quality score for reads was 36 in each library, corresponding to 99.97% base calling accuracy. Sequence lengths were distributed tightly around the expected 150bp and per sequence GC content agreed with the expected distribution. There was a low level of Illumina primer detected in reads and per-base sequence content was not stable in the first ~15bp. The first 15bp was removed from each read to eliminate the primers and reads were filtered by quality scores with a threshold score of 30. Roughly ~85% of paired reads survived trimming. Unpaired forward and reverse reads were discarded. Trimmed sample libraries each

had 15-25 million paired reads. Sequence lengths were distributed tightly around 125bp after trimming and the Illumina primer was eliminated in the post-trimming QC report.

Expression Matrix Generation

Reads were aligned to the *A. thaliana* TAIR10 reference genome assembly with the HISAT2 App. Alignment quality across samples was assessed with QualiMap reports produced by the HISAT2 App. In StringTie, aligned reads were assembled into transcripts with a minimum length of 200bp based on known transcript isoforms. StringTie identified 27,655 transcripts and estimated their abundances based on read counts.

Due to the small number of biological replicates, DESeq2 was used to assess differential expression at the gene level by modeling the relationship between expression variation and expression levels. Results from the DESeq2 App were downloaded from KBase and imported to R. The result tables were filtered to find expression changes between conditions that were statistically significant based on a q-value, or the p-value adjusted by using the Benjamini-Hochberg algorithm to find an upper bound for the expected false discovery rate ($q < 0.05$). Results were also filtered to find genes with fold change of at least 0.1. For genes with significant expression changes in all infection treatments, their fold changes realtive to the control lineage (Mock : Mock) were recorded in a heatmap.

Flux Balance Analysis

In order to generate a system-level phenotype for primary metabolism, we reconstructed a

network of primary metabolism from the *A. thaliana* TAIR10 genome by running the Reconstruct Plant Metabolism App. The PlantSEED framework was used to assign genes to enzyme functions based on a high-quality, manually curated set of annotations from databases and the literature. Those enzymes are assigned to reactions which convert between metabolites in the network. Reactions are further organized into subsystems and assigned to subcellular compartments. An expression score was computed for each reaction using the genes associated with it. Reactions that differed between control and infection conditions were detected by using the Mahalanobis distance to compute the variance of the expression scores around their geometric mean.

Using the metabolic network reconstructed from the genome, we performed flux balance analysis (FBA) to simulate the flow of metabolites into the system from a defined media and through the network. The defined media provided H_2O, O_2, HPO_4^{2-}, CO_2, NH_3, SO_4^{2-}, H^+ ,Cl^-, K^+, Mg, Fe^{2+}, as well as photons for reactions in the plant metabolic model. This media provided one key constraint on the simulations by limiting the nitrogen input. The “Run Flux Balance Analysis” App was used to find the reaction fluxes ($mmol \cdot h^{-1}$) that maximized a biomass reaction. The biomass reaction involved 148 building blocks-- including amino acids, nucleotides, cofactors, and sugars-- necessary to sustain the plant cell. This biomass reaction, together with upper and lower limits on reaction flux, provided another set of constraints on the simulations.

Transcription constraints on FBA

A final constraint on the simulations was added by integrating transcript abundance data. Transcript abundances from the control plants were used to constrain the simulation of flux (tFBA) by blocking reactions involving enzymes whose expression scores fell below a scaled expression score threshold (0.5) with given uncertainty (0.1). The “Run Flux Balance Analysis” App computes a normalized reaction expression score, accounting for the number of enzymatic isoforms and the composition of enzyme complexes associated with each reaction. An expression score threshold of 0.5 (corresponds with median of normalized expression values) and uncertainty of 0.1 were used to determine which reactions were blocked. Multiple simulations were run with constraints set by expression scores based on both control and microbe-exposed plants. Flux results from the control conditions were subtracted from others in order to find the reactions carrying different flux in response to microbial exposures. The absolute values of these flux changes were used to identify the reactions responsive to infection-specific constraints in the simulations. Custom code cells were used in the Narrative to access App outputs such as flux values and create figures.

RESULTS & DISCUSSION

Pathogenesis-related genes showed distinct responses to infection by individual pathogen strains and co-infection. Expression changes in plants with infection histories indicated that *P. syringae* NP29.1a and *P. viridiflava* RMX.31b both induced defense responses in *A. thaliana* and that some of these genes were perturbed across generations even when offspring remained

uninfected. Table 5.1 shows the number of genes with higher and lower expression in each infection condition than in the Mock : Mock control lineage. We selected the genes that were differentially expressed in response to all pathogen treatments in infected offspring as well as in the uninfected offspring of infected parents.

Expression change relative to Mock : Mock	P. syringae : Mock	P. viridiflava: Mock	Co-infection : Mock	Mock : P. syringae	Mock : P. viridiflava	Mock : Co-infection
increase	32	46	98	138	10	86
decrease	15	149	67	89	7	231

Table 5.1. The counts of genes in each infected lineage with increased or decreased expression compared to the control lineage based on mRNA transcripts per million. Column headers are infection histories and row names indicate the direction of the expression change between the infection lineage and the control lineage.

Four genes met these criteria (Figure 5.2, Table 5.2). Of these, two were pathogenesis-related proteins expressed in the rosette leaves throughout most of plant development (Mergner et al. 2020), At5g44430 (PDF1.2c) and At2g26010 (PDF1.3). Expression of PDF1.2a, a member of the same protein family, reportedly responds to biotic stressors*;* mutations of PDF1.2a alter sensitivity to the fungal pathogen *B. cinerea* (Sels et al. 2008).

Table 5.2. Names and functional categories for the genes differentially expressed in response to infection compared to the control lineage.

Gene	**Names**	**Predicted Function**
AT2G26010	PDF1.3, PLANT DEFENSIN 1.3	Defense
AT3G02400	FHA3, FORKHEADASSOCIATED DOMAIN 2	Transcription factor
AT3G57140	SDP1-LIKE, SUGAR-DEPENDENT 1-LIKE	Triacylglycerol lipase
AT5G44430	PDF1.2C, PLANT DEFENSIN 1.2C	Defense

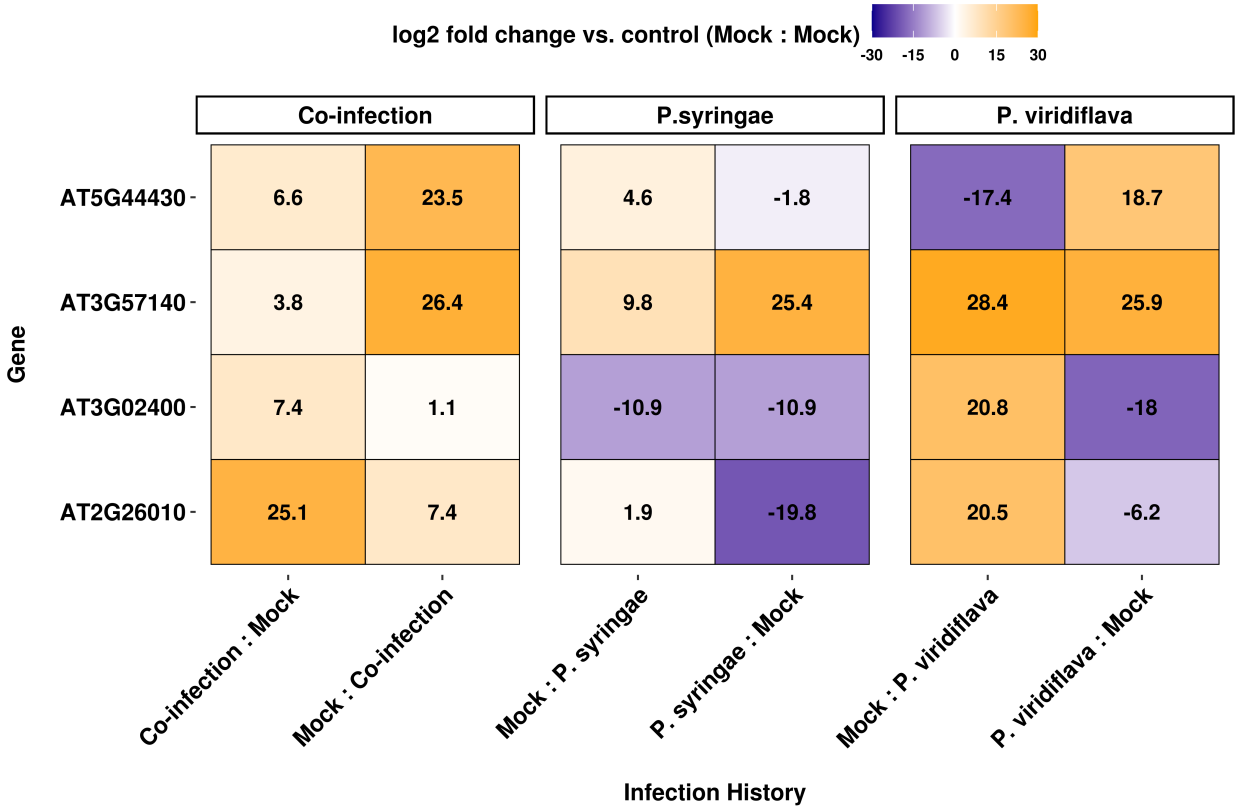

Figure 5.2 Gene expression changes in response to infection. For each plant gene (row), the expression change (log_2 fold change of transcripts per million) in each infection treatment (column) compared to the control is represented by a color indicating the direction and magnitude of the change. Positive changes in expression are orange and negative changes in expression are blue.

In the case of PDF1.2, the plant's immune responses to NP29.1a ad RMX3.1b differed from each other as well as from responses to co-infection. The response of PDF1.2c to *P. viridiflava* differed from that of the closely related PDF1.2a. The latter gene was increasingly transcribed in the days following infection according to previous qPCR data (Jakob et al. 2007), however log-transformed transcript counts per million (TPM) of the former gene declined 17.4 fold after *P.*

viridiflava infection (Figure 5.2). Yet in the naïve offspring of *P. viridiflava*-infected parents, PDF1.2c showed a 18.7 fold increase in TPM. The defensin showed a comparatively modest 4.6 fold increase following *P. syringae* inoculation and a 1.8 fold decrease in the uninfected offspring of *P. syringae*-exposed plants. After co-infection, mRNA counts for PDF1.2c increased 6.6 fold. In the uninfected offspring of co-infected plants, transcript counts increased 23.5 fold. Together, these results indicated distinct transcriptional responses to different pathogens at the PDF1.2c locus.

The pattern of plant defensin expression following co-infection was similar to that following infection with *P. viridiflava* in isolation (Figure 5.2), suggesting that *P. viridiflava* drove the immune response at the PDF1.3 locus during co-infection. When *P. viridiflava* was introduced by itself or during co-infection, PDF1.3 log-transformed TPM increased by 20.5 to 25.1 fold. In contrast, the generation following *P. syringae* infection saw a 19.8 fold decline in transcript abundance.

Triacylglycerol lipase expression increased in response to infections both within and across generations. The remaining two genes differentially expressed in response to all infection histories encoded a transcription factor and a lipase. While no additional information was available on the former, the latter, At3g57140 (SDP1-like), is a lipase which hydrolyzes the storage lipid triacylglycerol . Triacylglycerol hydrolysis (EC 3.1.1.3) is an early step in the path to jasmonic acid synthesis; thus, this enzyme is tied to the immune signaling pathways that regulate the defensins discussed above. The lipase is expressed in developing seeds, where it

breaks down oil body reserves to fuel germination (Eastmond et al. 2006). Triacylglycerol lipases are also expressed in rosette leaves, where oil bodies produce phytoalexins in response to fungal infections (Shimada et al. 2014).

Triacylglycerol lipase transcription appeared to be higher after infection both within and across generations, regardless of the pathogen treatment applied (Figure 5.2). The strongest response within a generation was to *P. viridiflava*, which induced an increase of 28.4 fold in the enzyme's TPM. Contrasting the variable responses of pathogenesis-related genes, this result suggests a more uniform metabolic response to pathogen infection in *A. thaliana*. In addition, it suggests that metabolic responses can be maintained across at least one generation after a pathogen encounter.

Transcription-constrained FBA predicted candidate reactions responsive to infections. To examine primary metabolism at the system level, rather than at the level of discrete genes, we next performed tFBA to simulate flux in a metabolic network with constraints set by the transcript abundances in each lineage. The tFBA reaction fluxes based on control plant data served as a baseline which we subtracted from reaction fluxes obtained by tFBA with expression scores from infected plants. Thus, our results predict changes in reaction fluxes in the simulation (Δ mmol $\cdot$ h^{-1}) that result from different infection histories. Of the 1066 reactions in the *A. thaliana* metabolic model, 460 were capable of carrying flux under the constraints of tFBA with the data from control plants. To select the variable reactions most influential in the simulations, we selected reactions with an absolute value for flux above 0.5. This yielded 37-40 candidate

reactions for each tFBA result. Most of these candidate reactions did not have large changes in flux given any of the six infection histories (Figure 5.3, Table 5.3).

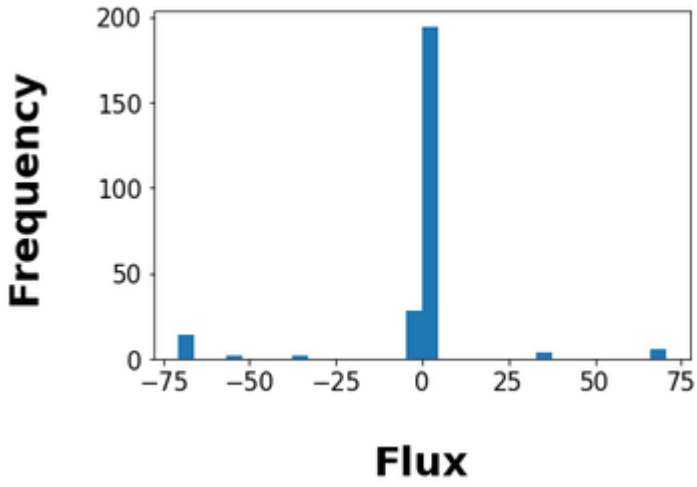

Figure 5.3 Most candidate reactions did not have large changes in flux in response to any of the six infection histories. The histogram shows the total number of reactions (y-axis) across all simulations with a given change in flux (Δ mmol $\cdot$ h^{-1} on the x-axis). The latter quantity is the difference in predicted flux values between simulations constrained with transcription data from infected versus control plant lineages. The histogram includes 271 flux values, 37-40 for each tFBA result, from reactions with scaled expression scores and absolute flux changes over a threshold of 0.5.

Table 5.3 Reactions with changes in simulated flux for the six infection histories compared to control conditions.

	P. syringae : Mock	**P. viridiflava: Mock**	**Co-infection : Mock**	**Mock : P. syringae**	**Mock : P. viridiflava**	**Mock : Co-infection**
rxn00069	70.76	70.48	70.48	70.48	70.48	70.48
rxn00018	-53.38	-53.38	0	0	0	0
rxn00248	2.5	-70.48	2.5	1.74	2.5	-70.48
rxn00747	-36.04	0	0	0	0	-36.04
rxn01116	36.04	36.04	0	36.04	36.04	0
rxn17196	-70.76	-70.48	-70.48	-70.48	-70.49	-70.48
rxn19828	-70.76	-70.49	-70.49	-70.49	-70.48	-70.49

Photosynthesis was part of a transgenerational infection response to multiple pathogens.

Previous studies on transgenerational defense induction showed common responses to multiple pathogen or herbivore species, suggesting that the responses were general to biotic stress and not specific to pathogens (Holeski et al. 2012). Our results partially support this idea of a general response to biotic stress because a few of the flux changes observed with transcription data from *P. syringae* and *P. viridiflava* infections overlaped. The activities of TPNH-ferredoxin reductase, ferredoxin-dependent glutamate synthase, and L-glutamate : NAD+ oxidoreductase were responsive to the tFBA constraints from infected plants (Figure 5.4). These reactions are all involved in photosynthesis and are linked by ferredoxins and glutamine (Table 5.4). Notably, the last of these reactions also showed expression changes in *S. lycopersicum* infected with *P. syringae* DC3000 in Chapter 4, further suggesting the existence of conserved responses of primary metabolism to biotic stress.

Table 5.4 Reactions with predicted flux differences under constraints from infected plants versus control plants.

Enzyme	Reaction	Subsystems	Compartment
TPNH-ferredoxin reductase (rxn17196)	NADP + H^+ + **2** Reduced-ferredoxins => NADPH + **2** Oxidized-ferredoxins	Photosystem I, Photosystem II, Cytochrome b6-f complex	Cytosol
ferredoxin-dependent glutamate synthase (rxn19828)	**2** L-Glutamate + **2** Oxidized-ferredoxins <=> 2-Oxoglutarate + L-Glutamine + **2** H^+ + **2** Reduced-ferredoxins	Glutamine, glutamate, aspartate, asparagine metabolism	Plastid
L-glutamate:NAD+ oxidoreductase (transaminating) (rxn00069)	NAD + **2** L-Glutamate <=> NADH + 2-Oxoglutarate + L-Glutamine + H^+	Glutamine, glutamate, aspartate, asparagine metabolism	Plastid
D-Ribulose-5-phosphate 3-epimerase (rxn01116)	D-Ribulose5-phosphate <=> D-Xylulose5-phosphate	Calvin-Benson-Bassham cycle in plants, Rubisco shunt in plants, Photorespiration (oxidative C2 cycle)	Plastid
3-phospho-D-glycerate carboxy-lyase (dimerizing; D-ribulose-1,5-bisphosphate-forming) (rxn00018)	H_2O + CO_2 + D-Ribulose 1,5-bisphosphate => **2** H^+ + **2** 3-Phosphoglycerate	Calvin-Benson-Bassham cycle in plants, Rubisco shunt in plants, Photorespiration (oxidative C2 cycle)	Plastid

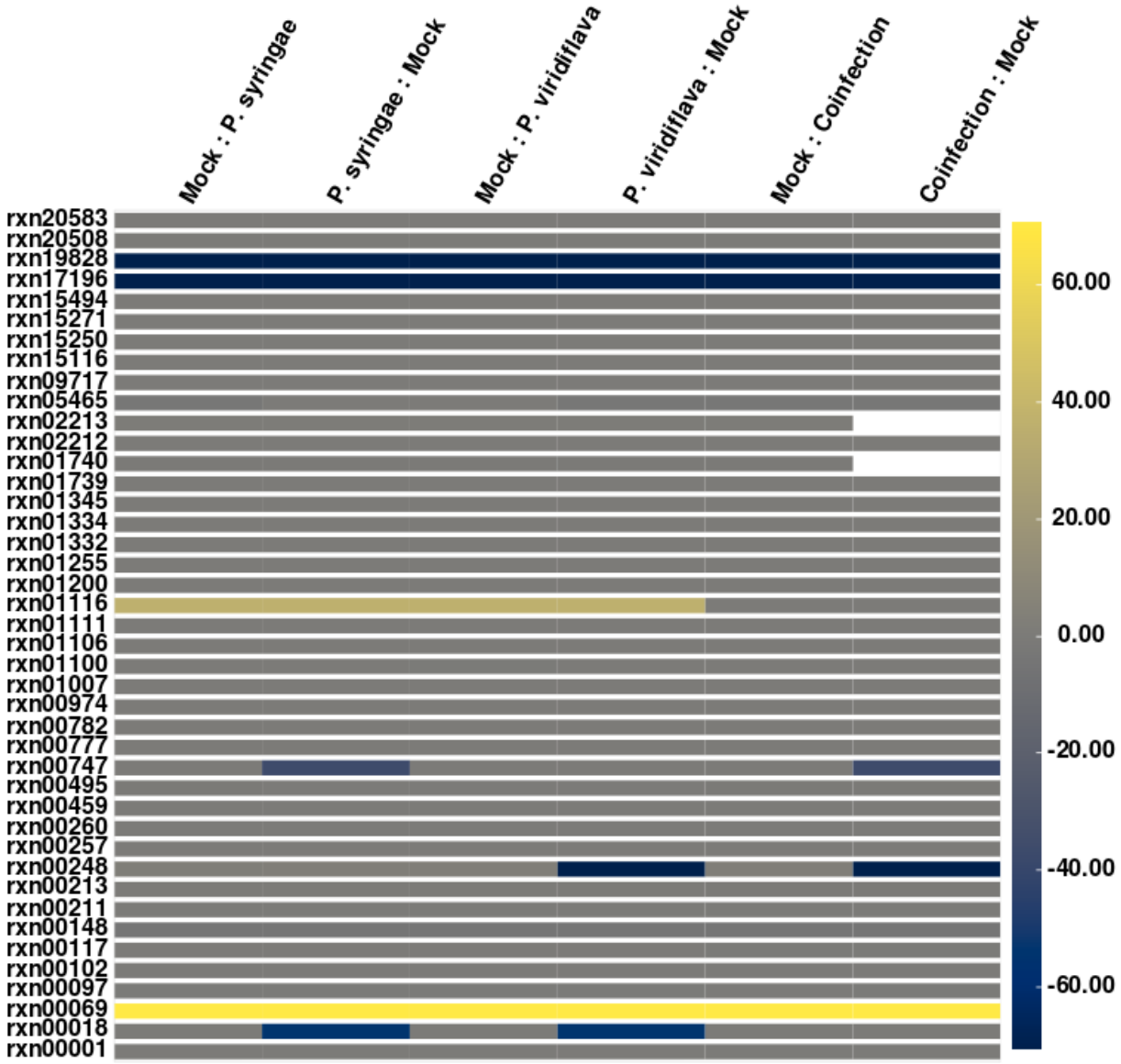

Figure 5.4 Flux changes predicted in response to infection. For each reaction (row) in primary metabolism, the change in flux (Δ mmol $\cdot$ h^{-1}) with each infection treatments (column) compared to the control is represented by a color indicating the direction and magnitude of the change. Positive changes in flux are yellow and negative changes in flux are blue. Since the magnitude of the change is indicated by the intensity of the color, changes closer to zero have little to no color.

Carbon assimilation responses distinguished single pathogen infections and co-infections. In addition to the trio of photosynthesis-associated reactions, two reactions involved in carbon assimilation were predicted to respond to infections by *P. syringae* and *P. viridiflava* (Table 5.4). Flux increased for D-Ribulose-5-phosphate 3-epimerase under constraints from both individual pathogens within and between generations, while for 3-phospho-D-glycerate carboxy-lyase it decreased only within rather than across generations. However, the flux changes seen in these reactions with individual pathogens were not repeated in simulations constrained with data from co-infection conditions. Competition between the two bacterial strains, or antagonism between the distinct signaling pathways they activate, might cause the responses of plant metabolism to differ in the more complex biotic environment of a co-infection.

CONCLUSIONS & FUTURE WORK

We find evidence that *A. thaliana* defensins are differentially regulated, albeit with inconsistent direction and magnitude, in response to different histories of infection with *Pseudomonas* pathogens. The same genes that respond to infections within a generation do so in the uninfected progeny of infected plants, suggesting that transgenerational defense induction is at work. With respect to metabolism, we find at the gene level that expression of triacylglycerol lipase consistently increases upon infection with different *Pseudomonas* pathogens and that this response is maintained in at least one subsequent generation. Notably, this enzyme acts early in the synthesis of jasmonic acid. *P. syringae* NP29.1a and *P. viridiflava* RMX3.1b are thought to activate distinct channels of defense signaling based on the expression of downstream marker

genes, with *P. viridiflava* inducing JA-mediated responses. However, the potential for increased JA synthesis in response to all bacterial pathogens indicates that pathogens interact with plant hormonal signals at multiple levels. It is therefore important to examine responses of both defense proteins and primary metabolism to infection.

At the reaction level, we find that a handful of pathways involved in photosynthesis and carbon assimilation are predicted to be perturbed by infection. A previous meta-study of multiple biotic stresses in several plant species showed that transcript levels of genes in photosynthesis and the carbon cycle were reduced regardless of the type of stress. This reduction in transcripts related to photosynthesis and carbon assimilation was proposed to be part of the plant defense response (Bilgin et al. 2010). Yet at the level of reactions rather than genes, we predict the activity of ferredoxin-related enzymes to decrease while the activity of L-glutamate:NAD+ oxidoreductase is predicted to increase. Thus, the responses of metabolism to infection at the reaction level are more complex than an increase or decrease in the broad subsystem of photosynthesis. Simulations on a model of primary metabolism can help us characterize these more granular differences and make predictions about how they affect the distribution of plant resources.

The metabolic changes induced by infection appear to persist in the uninfected offspring of infected plants. An open question is whether this transgenerational maintenance of infection-induced metabolic reconfiguration is adaptive. Changes in photosynthesis in response to infection suggest a tradeoff between growth and defense (Bilgin et al. 2010). If offspring are likely to encounter the same pathogens as their parents, then maintaining the optimal balance of

growth and defense for an infected state could better position them to use available resources efficiently.

Predictions about the transgenerational responses of metabolism to infection will require validation by means of targeted expression or metabolite quantification, especially because changes in predicted reaction flux could result from constraints applied upstream or downstream in connected pathways rather than to the enzymes in the responsive reaction. Mechanisms for transgenerational defense induction, whether maternal or epigenetic, also remain to be elucidated. Finally, controlled studies of pathogen growth or microbiome assembly will be necessary to establish whether the transcriptional changes associated with different infection histories actually alter the habitats available for microbial colonists in plant tissues.

The comparative tFBA approach we deploy here can further be used to find candidate reactions responsible for altering plant metabolism in the face of other environmental stresses. With a method to track these reconfigurations of the leaf niche, we can also begin to explore how host genetic variation influences gene and reaction-level responses to the environment.

APPENDIX A

Supplementary MATERIALS AND METHODS for Chapter 1

Data filtering and reduction

Since DADA2 relies on some initial assumptions about the error rate, we first examined the effect of different EE thresholds on the resulting feature table. While lower thresholds generally let more features and samples through, we noticed that even with a loose error threshold (~1/100 bases) there was a group of samples with a very low fraction of reads passing the filter. After pruning the samples to retain only those with 50% or more of reads passing the EE4 filter, 1427 samples remained.

The 1427 sets of forward and reverse reads were loaded into a QIIME2 artifact to tally the amplicon sequence variants present in each sample and perform quality filtering on the count table. Within QIIME2, we used cutadapt to remove primers and ran DADA2 to model errors and find amplicon sequence variants (ASVs). The initial table included 1421 samples with 10,987 sequence variants. We used the taxonomy created with the procedure described below to filter out reads matching mitochondrial or chloroplast sequences (removed 35 sequence variants and 102 samples). We filtered samples with notes in the metadata indicating any irregularities in the collection (removed 45 samples). We then filtered sequence variants with a frequency lower than 2 counts and samples with fewer than 10 reads (removed 19 features and 8 samples). This left a table with 10,803 taxa and 1,272 samples.

To generate the taxonomy we used the QIIME-SILVA bacterial 16S sequence database, release 128. The database taxonomy was constrained to seven levels (domain, phylum, class, order, family, genus, species) and taxonomy was assigned at a given level if on 99% of sequences shared an assignment. The sequences for primers 799F and 1193R were used to extract reads of 16S sequences in the database. The extracted reads were used to build a classifier using QIIME2's naive-bayes method. The classifier was used with the sklearn algorithm in QIIME2 to generate taxonomy assignments for the sequence variants. These taxonomic assignments were used to filter the table and list of sequences for mitochondrial and chloroplast sequences.

To generate a phylogeny for the sequence variants, QIIME2 was used to align the sequences with MAFFT and to infer (fasttree) and root a phylogenetic tree (phylogeny midpoint-root). The tree was imported along with the DADA2-generated ASV count table, the taxonomy, and the metadata into a phyloseq object in R for analysis.

Analysis in R

Sequence data were analyzed in R (version 3.4.4). Count table transformations, pruning, and rarefaction as well as distance matrix calculation and ordination were performed with packages phyloseq and vegan. PERMANOVA tests were performed with the adonis function. The VST transformation was performed using the DESeq2 package. Phylogenetic analysis was performed with ape and picante. Figures and supplemental figures were produced with ggplot2, ggpubr,

ggrepel, and directlabels.

Two approaches were used to asses the relationship of assemblage membership and sample variables. First, the proportion of ASVs shared between samples in random comparisons and comparisons conditioned on sample variables were compared. For each of the 1195 plant samples in the dataset, five other plant samples were randomly drawn without replacement and the mean proportion of ASVs they shared with the focal sample was calculated. Then, the same procedure was repeated with the five samples drawn only from those of the same tissue type, stage, site, or year as the focal sample. Finally, if samples taken from different tissues of the same plant were available, the average proportion of ASVs shared among these samples was calculated. The frequency distributions of these proportions were compared with a one-sided Wilcoxon rank sum test.

The second approach used a permutational multivariate analysis of variance to test for associations between sample variables and assemblage composition. The ASV counts were randomly subsampled without replacement to obtain an even sampling depth of one thousand reads. Assemblage composition was then quantified with the Raup-Crick dissimilarity index, the Bray-Curtis index, or the unweighted UniFrac distance. These distance matrices were ordinated with principle coordinate analysis (PcoA) to examine how samples grouped. The robustness of this grouping to data filters and normalizations was tested by repeating the ordination with count data subject to a VST or TMM transformation, global abundance and prevalence filters for ASV inclusion, or without rarefaction. Variance in the matrix of sample dissimilarities or distances

was compared within and between groups for each sample variable and the variables significantly associated with composition ($\alpha = 0.001$) were used to construct a multivariate model for sample variation. In the multivariate model, tissue type was nested within stage because different tissues emerged at specific stages. Stage was nested within year because the stages sampled in each study year varied (Table 1). The terms representing effects from sample processing were nested because the sample processing plates used for DNA extractions and PCRs were divided across four sequencing runs. These nested terms were ordered by their cumulative R^2 values.

To determine the dependence of these associations on the coarseness of taxonomic grouping in the dataset, tests of the multivariate model were repeated with distance matrices calculated from counts of genera, families, orders, classes, and phyla among the plant samples. For each taxonomic level above ASV, the unassigned ASVs were removed before ASVs were grouped into coarser taxa.

ASVs associated with specific tissues or developmental stages were identified were identified using the signassoc function of the indicspecies package. This function calculated an indicator value index (IndVal) based on the product of two probabilities: (1) the probability that a sample belonged to a habitat given ASV presence and (2) the probability that an ASV was present if a sample was taken from a habitat. For the habitats defined by each variable (six tissues, six developmental stages, two sites, and two years), indices were calculated independently for each ASV. The null hypothesis that no relationship existed between ASVs and conditions was tested

by comparing the empirical index with a distribution generated by randomly permuting the ASVpresence-absence count table. A two-tail p-value was used to select ASVs that are significantly more or less frequently observed in sampled belonging to a given condition (α = 0.01).

To investigate how consistently the ASVs driving tissue and stage associations in the dataset behaved across sites and years, the maximum prevalence of an ASV within each replication of the experiment was found. If ASVs reached maximum prevalence in the same tissues at each site and in each year they were observed, then they were considered spatially consistent.

Three metrics for assemblage diversity were compared between samples across plant developmental stages. Phylogenetic distance was measured as the summed branch lengths on the 16S phylogenetic tree between ASVs in the sample, with branch lengths weighted by ASV abundance. Shannon-Wiener indices from rarefied samples were calculated as $H' = -\Sigma^{s}_{i=1}(p_i)(\log_2 p_i)$. Evenness was assessed by the distributions of ASV relative abundances in samples with at least one hundred counts.

The diversity of bacterial lineages present was compared within and between root and rosette leaf samples at each developmental stage. For the 698 plant samples with at least one hundred counts and twenty ASVs, a matrix of pairwise dissimilarities (Raup-Crick, Bray-Curtis, or UniFrac) was generated. The dissimilarities for samples within and across groups were selected from the resulting matrix and compared with probability density plots.

APPENDIX B

An overview of the *A. thaliana* microbiomes in Chapter 1

Assemblages in plant tissues were depleted of phylum Acidobacteria and the classes Thermoleophilia, Bacilli, and Gemmatimonadetes relative to the surrounding soil. Data from the second study year, during which both plants and soil were collected at each timepoint, were filtered to exclude rare ASVs (total counts less than one thousand) and the relative abundance of each ASV was calculated for both plant root and soil samples. In each class, the distributions of these variant relative abundances were compared between soil and plant samples with a Kruskal-Wallis rank test ($\alpha = 0.001$) and the mean relative abundances in soil and plant samples were compared with a 95% confidence interval (Figure B1A). The relative abundance distributions of variants in Acidobacteria, Acidobacteria subgroup 2, Thermoleophilia, Bacilli, and Gemmatimonadetes differed significantly between soil and plant samples. The mean relative abundances of variants in these classes were reduced in roots compared to soil, indicating that the host plant might select against them during colonization.

Assemblages in plant tissues were mostly composed of class Actinobacteria and the subphyla α, β, and γ of Proteobacteria. To examine which clades were most abundant in plant assemblages, raw sequence counts for ASVs were summed for all plant samples and the mean abundance for variants in each class was compared with a 95% confidence interval (Figure B1B). Variants of Actinobacteria, Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria on average

had higher counts than variants of other classes by roughly one thousand, indicating that members of these lineages were the most common colonists of plant tissues.

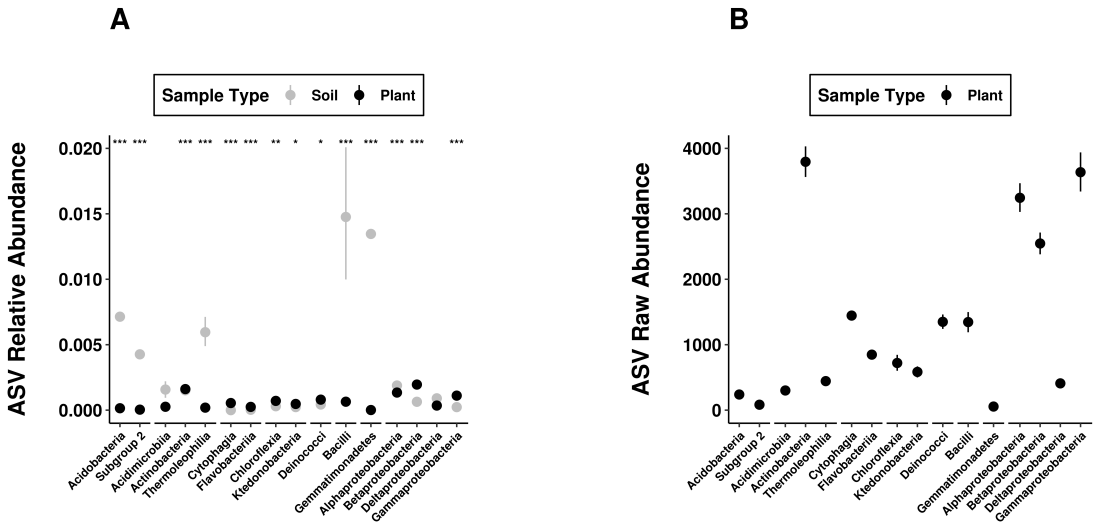

Figure B1 Plants hosted bacterial assemblages distinct from those in the surrounding soil. Plant assemblages were distinct from the surrounding soil and composed mostly of variants in class Actinobacteria and Proteobacteria subphyla α, β, and γ. These plots include observations of the 680 ASVs in the dataset with more than 000 total counts among 1181 plant and 77 soil samples. (A) The relative abundance of ASVs was calculated in both root samples and soil samples from all timepoints, sites, and years. The mean relative abundance for variants in each class, grouped by phylum on the x-axis, is plotted with a bootstrapped 95% confidence interval for both soil (gray) and plant (black) samples. A Kruskal-Wallis test indicates significant differences between soil and plant samples in the relative abundance distributions for variants of a class at alpha < 0.0001 (***), 0.001 (**), and 0.01 (*). (B) For plant samples, the counts of ASVs across tissues, stages, sites, and years were summed. The mean abundance for variants in each class, grouped by phylum on the x-axis, is plotted with a bootstrapped 95% confidence interval.

BIBLIOGRAPHY

Abdel-Farid IB, Jahangir M, van den Hondel CA, Kim HK, Choi YH, Verpoorte R. 2009. Fungal infection-induced metabolites in *Brassica rapa*. Plant Sci 176(5): 608-615. doi: 10.1016/j.plantsci.2009.01.017

Agler MT, Ruhe J, Kroll S, Morhenn C, Kim ST, Weigel D, Kemen EM. 2016. Microbial hub taxa link host and abiotic factors to plant microbiome variation. PLoS Biol 14(1). doi: 10.1371/journal.pbio.1002352

Ahn IP, Kim S, Lee YH. 2005. Vitamin B1 functions as an activator of plant disease resistance. Plant Physiol 138(3): 1505-1515. doi: 10.1104/pp.104.058693

Alamgir KM, Hojo Y, Christeller JT, Fukumoto K, Isshiki R, Shinya T, Baldwin IT, Galis I. 2016. Systematic analysis of rice (*Oryza sativa*) metabolic responses to herbivory. Plant Cell Environ 39(2): 453-466. doi: 10.1111/pce.12640

Albanese D, Donati C. 2017. Strain profiling and epidemiology of bacterial species from metagenomic sequencing. Nat Commun 8: 2260. doi: 10.1038/s41467-017-02209-5

Allesina S, Tang S. 2012. Stability criteria for complex ecosystems. Nature 483: 205–208. doi: 10.1038/nature10832

Amend AS, Cobian GM, Laruson AJ, Remple K, Tucker SJ, Poff KE, Antaky C, Boraks A, Jones CA, Kuehu D, Lensing BR. 2019. Phytobiomes are compositionally nested from the ground up. PeerJ 7: e6609. doi: 10.7717/peerj.6609

Anderson MJ. 2014. Permutational multivariate analysis of variance (PERMANOVA). Wiley statsref: statistics reference online, 1-15. doi: 10.1002/9781118445112.stat07841

Araki H, Tian D, Goss EM, Jakob K, Halldorsdottir SS, Kreitman M, Bergelson J. 2006. Presence/absence polymorphism for alternative pathogenicity islands in *Pseudomonas viridiflava*, a pathogen of *Arabidopsis*. Proc Natl Acad Sci 103: 5887–5892. doi: 10.1073/pnas.0601431103

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, Sneddon MW. 2018. KBase: the United States department of energy systems biology knowledgebase. Nat Biotechnol 36(7): 566. doi: 10.1038/nbt.4163

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT, Jiang R. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465(7298): 627-631. doi: 10.1038/nature08800

Bai Y, Müller DB, Srinivas G, Garrido-Oter R, Potthoff E, Rott M, Dombrowski N, Münch PC, Spaepen S, Remus-Emsermann M, et al. 2015. Functional overlap of the *Arabidopsis* leaf and root microbiota. Nature 528: 364. doi: 10.1038/nature16192

Banerjee S, Schlaeppi K, van der Heijden MG. Keystone taxa as drivers of microbiome structure and functioning. Nat Rev Microbiol 2018. 16(9): 567-76. doi: 10.1038/s41579-018-0024-1

Barabási AL, Albert R. Emergence of scaling in random networks. 1999. Science 286(5439): 509-12. doi: 10.1126/science.286.5439.509

Barak JD, Kramer LC, Hao LY. 2011. Colonization of tomato plants by *Salmonella enterica* is cultivar dependent, and type 1 trichomes are preferred colonization sites. Appl Environ Microbiol 77(2): 498-504. doi: 10.1128/aem.01661-10

Barberán A, Bates ST, Casamayor EO, Fierer N. 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J 6(2): 343-351. doi: 10.1038/ismej.2011.119

Barret M, Briand M, Bonneau S, Préveaux A, Valière S, Bouchez O, Hunault G, Simoneau P, Jacques M-A. 2015. Emergence shapes the structure of the seed microbiota. Appl Environ Microbiol 81: 1257-1266. doi: 10.1128/AEM.03722-14

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A. 2012. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41(D1): D991-D995. doi: 10.1093/nar/gks1193

Bartoli C, Frachon L, Barret M, Rigal M, Huard-Chauveau C, Mayjonade B, Zanchetta C, Bouchez O, Roby D, Carrère S, Roux F. 2018. In situ relationships between microbiota and potential pathobiota in *Arabidopsis thaliana*. ISME J 12(8): 2024-2038. doi: 10.1038/s41396-018-0152-7

Bartoli C, Frachon L, Barret M, Rigal M, Huard-Chauveau C, Mayjonade B, Zanchetta C, Bouchez O, Roby D, Carrère S, Roux F. 2018. In situ relationships between microbiota and potential pathobiota in *Arabidopsis thaliana*. ISME J 12(8): 2024-2038. doi: 10.1038/s41396-018-0152-7

Bartoli C, Roux F. 2017. Genome-wide association studies in plant pathosystems: Toward an ecological genomics approach. Front Plant Sci 8: 763. doi: 10.3389/fpls.2017.00763

Basilio AM, Medan D, Torretta JP, Bartoloni NJ. 2006. A year-long plant-pollinator network. Austral Ecology 31(8): 975-983. doi: 10.1111/j.1442-9993.2006.01666.x

Beattie GA, Lindow SE. 1999. Bacterial colonization of leaves: a spectrum of strategies. Phytopathology 89(5): 353-359. doi: 10.1094/PHYTO.1999.89.5.353

Beckers V, Dersch LM, Lotz K, Melzer G, Bläsing OE, Fuchs R, Ehrhardt T, Wittmann C. 2016. In silico metabolic network analysis of *Arabidopsis* leaves. BMC Syst Biol 10(1): 1-19. doi: 10.1186/s12918-016-0347-3

Beilsmith K, Perisin M, Bergelson J. 2020. Natural bacterial assemblages in *Arabidopsis thaliana* tissues become more distinguishable and diverse during host development. bioRxiv. doi: 10.1101/2020.03.04.958165

Berg G, Köberl M, Rybakova D, Müller H, Grosch R, Smalla K. 2017. Plant microbial diversity is suggested as the key to future biocontrol and health trends. FEMS Microbiol Ecol 93(5). doi: 10.1093/femsec/fix050

Bergelson J, Mittelstrass J, Horton MW. 2019. Characterizing both bacteria and fungi improves understanding of the *Arabidopsis* root microbiome. Sci Rep 9 (1): 24. doi: 10.1038/s41598-018-37208-z

Berger S, Papadopoulos M, Schreiber U, Kaiser W, Roitsch T. 2004. Complex regulation of gene expression, photosynthesis and sugar levels by pathogen infection in tomato. Physiol Plant 122(4): 419-428. doi: 10.1111/j.1399-3054.2004.00433.x

Berger S, Sinha AK, Roitsch T. 2007. Plant physiology meets phytopathology: plant primary metabolism and plant–pathogen interactions. J Exp Bot 58(15-16): 4019-4026. doi: 10.1093/jxb/erm298

Berry D, Widder S. 2014. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol 5: 219. doi: 10.3389/fmicb.2014.00219

Bilgin DD, Zavala JA, Zhu JI, Clough SJ, Ort DR, DeLucia EH. 2010. Biotic stress globally downregulates photosynthesis genes. Plant Cell Environ 33(10): 1597-1613. doi: 10.1111/j.1365-3040.2010.02167.x

Blekhman R, Goodrich JK, Huang K, Sun Q, Bukowski R, Bell JT, Spector TD, Keinan A, Ley RE, Gevers D, Clark AG. 2015. Host genetic variation impacts microbiome composition across human body sites. Genome Biol 16: 191. doi: 10.1186/s13059-015-0759-1

Bodenhausen N, Bortfeld-Miller M, Ackermann M, Vorholt JA. 2014. A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. PLoS Genet 10: e1004283. doi: 10.1371/journal.pgen.1004283

Bodenhausen N, Horton MW, Bergelson J. 2013. Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. PloS One 8(2), e56329. doi: 10.1371/journal.pone.0056329

Bohan DA, Vacher C, Tamaddoni-Nezhad A, Raybould A, Dumbrell AJ, Woodward G. 2017. Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. Trends Ecol Evol 32(7): 477-487. doi: 10.1016/j.tree.2017.03.001

Bolton MD. Primary metabolism and plant defense—fuel for the fire. 2009. Mol Plant Microbe Interact 22(5): 487-497. doi: 10.1094/MPMI-22-5-0487

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MG, Lee J, Ley R, Liu Y, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson, II MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CH, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2019. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. Nat Biotechnol 37(8): 852-857. doi: 10.7287/peerj.preprints.27295v2

Bonder MJ, Kurilshikov A, Tigchelaar EF, Mujagic Z, Imhann F, Vila AV, Deelen P, Vatanen T, Schirmer M, Smeekens SP, Zhernakova DV. 2016. The effect of host genetics on the gut microbiome. Nat Genet 48: 1407–1412. doi: 10.1038/ng.3663

Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F. 2010. Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. PLoS Genet 6(5). doi: 10.1371/journal.pgen.1000940

Brachi B, Morris GP, Borevitz JO. 2011. Genome-wide association studies in plants: The missing heritability is in the field. Genome Biol 12: 232. doi: 10.1186/gb-2011-12-10-232

Brandl MT, Lindow SE. 1998. Contribution of indole-3-acetic acid production to the epiphytic fitness of *Erwinia herbicola*. Appl Environ Microbiol 64(9): 3256-3263. doi: 10.1128/aem.64.9.3256-3263.1998

Bray JR, Curtis JT. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol Monogr 27(4): 325-349. doi: 10.2307/1942268

Broido AD, Clauset A. 2019. Scale-free networks are rare. Nat Commun 10(1): 1-10. doi: 10.1038/s41467-019-08746-5

Brown JH, Gillooly JF, Allen AP, Savage VM, West GB. 2004. Toward a metabolic theory of ecology. Ecology 85(7): 1771-1789. doi: 10.1890/03-9000

Bulgarelli D, Garrido-Oter R, Münch PC, Weiman A, Dröge J, Pan Y, McHardy AC, Schulze-Lefert P. 2015. Structure and function of the bacterial root microbiota in wild and domesticated barley. Cell Host Microbe 17(3): 392-403. doi: 10.1016/j.chom.2015.01.011

Bulgarelli D, Rott M, Schlaeppi K, van Themaat EVL, Ahmadinejad N, Assenza F, Rauf P, Huettel B, Reinhardt R, Schmelzer E, Peplies J. 2012. Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. Nature 488(7409): 91. doi: 10.1038/nature11336

Burch AY, Finkel OM, Cho JK, Belkin S, Lindow SE. 2013. Diverse microhabitats experienced by *Halomonas variabilis* on salt-secreting leaves. Appl Environ Microbiol 79: 845–852. doi: 10.1128/AEM.02791-12

Burkle LA, Alarcón R. 2011. The future of plant–pollinator diversity: understanding interaction networks across time, space, and global change. Am J Bot 98(3): 528-538. doi: 10.3732/ajb.1000391

Cáceres MD, Legendre P. 2009. Associations between species and groups of sites: indices and statistical inference. Ecology 90(12): 3566-3574. doi: 10.1890/08-1823.1

Cáceres, M. D. 2020. How to use theindicspeciespackage (ver. 1.7.8) cran.r-project.org/web/packages/indicspecies/vignettes/indicspeciesTutorial.pdf

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods 13(7): 581. doi: 10.1038/nmeth.3869

Canarini A, Kaiser C, Merchant A, Richter A, Wanek W. 2019. Root exudation of primary metabolites: mechanisms and their roles in plant responses to environmental stimuli. Front Plant Sci 10: 157. doi: 10.3389/fpls.2019.00157

Chaparro JM, Badri DV, Vivanco JM. 2014. Rhizosphere microbiome assemblage is affected by plant development. ISME J 8(4): 790. doi: 10.1038/ismej.2013.196

Chaparro JM, Sheflin AM, Manter DK, Vivanco JM. 2012. Manipulating the soil microbiome to increase soil health and plant fertility. Biol Fertil Soils 48: 489–499. doi: 10.1007/s00374-012-0691-4

Chinchilla D, Bauer Z, Regenass M, Boller T, Felix G. 2006. The *Arabidopsis* receptor kinase FLS2 binds flg22 and determines the specificity of flagellin perception. Plant Cell 18(2): 465-

476. doi: 10.1105/tpc.105.036574

Choi K, Choi J, Lee PA, Roy N, Khan R, Lee HJ, Weon HY, Kong HG, Lee SW. 2020. Alteration of Bacterial Wilt Resistance in Tomato Plant by Microbiota Transplant. Front Plant Sci 11: 1186. doi: 10.3389/fpls.2020.01186

Clarke CR, Hayes BW, Runde BJ, Markel E, Swingle BM, Vinatzer BA. 2016. Comparative genomics of *Pseudomonas syringae* pathovar tomato reveals novel chemotaxis pathways associated with motility and plant pathogenicity. PeerJ 4: e2570. doi: 10.7717/peerj.2570

Closs GP, Lake PS. 1994. Spatial and temporal variation in the structure of an intermittent-stream food web. Ecol Monogr 64(1): 1-21. doi: 10.2307/2937053

Coleman-Derr D, Desgarennes D, Fonseca-Garcia C, Gross S, Clingenpeel S, Woyke T, North G, Visel A, Partida-Martinez LP, Tringe SG. 2016. Plant compartment and biogeography affect microbiome composition in cultivated and native *Agave* species. New Phytol 209(2): 798-811. doi: 10.1111/nph.13697

Conway JR, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33(18): 2938-40. doi: 10.1093/bioinformatics/btx364

Coolen S, Proietti S, Hickman R, Davila Olivas NH, Huang PP, Van Verk MC, Van Pelt JA, Wittenberg AH, De Vos M, Prins M, Van Loon JJ. 2016. Transcriptome dynamics of *Arabidopsis* during sequential biotic and abiotic stresses. Plant J 86(3): 249-67. doi: 10.1111/tpj.13167

Coyte KZ, Schluter J, Foster KR. 2015. The ecology of the microbiome: networks, competition, and stability. Science 350(6261): 663-6. doi: 10.1126/science.aad2602

Csardi G, Nepusz T. The igraph software package for complex network research. 2006. igraph.org/r

Curtin SJ, Tiffin P, Guhlin J, Trujillo DI, Burghardt LT, Atkins P, Baltes NJ, Denny R, Voytas DF, Stupar RM, Young ND. 2017. Validating genome-wide association candidates controlling quantitative variation in nodulation. Plant Physiol 173: 921–931. doi: 10.1104/pp.16.01923

Davenport ER, Cusanovich DA, Michelini K, Barreiro LB, Ober C, Gilad Y. 2015. Genome-wide association studies of the human gut microbiota. PLoS One 10: e0140301. doi: 10.1371/journal.pone.0140301

de Vries FT, Griffiths RI, Bailey M, Craig H, Girlanda M, Gweon HS, Hallin S, Kaisermann A, Keith AM, Kretzschmar M, Lemanceau P. 2018. Soil bacterial networks are less stable under drought than fungal networks. Nat Commun 9(1): 3033. doi: 10.1038/s41467-018-05516-7

Delmas E, Besson M, Brice MH, Burkle LA, Dalla Riva GV, Fortin MJ, Gravel D, Guimarães Jr, PR, Hembry DH, Newman EA, Olesen JM. 2019. Analysing ecological networks of species interactions. Biol Rev 94(1): 16-36. doi: https://doi.org/10.1111/brv.12433

Demmitt BA, Corley RP, Huibregtse BM, Keller MC, Hewitt JK, McQueen MB, Knight R, McEdwards J, Johnson C, Santos-Medellín C, Lurie E, Podishetty NK, Bhatnagar S, Eisen JA, Sundaresan V. 2015. Structure, variation, and assembly of the root-associated microbiomes of rice. Proc Natl Acad Sci 112: E911–E920. doi: 10.1073/pnas.1414592112

Derocles SA, Bohan DA, Dumbrell AJ, Kitson JJ, Massol F, Pauvert C, Plantegenest M, Vacher C, Evans DM. 2018. Biomonitoring for the 21st century: integrating next-generation sequencing into ecological network analysis. In Advances in ecological research, vol. 58, pp. 1-62. Academic Press. doi: 10.1016/bs.aecr.2017.12.001

Eastmond PJ. 2006. SUGAR-DEPENDENT1 encodes a patatin domain triacylglycerol lipase that initiates storage oil breakdown in germinating *Arabidopsis* seeds. Plant Cell 18(3): 665-675. doi: 10.1105/tpc.105.040543

Edwards JA, Santos-Medellín CM, Liechty ZS, Nguyen B, Lurie E, Eason S, Phillips G, Sundaresan V. Compositional shifts in root-associated bacterial and archaeal microbiota track the plant life cycle in field-grown rice. 2018. PLoS Biol 16(2): e2003862. doi: 10.1371/journal.pbio.2003862

Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, Brachi B, Hagmann J, Grimm DG, Chen J, Busch W. 2018. The rate and potential relevance of new mutations in a colonizing plant lineage. PLoS Genet 14(2): e1007155. doi: 10.1371/journal.pgen.1007155

Fahimipour AK, Kardish MR, Lang JM, Green JL, Eisen JA, Stachowicz JJ. 2017. Global-scale structure of the Eelgrass microbiome. Appl Environ Microbiol 83: e03391-16. doi: 10.1128/aem.03391-16

Fan K, Weisenhorn P, Gilbert JA, Shi Y, Bai Y, Chu H. 2018. Soil pH correlates with the co-occurrence and assemblage process of diazotrophic communities in rhizosphere and bulk soils of wheat fields. Soil Biol Biochem 121: 185-192. doi: 10.1016/j.soilbio.2018.03.017

Faust K, Lahti L, Gonze D, De Vos WM, Raes J. 2015. Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr Opin Microbiol 25: 56-66. doi: 10.1016/j.mib.2015.04.004

Fernandez O, Theocharis A, Bordiec S, Feil R, Jacquens L, Clément C, Fontaine F, Barka EA. 2012. *Burkholderia phytofirmans* PsJN acclimates grapevine to cold by modulating carbohydrate metabolism. Mol Plant Microbe Interact 25(4): 496-504. doi: 10.1094/MPMI-09-11-0245

Figueiredo ART, Kramer J. 2020. Cooperation and conflict within the microbiota and their effects on animal hosts. Front Ecol Evol 8: 132. doi: 10.3389/fevo.2020.00132

Fitzpatrick CR, Copeland J, Wang PW, Guttman DS, Kotanen PM, Johnson MT. 2018. Assembly and ecological function of the root microbiome across angiosperm plant species. Proc Natl Acad Sci 115(6): E1157-E1165. doi: 10.1073/pnas.1717617115

Fitzpatrick CR, Schneider AC. 2020. Unique bacterial assembly, composition, and interactions in a parasitic plant and its host. J Exp Bot 71(6): 2198-2209. doi: 10.1093/jxb/erz572

Fitzpatrick TB, Chapman LM. The importance of thiamine (vitamin B1) in plant health: From crop yield to biofortification. 2020. J Biol Chem 295(34): 12002-12013. doi: 10.1074/jbc.REV120.010918

Fortuna MA, Barbour MA, Zaman L, Hall AR, Buckling A, Bascompte J. 2019. Coevolutionary dynamics shape the structure of bacteria-phage infection networks. Evolution 73(5): 1001-1011. doi: 10.1111/evo.13731

Foster KR, Schluter J, Coyte KZ, Rakoff-Nahoum S. 2017. The evolution of the host microbiome as an ecosystem on a leash. Nature 548: 43–51. doi: 10.1038/nature23292

Fouhy F, Clooney AG, Stanton C, Claesson MJ, Cotter PD. 2016. 16S rRNA gene sequencing of mock microbial populations- impact of DNA extraction method, primer choice and sequencing platform. BMC Microbiol 16: 123 . doi: 10.1186/s12866-016-0738-z

Frachon L, Mayjonade B, Bartoli C, Hautekèete NC, Roux F. 2019. Adaptation to plant communities across the genome of *Arabidopsis thaliana*. Mol Biol Evol 36(7): 1442-1456. doi: 10.1093/molbev/msz078

Franche C, Lindström K, Elmerich C. 2009. Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. Plant Soil 321(1-2): 35-59. doi: 10.1007/s11104-008-9833-8

Freilich S, Zarecki R, Eilam O, Segal ES, Henry CS, Kupiec M, Gophna U, Sharan R, Ruppin E. 2011. Competitive and cooperative metabolic interactions in bacterial communities. Nat Commun 2(1): 1-7. doi: 10.1038/ncomms1597

Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. 2012. PLoS Comput Biol 8(9): e1002687. doi: 10.1371/journal.pcbi.1002687

Gallagher MD, Chen-Plotkin AS. 2018. The post-GWAS era: From association to function. Am J Hum Genet 102: 717–730. doi: 10.1016/j.ajhg.2018.04.002

Ganguly DR, Crisp PA, Eichten SR, Pogson BJ. 2017. The *Arabidopsis* DNA methylome is stable under transgenerational drought stress. Plant Physiol 175(4): 1893-1912. doi:

10.1104/pp.17.00744

Ganjgahi H, Winkler AM, Glahn DC, Blangero J, Donohue B, Kochunov P, Nichols TE. 2018. Fast and powerful genome wide association of dense genetic data with high dimensional imaging phenotypes. Nat Commun 9(1): 1-3. doi: https://doi.org/10.1038/s41467-018-05444-6

Germaine K, Keogh E, Garcia-Cabellos G, Borremans B, Van Der Lelie D, Barac T, Oeyen L, Vangronsveld J, Moore FP, Moore ER, Campbell CD. 2004. Colonisation of poplar trees by gfp expressing bacterial endophytes. FEMS Microbiol Ecol 48(1): 109-118. doi: 10.1016/j.femsec.2003.12.009

Gohl DM, Vangay P, Garbe J, MacLean A, Hauge A, Becker A, Gould TJ, Clayton JB, Johnson TJ, Hunter R, Knights D. 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. Nat Biotechnol 34(9): 942-949. doi: 10.1038/nbt.3601

Goldford JE, Lu N, Bajić D, Estrela S, Tikhonov M, Sanchez-Gorostiaga A, Segrè D, Mehta P, Sanchez A. 2018. Emergent simplicity in microbial community assembly. Science 361(6401): 469-474. doi: 10.1126/science.aat1168

Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, Clark AG, Ley RE. 2016. Genetic determinants of the gut microbiome in UK twins. Cell Host Microbe 19: 731–743. doi: 10.1016/j.chom.2016.04.017

Gould AL, Zhang V, Lamberti L, Jones EW, Obadia B, Korasidis N, Gavryushkin A, Carlson JM, Beerenwinkel N, Ludington WB. 2018. Microbiome interactions shape host fitness. Proc Natl Acad Sci 115(51): E11951-E11960. doi: 10.1073/pnas.1809349115

Grady KL, Sorensen JW, Stopnisek N, Guittar J, Shade A. 2019. Abundance-occupancy distributions to prioritize plant core microbiome membership. Curr Opin Microbiol 49: 50-58. doi: 10.1016/j.mib.2019.09.008

Grilli J, Rogers T, Allesina S. Modularity and stability in ecological communities. 2016. Nat Commun 7(1): 1-0. doi: 10.0.4.14/ncomms12031

Gurney J, Aldakak L, Betts A, Gougat-Barbera C, Poisot T, Kaltz O, Hochberg ME. 2017. Network structure and local adaptation in co-evolving bacteria–phage interactions. Mol Ecol 26(7): 1764-1777. doi: 10.1111/mec.14008

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hackinger S, Zeggini E. 2017. Statistical methods to detect pleiotropy in human complex traits. Open Biol 7: 170125. doi: 10.1098/rsob.170125

Hagen M, Kissling WD, Rasmussen C, De Aguiar MA, Brown LE, Carstensen DW, Alves-Dos-

Santos I, Dupont YL, Edwards FK, Genini J, Guimaraes Jr PR. 2012. Biodiversity, species interactions and ecological networks in a fragmented world. In Advances in ecological research, vol. 46, pp. 89-210. Academic Press. doi: 10.1016/B978-0-12-396992-7.00002-2

Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, Kay DM, Tenesa A, Kusel VI, Sheehan P, Eaaswarkhanth M, Yearout D. 2011. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. PLoS Genet 7: e1002237. doi: 10.1371/journal.pgen.1002237

Hannula SE, Kielak AM, Steinauer K, Huberty M, Jongen R, Jonathan R, Heinen R, Bezemer TM. 2019. Time after time: temporal variation in the effects of grass and forb species on soil bacterial and fungal communities. mbio 10(6). doi: 10.1128/mBio.02635-19

Hardoim PR, van Overbeek LS, Berg G, Pirttilä AM, Compant S, Campisano A, Döring M, Sessitsch A. 2015. The Hidden World within Plants: Ecological and Evolutionary Considerations for Defining Functioning of Microbial Endophytes. Microbiol Mol Biol Rev 79(3): 293-320. doi: 10.1128/MMBR.00050-14

Henry CS, DeJongh M, Best AA, Frybarger PM, Linsay B, Stevens RL. 2010. High-throughput generation, optimization and analysis of genome-scale metabolic models. Nat Biotechnol 28(9): 977-982. doi: 10.1038/nbt.1672

Herrera Paredes S, Gao T, Law TF, Finkel OM, Mucyn T, Teixeira PJ, Salas González I, Feltcher ME, Powers MJ, Shank EA, Jones CD. 2018. Design of synthetic bacterial communities for predictable plant phenotypes. PLoS Biol 16: e2003962. doi: 10.1371/journal.pbio.2003962
Herrera, C.M., 1988. Variation in mutualisms: the spatiotemporal mosaic of a pollinator assemblage. Biol J Linn Soc Lond 35(2): 95-125. doi: 10.1111/j.1095-8312.1988.tb00461.x

Hillebrand H, Bennett DM, Cadotte MW. 2008. Consequences of dominance: a review of evenness effects on local and regional ecosystem processes. Ecology 89: 1510–1520. doi: 10.1890/07-1053.1

Hirano H, Takemoto K. 2019. Difficulty in inferring microbial community structure based on co-occurrence network approaches. BMC Bioinformatics 20(1): 1-14. doi: 10.1186/s12859-019-2915-1

Hirano SS, Upper CD. 2000. Bacteria in the leaf ecosystem with emphasis on *Pseudomonas syringae*—a pathogen, ice nucleus, and epiphyte. Microbiol Mol Biol Rev 64(3): 624-653. doi: 10.1128/mmbr.64.3.624-653.2000

Holeski LM, Jander G, Agrawal AA. 2012. Transgenerational defense induction and epigenetic inheritance in plants. Trends Ecol Evol 27(11): 618-626. doi: https://doi.org/10.1016/j.tree.2012.07.011

Horton MW, Bodenhausen N, Beilsmith K, Meng D, Muegge BD, Subramanian S, Vetter MM, Vilhjálmsson BJ, Nordborg M, Gordon JI, Bergelson J. 2014. Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. Nat Commun 5: 5320. doi: 10.1038/ncomms6320

Howard BE, Hu Q, Babaoglu AC, Chandra M, Borghi M, Tan X, He L, Winter-Sederoff H, Gassmann W, Veronese P, Heber S. 2013. High-throughput RNA sequencing of *Pseudomonas*-infected *Arabidopsis* reveals hidden transcriptome complexity and novel splice variants. PLoS One 8(10): e74183. doi: 10.1371/journal.pone.0074183

Hurley B, Lee D, Mott A, Wilton M, Liu J, Liu YC, Angers S, Coaker G, Guttman DS, Desveaux D. 2014. The Pseudomonas syringae type III effector HopF2 suppresses *Arabidopsis* stomatal immunity. PLoS One 9(12). doi: 10.1371/journal.pone.0114921

Hwang MS, Morgan RL, Sarkar SF, Wang PW, Guttman DS. 2005. Phylogenetic characterization of virulence and resistance phenotypes of *Pseudomonas syringae*. Appl Environ Microbiol 71(9): 5182-5191. doi: 10.1128/AEM.71.9.5182-5191.2005

Igartua C, Davenport ER, Gilad Y, Nicolae DL, Pinto J, Ober C. 2017. Host genetic variation in mucosal immunity pathways influences the upper airway microbiome. Microbiome 5: 16. doi: 10.1186/s40168-016-0227-5

Ings TC, Montoya JM, Bascompte J, Blüthgen N, Brown L, Dormann CF, Edwards F, Figueroa D, Jacob U, Jones JI, Lauridsen RB. 2009. Ecological networks–beyond food webs. J Anim Ecol 78(1): 253-269. doi: 10.1111/j.1365-2656.2008.01460.x

Jakob K, Goss EM, Araki H, Van T, Kreitman M, Bergelson J. 2002. *Pseudomonas viridiflava* and *P. syringae*—natural pathogens of *Arabidopsis thaliana*. Mol Plant Microbe Interact 15(12): 1195-1203. doi: 10.1094/MPMI.2002.15.12.1195

Jakob K, Kniskern JM, Bergelson J. 2007. The role of pectate lyase and the jasmonic acid defense response in *Pseudomonas viridiflava* virulence. Mol Plant Microbe Interact 20(2): 146-158. doi: 10.1094/MPMI-20-2-0146

Janda JM, Abbott SL. 2007. 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. J Clin Microbiol 45(9): 2761-4. doi: 10.1128/JCM.01228-07

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45: D353–D361. doi: 10.1093/nar/gkw1092

Karasov TL, Almario J, Friedemann C, Ding W, Giolai M, Heavens D, Kersten S, Lundberg DS, Neumann M, Regalado J, Neher RA. 2018. *Arabidopsis thaliana* and *Pseudomonas* pathogens

exhibit stable associations over evolutionary timescales. Cell Host Microbe 24(1): 168-179. doi: 10.1016/j.chom.2018.06.011

Karasov TL, Chae E, Herman JJ, Bergelson J. 2017. Mechanisms to mitigate the trade-off between growth and defense. Plant Cell 29(4): 666-680. doi: 10.1105/tpc.16.00931

Karasov TL, Kniskern JM, Gao L, DeYoung BJ, Ding J, Dubiella U, Lastra RO, Nallu S, Roux F, Innes RW, Barrett LG. 2014. The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature 512: 436–440. doi: 10.1038/nature13439

Kassambara A. 2018. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.1.8. CRAN.R-project.org/package=ggpubr

Katagiri F, Thilmony R, He SY. 2002. The *Arabidopsis thaliana-Pseudomonas syringae* interaction. The Arabidopsis Book/American Society of Plant Biologists. The *Arabidopsis* Book/American Society of Plant Biologists 1. doi: 10.1199/tab.0039

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4): 772-780. doi: 10.1093/molbev/mst010

Keegan KP, Glass EM, Meyer F. 2016. MG-RAST, a metagenomics service for analysis of microbial community structure and function. Methods Mol Biol 1399: 207–233. doi: 10.1007/978-1-4939-3369-3_13

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. Bioinformatics 26: 1463-1464. doi: 10.1093/bioinformatics/btq166

Kembel SW, O'Connor TK, Arnold HK, Hubbell SP, Wright SJ, Green JL. 2014. Relationships between phyllosphere bacterial communities and plant functional traits in a neotropical forest. Proc Natl Acad Sci 111(38): 13715-13720. doi: 10.1073/pnas.1216057111

Khare E, Mishra J, Arora NK. 2018. Multifaceted Interactions Between Endophytes and Plant: Developments and Prospects. Front Microbiol 9: 2732. doi: 10.3389/fmicb.2018.02732

Kliebenstein DJ. 2004. Secondary metabolites and plant/environment interactions: a view through *Arabidopsis thaliana* tinged glasses. Plant Cell Environ 27(6): 675-684. doi: 10.1111/j.1365-3040.2004.01180.x

Kloepper JW, Ryu C-M, Zhang S. 2004. Induced systemic resistance and promotion of plant growth by *Bacillus* spp. Phytopathology 94: 1259-1266. doi: 10.1094/PHYTO.2004.94.11.1259

Kloepper JW, Ryu C. Bacterial endophytes as elicitors of induced systemic resistance. In

Microbial root endophytes, pp. 33-52. Springer, Berlin, Heidelberg, 2006. doi: 10.1007/3-540-33526-9_3

Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, Kosciolek T, McCall LI, McDonald D, Melnik AV. 2018. Best practices for analysing microbiomes. Nat Rev Microbiol 16: 410–422. doi: 10.1038/s41579-018-0029-9

Knights D, Silverberg MS, Weersma RK, Gevers D, Dijkstra G, Huang H, Tyler AD, Van Sommeren S, Imhann F, Stempak JM, Huang H. 2014. Complex host genetics influence the microbiome in inflammatory bowel disease. Genome Med 6: 107. doi: 10.1186/s13073-014-0107-1

Kniskern JM, Traw MB, Bergelson J. 2007. Salicylic acid and jasmonic acid signaling defense pathways reduce natural bacterial diversity on *Arabidopsis thaliana*. Mol Plant Microbe Interact 20(12): 1512-22. doi: 10.1094/MPMI-20-12-1512

Kolde R, Franzosa EA, Rahnavard G, Hall AB, Vlamakis H, Stevens C, Daly MJ, Xavier RJ, Huttenhower C. 2018. Host genetic variation and its microbiome interactions within the Human Microbiome Project. Genome Med 10: 6. doi: 10.1186/s13073-018-0515-8

Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. 2012. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet 44: 1066-1071. doi: 10.1038/ng.2376

Kowalchuk GA, Yergeau E, Leveau JH, Sessitsch A, Bailey M, Liu W. Plant-associated microbial communities. In Environmental molecular microbiology, pp. 131-148. New York, Caister Academic Press, 2010.

Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss, PD. 2013. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol 79(17): 5112-5120. doi: 10.1128/AEM.01043-13

Krämer U. 2015. The natural history of model organisms: Planting molecular functions in an ecological context with *Arabidopsis thaliana*. Elife 4: e06100. doi: 10.7554/eLife.06100

Krüger D, Buscot F. 2016. Life in leaf litter: novel insights into community dynamics of bacteria and fungi during litter decomposition. Mol Ecol 25(16): 4059-4074. doi: 10.1111/mec.13739

Kuiper I, Bloemberg GV, Lugtenberg B.J. 2001. Selection of a plant-bacterium pair as a novel tool for rhizostimulation of polycyclic aromatic hydrocarbon-degrading bacteria. Mol Plant Microbe Interact 14(10): 1197-1205. doi: 10.1094/MPMI.2001.14.10.1197

Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. 2015. Sparse and

compositionally robust inference of microbial ecological networks. PLoS Comput Biol 11(5): e1004226. doi: 10.1371/journal.pcbi.1004226

Laforest-Lapointe I, Messier C, Kembel SW. 2016. Tree phyllosphere bacterial communities: exploring the magnitude of intra-and inter-individual variation among host species. PeerJ 4: e2367. doi: 10.7717/peerj.2367

Laine AL, Hanski I. 2006. Large-scale spatial dynamics of a specialist plant pathogen in a fragmented landscape. J Ecol 94(1): 217-226. doi: 10.1111/j.1365-2745.2005.01075.x

Lambais MR, Crowley DE, Cury JC, Büll RC, Rodrigues RR. 2006. Bacterial diversity in tree canopies of the Atlantic forest. Science 312: 1917. doi: 10.1126/science.1124696

Lassalle F, Campillo T, Vial L, Baude J, Costechareyre D, Chapulliot D, Shams M, Abrouk D, Layeghifard M, Hwang DM, Guttman DS. 2017. Disentangling interactions in the microbiome: a network perspective. Trends Microbiol 25(3): 217-228. doi: 10.1016/j.tim.2016.11.008

Lazzaro BP, Fox GM. 2017. Host–Microbe Interactions: Winning the Colonization Lottery. Curr Biol 27(13): R642-R644. doi: 10.1016/j.cub.2017.05.068

Lebeis SL, Paredes SH, Lundberg DS, Breakfield N, Gehring J, McDonald M, Malfatti S, Del Rio TG, Jones CD, Tringe SG, Dangl JL. 2015. Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. Science 349(6250): 860-864. doi: 10.1126/science.aaa8764

Leff JW, Del Tredici P, Friedman WE, Fierer N. 2015. Spatial structuring of bacterial communities within individual *Ginkgo biloba* trees. Environ Microbiol 17(7): 2352-2361. doi: 10.1111/1462-2920.12695

Levy A, Conway JM, Dangl JL, Woyke T. 2018. Elucidating bacterial gene functions in the plant microbiome. Cell Host Microbe 24: 475–485. doi: 10.1016/j.chom.2018.09.005

Li JR, Liu CC, Sun CH, Chen YT. 2018. Plant stress RNA-seq Nexus: a stress-specific transcriptome database in plant cells. BMC Genomics 19(1): 966. doi: 10.1186/s12864-018-5367-5

Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO. 2010. Association mapping of local climate-sensitive quantitative trait loci in *Arabidopsis thaliana*. Proc Natl Acad Sci 107: 21199–21204. doi: 10.1073/pnas.1007431107

Lindow SE, Brandl MT. 2003. Microbiology of the phyllosphere. Appl Environ Microbiol 69: 1875-1883. doi: 10.1128/aem.69.4.1875-1883.2003

Liu H, Roeder K, Wasserman L. 2010. Stability approach to regularization selection (StARS) for

high dimensional graphical models. In Advances in Neural Information Processing Systems, 24(2): 1432-1440.

López-Gresa MP, Maltese F, Bellés JM, Conejero V, Kim HK, Choi YH, Verpoorte R. 2010. Metabolic response of tomato leaves upon different plant–pathogen interactions. Phytochem Anal 21(1): 89-94. doi: 10.1002/pca.1179

Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71(12): 8228-8235. doi: 10.1128/AEM.71.12.8228-8235.2005

Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. 2012. Diversity, stability and resilience of the human gut microbiota. Nature 489(7415): 220-230. doi: 10.1038/nature11550

Luna E, Bruce TJ, Roberts MR, Flors V, Ton J. 2012. Next-generation systemic acquired resistance. Plant Physiol 158(2): 844-853. doi: https://doi.org/10.1104/pp.111.187468

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S, Tremblay J, Engelbrektson A, Kunin V, Del Rio TG, Edgar RC. 2012. Defining the core *Arabidopsis thaliana* root microbiome. Nature 488(7409): 86. doi: 10.1038/nature11237

Ma F, Lin P, Chen Q, Lu X, Zhang YE, Wu C-I. 2018. Direct measurement of pervasive weak repression by microRNAs and their role at the network level. BMC Genomics 19: 362. doi: 10.1186/s12864-018-4757-z

Maciá-Vicente JG, Nam B, Thines M. 2020. Community sequencing on a natural experiment reveals little influence of host species and timing but a strong influence of compartment on the composition of root endophytes in three annual Brassicaceae. bioRxiv. doi: 10.1101/2020.01.24.918177

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(1): e00682-13. doi: 10.1128/mBio.00682-13

Malinovsky FG, Fangel JU, Willats WG. 2014. The role of the cell wall in plant immunity. Front Plant Sci 5: 178. doi: 10.3389/fpls.2014.00178

Marshall DD, Powers R. 2017. Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. Prog Nucl Magn Reson Spectrosc 100: 1-16. doi: 10.1016/j.pnmrs.2017.01.001

Marti G, Erb M, Boccard J, Glauser G, Doyen GR, Villard N, Robert CA, Turlings TC, Rudaz S, Wolfender JL. 2013. Metabolomics reveals herbivore-induced metabolites of resistance and susceptibility in maize leaves and roots. Plant Cell Environ 36(3): 621-639. doi:

10.1111/pce.12002

Marulanda A, Azcón R, Chaumont F, Ruiz-Lozano JM, Aroca R. 2010. Regulation of plasma membrane aquaporins by inoculation with a *Bacillus megaterium* strain in maize (*Zea mays* L.) plants under unstressed and salt-stressed conditions. Planta 232(2): 533-543. doi: 10.1007/s00425-010-1196-8

Massoni J, Bortfeld-Miller M, Jardillier L, Salazar G, Sunagawa S, Vorholt, JA. 2019. Consistent host and organ occupancy of phyllosphere bacteria in a community of wild herbaceous plant species. ISME J 14(1): 245-258. doi: 10.1038/s41396-019-0531-8

Mauch-Mani B, Baccelli I, Luna E, Flors V. 2017. Defense priming: an adaptive part of induced resistance. Annu Rev Plant Biol 68: 485-512. doi: 10.1146/annurev-arplant-042916-041132

McCann KS. 2000. The diversity-stability debate. Nature 405: 228–233. doi: 10.1038/35012234

McCoy CO, Matsen IV FA. 2013. Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. PeerJ 1: e157. doi: 10.7717/peerj.157

McIntyre AB, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, Minot SS, Danko D, Foox J, Ahsanuddin S, Tighe S. 2017. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol 18: 182. doi: 10.1186/s13059-017-1299-7

McMurdie PJ, Holmes S. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS One 8(4): e61217. doi: 10.1371/journal.pone.0061217

Meaden S, Metcalf CJE, Koskella B. 2016. The effects of host age and spatial location on bacterial community composition in the English Oak tree (*Quercus robur*). Environ Microbiol Rep 8: 649–658. doi: 10.1111/1758-2229.12418

Melotto M, Zhang L, Oblessuc PR, He SY. 2017. Stomatal defense a decade later. Plant Physiol 174(2): 561-571. doi: 10.1104/pp.16.01853

Mendez LW, Raaijmakers JM, de Hollander M, Mendes R, Tsai SM. 2018. Influence of resistance breeding in common bean on rhizosphere micorbiome composition and function. ISME J 12: 212–224. doi: 10.1038/ismej.2017.158

Mercado-Blanco J, Rodrıguez-Jurado D, Hervás A, and Jiménez-Dıaz RM. 2004. Suppression of *Verticillium* wilt in olive planting stocks by root-associated fluorescent *Pseudomonas* spp. Biol Control 30(2): 474-486. doi: 10.1016/j.biocontrol.2004.02.002

Mergner J, Frejno M, List M, Papacek M, Chen X, Chaudhary A, Samaras P, Richter S, Shikata H, Messerer M, Lang D. 2020. Mass-spectrometry-based draft of the *Arabidopsis* proteome.

Nature 579(7799): 409-414. doi: 10.1038/s41586-020-2094-2

Meyer KM, Leveau JH. 2012. Microbiology of the phyllosphere: a playground for testing ecological concepts. Oecologia 168(3): 621-629. doi: 10.1007/s00442-011-2138-2

Miao Z, Han Z, Zhang T, Chen S, Ma C. 2017. A systems approach to a spatio-temporal understanding of the drought stress response in maize. Sci Rep 7(1): 1-14. doi: 10.1038/s41598-017-06929-y

Mighell K, Saltonstall K, Turner BL, Espinosa-Tasón J, Van Bael, SA. 2019. Abiotic and biotic drivers of endosymbiont community assembly in *Jatropha curcas*. Ecosphere 10(11). doi: 10.1002/ecs2.2941

Mitchell-Olds T, Schmitt J. 2006. Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. Nature 441(7096): 947-952. doi: 10.1038/nature04878

Mitchell-Olds T. 2001. *Arabidopsis thaliana* and its wild relatives: a model system for ecology and evolution. Trends Ecol Evol 16 (12): 693-700. doi: 10.1016/S0169-5347(01)02291-1

Morella NM, Gomez AL, Wang G, Leung MS, Koskella B. 2018. The impact of bacteriophages on phyllosphere bacterial abundance and composition. Mol Ecol 27: 2025–2038. doi: 10.1111/mec.14542

Morgan JL, Darling AE, Eisen JA. 2010. Metagenomic sequencing of an in vitro-simulated microbial community. PLoS One 5: e10209. doi: 10.1371/journal.pone.0010209

Morris CE, Lamichhane JR, Nikolić I, Stanković S, Moury B. 2019. The overlapping continuum of host range among strains in the *Pseudomonas syringae* complex. Phytopathology Research 1(1): 4. doi: 10.1186/s42483-018-0010-6

Nawrath C, Schreiber L, Franke RB, Geldner N, Reina-Pinto JJ, Kunst L. 2013. Apoplastic diffusion barriers in *Arabidopsis*. *Arabidopsis* Book (American Society of Plant Biologists), 11. doi: 10.1199/tab.0167

Naylor D, DeGraaf S, Purdom E, Coleman-Derr D. 2017. Drought and host selection influence bacterial community dynamics in the grass root microbiome. ISME J 11(12): 2691-704. doi: 10.1038/ismej.2017.118

Nelissen H, Moloney M, Inzé D. Translational research: from pot to plot. Plant biotechnology journal. 2014. Plant Biotechnol J 12(3): 277-285. doi: 10.1111/pbi.12176

Nomura K, Mecey C, Lee YN, Imboden LA, Chang JH, He SY. 2011. Effector-triggered immunity blocks pathogen degradation of an immunity-associated vesicle traffic regulator in Arabidopsis. Proc Natl Acad Sci 108(26): 10774-10779. doi: 10.1073/pnas.1103338108

Obadia B, Güvener ZT, Zhang V, Ceja-Navarro JA, Brodie EL, William WJ, Ludington WB. 2017. Probabilistic invasion underlies natural gut microbiome stability. Curr Biol 27(13): 1999-2006. doi: 10.1016/j.cub.2017.05.034

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2017. vegan: Community Ecology Package. R package version 2.4-5. CRAN.R-project.org/package=vegan

Olesen JM, Bascompte J, Elberling H, Jordano P. 2008. Temporal dynamics in a pollination network. Ecology 89(6): 1573-1582. doi: 10.1890/07-0451.1

Olesen JM, Dupont YL, O'Gorman, E., Ings, T.C., Layer, K., Melián, C.J., Trøjelsgaard, K., Pichler, D.E., Rasmussen, C. and Woodward, G. 2010. From Broadstone to Zackenberg: space, time and hierarchies in ecological networks. In Advances in ecological research, vol. 42, pp. 1-69. Academic Press. doi: 10.1016/B978-0-12-381363-3.00001-0

Olesen JM, Stefanescu C, Traveset A. 2011. Strong, long-term temporal dynamics of an ecological network. PloS One 6(11): e26455. doi: 10.1371/journal.pone.0026455

Org E, Parks BW, Joo JW, Emert B, Schwartzman W, Kang EY, Mehrabian M, Pan C, Knight R, Gunsalus R, Drake TA. 2015. Genetic and environmental control of host-gut microbiota interactions. Genome Research 25: 1558–1569. doi: 10.1101/gr.194118.115

Orth JD, Thiele I, Palsson BØ. 2010. What is flux balance analysis? Nat Biotechnol. 28(3): 245-248. doi: 10.1038/nbt.1614

Paradis E, Schliep K. 2018. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35(3): 526-528. doi: 10.1093/bioinformatics/bty633

Pascale A, Proietti S, Pantelides IS, Stringlis IA. 2020. Modulation of the root microbiome by plant molecules: the basis for targeted disease suppression and plant growth promotion. Front Plant Sci 10: 1741. doi: 10.3389/fpls.2019.01741

Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL, Buckler ES, Ley RE. 2013. Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc Natl Acad Sci 110(16): 6548-6553. doi: 10.1073/pnas.1302837110

Peredo EL, Simmons SL. 2018. Leaf-FISH: Microscale imaging of bacterial taxa on phyllosphere. Front Microbiol 8: 2669. doi: 10.3389/fmicb.2017.02669

Perisin M. 2016. The dynamics of bacterial communities associated with *Arabidopsis thaliana*. ProQuest Dissertations Publishing 10052860.

Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. Curr Opin Microbiol 6: 498–505. doi: 10.1016/j.mib.2003.09.008

Piasecka A, Kachlicki P, Stobiecki M. 2019. Analytical methods for detection of plant metabolomes changes in response to biotic and abiotic stresses. Int J Mol Sci 20(2): 379. doi: 10.3390/ijms20020379

Ploch S, Rose LE, Bass D, Bonkowski M. 2016. High diversity revealed in leaf-associated protists (Rhizaria: Cercozoa) of Brassicaceae. J Eukaryot Microbiol 63(5), 635-641. doi: 10.1111/jeu.12314

Poisot T, Stouffer DB, Gravel, D. 2015. Beyond species: why ecological interaction networks vary through space and time. Oikos 124(3): 243-251. doi: 10.1111/oik.01719

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PloS One 5(3): e9490. doi: 10.1371/journal.pone.0009490

Props R, Kerckhof F-M, Rubbens P, De Vrieze J, Hernandez Sanabria E, Waegeman W, Monsieurs P, Hammes F, Boon N. 2017. Absolute quantification of microbial taxon abundances. ISME J 11: 584–587. doi: 10.1038/ismej.2016.117

Qi G, Ma G, Chen S, Lin C, Zhao X. 2019. Microbial network and soil properties are changed in bacterial wilt-susceptible soil. Appl Environ Microbiol 85(13). doi: 10.1128/AEM.00162-19

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41(D1): D590-D596. doi: 10.1093/nar/gks1219

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.R-project.org/.

Ramírez-Carrasco G, Martínez-Aguilar K, Alvarez-Venegas R. 2017. Transgenerational defense priming for crop protection against plant pathogens: a hypothesis. Front Plant Sci 8(696). doi: 10.3389/fpls.2017.00696

Rascovan N, Carbonetto B, Perrig D, Díaz M, Canciani W, Abalo M, Alloati J, González-Anta G, Vazquez MP. 2016. Integrated analysis of root microbiomes of soybean and wheat from agricultural fields. Sci Rep 6: 28084. doi: 10.1038/srep28084

Rasmann S, De Vos M, Casteel CL, Tian D, Halitschke R, Sun JY, Agrawal AA, Felton GW, Jander G. 2012. Herbivory in the previous generation primes plants for enhanced insect resistance. Plant Physiol 158(2): 854-863. doi: 10.1104/pp.111.187831

Rasmussen C, Dupont YL, Mosbacher JB, Trøjelsgaard K, Olesen JM. 2013. Strong impact of

temporal resolution on the structure of an ecological network. PloS One 8(12): e81694. doi: 10.1371/journal.pone.0081694

Rastogi G, Coaker GL, Leveau JHJ. 2013. New insights into the structure and function of phyllosphere microbiota through high-throughput molecular approaches. FEMS Microbiol Lett 348: 1-10. doi: 10.1111/1574-6968.12225

Raup DM, Crick RE. 1979. Measurement of faunal similarity in paleontology. J Paleontol 1213-1227.

Ravasz E, Barabási AL. 2003. Hierarchical organization in complex networks. Phys Rev E 67(2): 026112. doi: 10.1103/PhysRevE.67.026112

Rezzonico F, Rupp O, Fahrentrapp J. 2017. Pathogen recognition in compatible plant-microbe interactions. Sci Rep 7(1): 1-2. doi: 10.1038/s41598-017-04792-5

Rico A, McCraw SL, Preston GM. 2011. The metabolic interface between *Pseudomonas syringae* and plant cells. Curr Opin Microbiol 14(1): 31-38. doi: 10.1016/j.mib.2010.12.008

Rivero J, Gamir J, Aroca R, Pozo MJ, Flors V. 2015. Metabolic transition in mycorrhizal tomato roots. Front Microbiol 6: 598. doi: 10.3389/fmicb.2015.00598

Rodríguez H, Fraga R. 1999. Phosphate solubilizing bacteria and their role in plant growth promotion. Biotechnol Adv 17(4–5): 319-339. doi: 10.1016/S0734-9750(99)00014-2

Rojas CM, Senthil-Kumar M, Tzin V, Mysore K. 2014. Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense. Front Plant Sci 5:17. doi: 10.3389/fpls.2014.00017

Rosli HG, Zheng Y, Pombo MA, Zhong S, Bombarely A, Fei Z, Collmer A, Martin GB. 2013. Transcriptomics-based screen for genes induced by flagellin and repressed by pathogen effectors identifies a cell wall-associated kinase involved in plant immunity. Genome Biol 12: 1-5. doi: 10.1186/gb-2013-14-12-r139

Rothschild D, Weissbrod O, Barkan E, Kurilshikov A, Korem T, Zeevi D, Costea PI, Godneva A, Kalka IN, Bar N, et al. 2018. Environment dominates over host genetics in shaping human gut microbiota. Nature 555: 210–215. doi: 10.1038/nature25973

Röttjers L, Faust K. 2018. From hairballs to hypotheses–biological insights from microbial networks. FEMS Microbiol Rev 42(6): 761-80. 10.1093/femsre/fuy030

Röttjers L, Faust K. 2019. Can we predict keystones? Nat Rev Microbiol 17(3): 193. doi: 10.1038/s41579-018-0132-y

Rudrappa T, Czymmek KJ, Paré PW, Bais HP. 2008. Root-secreted malic acid recruits beneficial soil bacteria. Plant Physiol 148: 1547–1556. doi: 10.1104/pp.108.127613

Ruiz-Pérez CA, Restrepo S, Zambrano MM. 2016. Microbial and functional diversity within the phyllosphere of *Espeletia* Species in an Andean high-mountain ecosystem. Appl Environ Microbiol 82: 1807–1817. doi: 10.1128/AEM.02781-15

Samarakoon T, Wang SY, Alford MH. 2013. Enhancing PCR amplification of DNA from recalcitrant plant specimens using a trehalose-based additive. Appl Plant Sci 1(1): 1200236. doi: 10.3732/apps.1200236

Sapp M, Ploch S, Fiore-Donno AM, Bonkowski M, Ros LE. 2018. Protists are an integral part of the *Arabidopsis thaliana* microbiome. Environ Microbiol 20: 30–43. doi: 10.1111/1462-2920.13941

Sasaki E, Zhang P, Atwell S, Meng D, Nordborg M. 2015. "Missing" G x E variation controls flowering time in *Arabidopsis thaliana*. PLoS Genet 11: e1005597. doi: 10.1371/journal.pgen.1005597

Savatin DV, Gramegna G, Modesti V and Cervone F. 2014. Wounding in the plant tissue: the defense of a dangerous passage. Front Plant Sci 5: 470. doi: 10.3389/fpls.2014.00470

Schlaeppi K, Dombrowski N, Oter RG, van Themaat EVL, Schulze-Lefert P. 2014. Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. Proc Natl Acad 111(2): 585-592. doi: 10.1073/pnas.1321597111

Schmidt JE, Kent AD, Brisson VL, Gaudin AC. 2019. Agricultural management and plant selection interactively affect rhizosphere microbial community structure and nitrogen cycling. Microbiome 7(1): 1-18. doi: 10.1186/s40168-019-0756-9

Schoenly K, Cohen JE. 1991. Temporal variation in food web structure: 16 empirical cases. Ecol Monogr 61(3): 267-298. doi: 10.2307/2937109

Schwachtje J, Fischer A, Erban A, Kopka J. 2018. Primed primary metabolism in systemic leaves: a functional systems analysis. Sci Rep 8(1): 1-11. doi: 10.1038/s41598-017-18397-5

Scofield SR, Tobias CM, Rathjen JP, Chang JH, Lavelle DT, Michelmore RW, Staskawicz BJ. 1996. Molecular basis of gene-for-gene specificity in bacterial speck disease of tomato. Science 274(5295): 2063-2065. doi: 10.1126/science.274.5295.2063

Seaver S, Bradbury LM, Frelin O, Zarecki R, Ruppin E, Hanson AD, Henry CS. 2015. Improved evidence-based genome-scale metabolic models for maize leaf, embryo, and endosperm. Front Plant Sci 6: 142. doi: 10.3389/fpls.2015.00142

Seaver SM, Gerdes S, Frelin O, Lerma-Ortiz C, Bradbury LM, Zallot R, Hasnain G, Niehaus TD, El Yacoubi B, Pasternak S, Olson R. 2014. High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. Proc Natl Acad Sci 111(26): 9645-9650. doi: 10.1073/pnas.1401329111

Seaver SM, Lerma-Ortiz C, Conrad N, Mikaili A, Sreedasyam A, Hanson AD, Henry CS. 2018. PlantSEED enables automated annotation and reconstruction of plant primary metabolism with improved compartmentalization and comparative consistency. Plant J 95(6): 1102-1113. doi: 10.1111/tpj.14003

Seifi HS, Van Bockhaven J, Angenon G, Höfte M. 2013. Glutamate metabolism in plant disease and defense: friend or foe? Mol Plant Microbe Interact 26(5): 475-485. doi: 10.1094/MPMI-07-12-0176-CR

Sels J, Mathys J, De Coninck BM, Cammue BP, De Bolle MF. 2008. Plant pathogenesis-related (PR) proteins: a focus on PR peptides. Plant Physiol Biochem 46(11): 941-950. doi: 10.1016/j.plaphy.2008.06.011

Serrano M, Coluccia F, Torres M, L'Haridon F, Métraux JP. 2014. The cuticle and plant defense to pathogens. Front Plant Sci 5: 274. doi: 10.3389/fpls.2014.00274

Shannon CE. 1948. A mathematical theory of communication. Bell system technical journal 27(3): 379-423.

Shi S, Nuccio EE, Shi ZJ, He Z, Zhou J, Firestone MK. 2016. The interconnected rhizosphere: high network complexity dominates rhizosphere assemblages. Ecol Lett 19(8): 926-936. doi: 10.1111/ele.12630

Shimada TL, Takano Y, Shimada T, Fujiwara M, Fukao Y, Mori M, Okazaki Y, Saito K, Sasaki R, Aoki K, Hara-Nishimura I. 2014. Leaf oil body functions as a subcellular factory for the production of a phytoalexin in *Arabidopsis*. Plant Physiol 164(1): 105-118. doi: 10.1104/pp.113.230185

Silva HSA, Romeiro RDS, Mounteer, A. 2003. Development of a root colonization bioassay for rapid screening of rhizobacteria for potential biocontrol agents. J Phytopathol 151(1): 42-46. doi: 10.1046/j.1439-0434.2003.00678.x

Slaughter A, Daniel X, Flors V, Luna E, Hohn B, Mauch-Mani B. 2012. Descendants of primed *Arabidopsis* plants exhibit resistance to biotic stress. Plant Physiol 158(2): 835-843. doi: 10.1104/pp.111.191593

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet 16: 472-482. doi: 10.1038/nrg3962

Tan J, Bednarek P, Liu J, Schneider B, Svatoš A, Hahlbrock K. 2004. Universally occurring phenylpropanoid and species-specific indolic metabolites in infected and uninfected *Arabidopsis thaliana* roots and leaves. Phytochemistry 65(6): 691-699. doi: 10.1016/j.phytochem.2003.12.009

Thiergart T, Durán P, Ellis T, Vannier N, Garrido-Oter R, Kemen E, Roux F, Alonso-Blanco C, Ågren J, Schulze-Lefert P, Hacquard S. 2020. Root microbiota assembly and adaptive differentiation among European *Arabidopsis* populations. Nat Ecol Evol 4(1): 122-131. doi: 10.1038/s41559-019-1063-3

Traw MB, Kniskern JM, Bergelson J. 2007. SAR increases fitness of *Arabidopsis thaliana* in the presence of natural bacterial pathogens. Evolution 61(10): 2444-2449. doi: 10.1111/j.1558-5646.2007.00211.x

Tsuchimatsu T, Goubet PM, Gallina S, Holl AC, Fobis-Loisy I, Bergès H, Marande W, Prat E, Meng D, Long Q, Platzer A. 2017. Patterns of polymorphism at the self-incompatibility locus in 1,083 *Arabidopsis thaliana* genomes. Mol Biol Evol 34(8): 1878-1889. doi: 10.1093/molbev/msx122

Turpin W, Espin-Garcia O, Xu W, Silverberg MS, Kevans D, Smith MI, Guttman DS, Griffiths A, Panaccione R, Otley A, Xu L. 2016. Association of host genome with intestinal microbial composition in a large healthy cohort. Nat Genet 48: 1413–1417. doi: 10.1038/ng.3693

Tylianakis JM, Morris RJ. 2017. Ecological networks across environmental gradients. Annu Rev Ecol Evol Syst 48. doi: 10.1146/annurev-ecolsys-110316-022821

Underwood W. 2012. The plant cell wall: a dynamic barrier against pathogen invasion. Front Plant Sci 3: 85. doi: 10.3389/fpls.2012.00085

Uroz S, Buée M, Deveau A, Mieszkin S, Martin F. 2016. Ecology of the forest microbiome: Highlights of temperate and boreal ecosystems. Soil Biol Biochem 103: 471–488. doi: 10.1016/j.soilbio.2016.09.006

Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S. 2017. Quantitative microbiome profiling links gut community variation to microbial load. Nature 551: 507–511. doi: 10.1038/nature24460

Venturelli OS, Carr AV, Fisher G, Hsu RH, Lau R, Bowen BP, Hromada S, Northen T, Arkin, AP. 2018. Deciphering microbial interactions in synthetic human gut microbiome communities. Mol Syst Biol 14(6): e8157. doi: 10.15252/msb.20178157

Vollmers J, Wiegand S, Kaster AK. 2017. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! PLoS One 12: e0169662. doi: 10.1371/journal.pone.0169662

Vorholt JA. 2012. Microbial life in the phyllosphere. Nat Rev Microbiol 10(12): 828-840. doi: 10.1038/nrmicro2910

Wagner MR, Lundberg DS, Tijana G, Tringe SG, Dangl JL, Mitchell-Olds T. 2016. Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. Nat Commun 7: 12151. doi: 10.1038/ncomms12151

Wang J, Thingholm LB, Skiecevičienė J, Rausch P, Kummen M, Hov JR, Degenhardt F, Heinsen F-A, Rühlemann MC, Szymczak S. 2016. Genome-wide association analysis identifies variation in vitamin D receptor and other host factors influencing the gut microbiota. Nat Genet 48: 1396–1406. doi: 10.1038/ng.3695

Wang M, Roux F, Bartoli C, Huard-Chauveau C, Meyer C, Lee H, Roby D, McPeek MS, Bergelson J. 2018. Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. Proc Natl Acad Sci 115: E5440-E5449. doi: 10.1073/pnas.1710980115

Ward JL, Forcat S, Beckmann M, Bennett M, Miller SJ, Baker JM, Hawkins ND, Vermeer CP, Lu C, Lin W, Truman WM. 2010. The metabolic transition during disease following infection of *Arabidopsis thaliana* by *Pseudomonas syringae* pv. tomato. Plant J 63(3): 443-457. doi: 10.1111/j.1365-313X.2010.04254.x

Wei Z, Yang T, Friman VP, Xu Y, Shen Q, Jousset A. 2015. Trophic network architecture of root-associated bacterial communities determines pathogen invasion and plant health. Nat Commun 6(1): 1-9. doi: 10.1038/ncomms9413

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5(1): 27. doi: 10.1186/s40168-017-0237-y

Westcott SL, Schloss PD. 2015. De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. PeerJ 3: e1487. doi: 10.7717/peerj.1487

Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016. ggplot2-book.org/

Widmer C, Lippert C, Weissbrod O, Fusi N, Kadie C, Davidson R, Listgarten J, Heckerman D. 2014. Further improvements to linear mixed models for genome-wide association studies. Sci Rep 4: 6874. doi: 10.1038/srep06874

Williams RJ, Howe A, Hofmockel KS. 2014. Demonstrating microbial co-occurrence pattern analyses within and between ecosystems. Front Microbiol 5: 358. doi:

10.3389/fmicb.2014.00358

Wu S, Tohge T, Cuadros-Inostroza Á, Tong H, Tenenboim H, Kooke R, Méret M, Keurentjes JB, Nikoloski Z, Fernie AR, Willmitzer L. 2018. Mapping the *Arabidopsis* metabolic landscape by untargeted metabolomics at different environmental conditions. Mol Plant 11: 118–134. doi: 10.1016/j.molp.2017.08.012

Xian WD, Salam N, Li MM, Zhou EM, Yin YR, Liu ZT, Ming YZ, Zhang XT, Wu G, Liu L, Xiao M. 2020. Network-directed efficient isolation of previously uncultivated Chloroflexi and related bacteria in hot spring microbial mats. NPJ Biofilms Microbiomes 6(1): 1-10. doi: 10.1038/s41522-020-0131-4

Xiao Y, Liu H, Wu L, Warburton M, Yan J. 2017. Genome-wide Association Studies in Maize: Praise and Stargaze. Mol Plant 10: 359–374. doi: 10.1016/j.molp.2016.12.008

Xin XF, Kvitko B, He SY. 2018. *Pseudomonas syringae*: what it takes to be a pathogen. Nat Rev Microbiol 16(5): 316. doi: 10.1038/nrmicro.2018.17

Yu X, Lund SP, Scott RA, Greenwald JW, Records AH, Nettleton D, Lindow SE, Gross DC, Beattie GA. 2013. Transcriptional responses of *Pseudomonas syringae* to growth in epiphytic versus apoplastic leaf sites. Proc Natl Acad Sci 110: E425–E434. doi: 10.1073/pnas.1221892110

Yuan H, Cheung CY, Hilbers PA, van Riel NA. 2016. Flux balance analysis of plant metabolism: the effect of biomass composition and model structure on model predictions. Front Plant Sci 7: 537. doi: 10.3389/fpls.2016.00537

Zhang J, Li W, Xiang T, Liu Z, Laluk K, Ding X, Zou Y, Gao M, Zhang X, Chen S, Mengiste T. 2010. Receptor-like cytoplasmic kinases integrate signaling from multiple plant immune receptors and are targeted by a *Pseudomonas syringae* effector. Cell Host Microbe 7(4): 290-301. doi: 10.1016/j.chom.2010.03.007