UNDERSTANDING THE PLANT MICROBIOME

Sur Herrera Paredes

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2017

Approved by:

Jeffery L. Dangl

Corbin D. Jones

Adrian Marchetti

Thomas Mitchell-Olds

Elizabeth Shank

ABSTRACT

Sur Herrera Paredes: Understanding the plant microbiome
(Under the direction of Jeffery L. Dangl)

Plants live in a microbial world and microbes have been known to influence plant health for more than a century. Remarkable progress has been made in elucidating the molecular, physiological and ecological processes in various instances of plant-microbe interactions. This has been possible thanks to a reductionist paradigm that emphasizes testing binary interactions involving only one type of microbe and one type of plant at the same time. In recent years, it has become increasingly clear that plants harbour an enourmous diversity of microbes. These observations raise important questions such as: what is the microbial diversity of the plant associated microbiota? How is the microbial diversity in the plant determined by external factors like soil biodiversity and nutritional composition? What is the role that the plant host plays in structuring the observed microbial biodiversity patterns? What are the plant genes and pathways that modulate the root microbiome and how do those interact with the environment? Finally, what is the function that the plant microbiome performs for the host? How does it influence phenotypic plasticity? and how can we manipulate the plant microbiome to modulate plant phenotypes?

The work described in this dissertation provides some answers to those main questions. We characterized the bacterial diversity in and around Arabidopsis roots, and we showed that the root environment reproducibly selects for a subset of soil taxa, but we also established that the soil type is the second most important factor in explaining the observed communities inside the plant (Chapter 2). We also showed that there is weak but statistically significant effect of plant developmental stage and genotype in the root microbiome (Chapter 2). These results have been reproduced multiple times, in a variety of contexts, and represent the

overarching principles of root microbiome assembly. These principles are reviewed in chapter 1 in light of current data from us and others. While natural variation revealed limited differences in root microbiome, reverse genetics approaches showed stronger effects (Chapters 4 and 5). We used mutant panels in a natural soil to find that the plant phytohormone salicylic acid, which controls a sector of plant immunity, modulates the abundance of specific taxa in the root (Chapter 4). A similar approach, found that an intact phosphate starvation response in Arabidopsis is required to assemble a wild-type root microbiome (Chapter 5). Our studies based on natural soil surveys, while useful, are limited by a lack of genomic context that is inherent to single marker surveys. To overcome this limitation, we pioneered a synthetic community approach by leveraging a large collection of wild root isolates. We have shown that we can use this approach to separate the host and environmental effects on the root microbiome (Chapter 3). We have used this synthetic community approach to delve deeper into the insights obtained in the natural soil surveys. We have shown that there is natural host genetic variation that is associated with the abundance of specific bacterial strains (Chapter 3); that plants deficient for various aspects the salicylic acid pathway can be colonized by bacteria that would be normally excluded, and that salicylic acid exerts its effect on specific strains in a direct manner (Chapter 4); finally, we have shown that a bacterial community can induce the activation of the plant phosphate starvation response, and that the master transcriptional regulator of this response is also a negative regulator of immunity (Chapter 5). Most of the synthetic community work, by us and others, is based on single synthetic communities that try to maximize diversity. While this approach has been successful, it cannot differentiate between correlation and causation, and it limits the questions that can be asked. We have developed experimental designs, and analytical pipelines that allow us to overcome these limitations. By systematically varying the microbial community composition we have shown that we can directly estimate how bacterial groups (Chapter 6) or strains (Chapter 7) will influence plant phenotypes. We can do this from a community context thereby obviating the need for binary association assays. We have shown

that bacterial groups act mostly additively (Chapter 6) and that bacterial strains can act either additively or interactively depending on the plant phenotype (Chapter 7). Finally, we have shown that we can manipulate plant phenotypes by designing novel bacterial consortia (Chapters 6 and 7).

Understanding plant-microbe interactions is essential for plant health and, by extension, for human health. Abating hunger is one of the great unsolved challenges of humanity. Currently, about one in nine people on Earh (∼800 million people) are hungry every day. The consequences of hunger are devastating and long-lasting. A sustainable increase in agricultural productivity is neccessary to reduce hunger and sustain projected population growth over the next century and beyond. The work described here attempts to bring together the best of reductionist and systems-level approaches, and provides key insights into plant microbiome function and manipulation that will impact conservation, mangement and agriculture.

ACKNOWLEDGEMENTS

The work presented here was only possible thanks to the numerous contributions by multiple people. I want to thank the Dangl Lab, both current and former members have created a unique environment for science to thrive. This environment is only possible thanks to the leadership of Prof. Jeff Dangl who inspires everyone to be the best scientist possible. I also want to tank the lab manager Terry Law for efficiently running the day-to-day functioning of the lab. I thank the NLR and Effector branches of the Dangl lab for their continuous interest, critical insight and participation, in particular during my lab seminars. I also thank the microbiome group of the lab, for all the support and willingness to collaborate all these years.

I thank my co-advisors, Prof. Jeff Dangl and Prof. Corbin Jones, for their patience, encouragment, advice and trust through my PhD. The work described here wouldn't have been possible without their committed enthusiasm. I also thank the rest of my thesis committee: Prof. Adrian Marchetti, Prof. Thomas Mitchell-Olds, and Prof. Elizabeth Shank, for dedicating some of their valuable time to me, and for their thoughtful comments and advice.

I also want to thank my scientific collaborators and co-authors from outside the lab, at UNC and elsewhere. A special mention goes to the group at the DOE Joint Genome Institute and the Max Planck Institute at Cologne. They have enriched all my projects with the most valuable critiques.

I thank my PhD progam, the Bioinformatics and Computational Biology Curriculum, for the chance they gave me, and in particular the administrative staff: John Cornett and Cara Marlow, for their smooth support and interest.

I also thank the different institutions that provided financial and material support for

the work described in this dissertation. These include the University of North Carolina at Chapel Hill, and its Graduate School and Department of Biology; the DOE Joint Genome Institute; the National Science Foundation; the Howard Hughes Medical Institute; and the Gordon and Betty Moore Foundation.

Most of all I would like to thank my family, which has always provided unconditional support in all my endeavors; my friends, who have made me a better person; and Mariana, who is my life and inspiration.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Giving back to the community: microbial mechanisms of plant-soil interactions[1]

The role of both plants and soil microbes on ecosystem functioning has been long recognized, but the precise feedback mechanisms between them are more elusive. Definition of these interactions is critical if we aim to achieve an integral understanding of ecosystem functioning, and ultimately explain natural, agricultural and synthetic systems.

Advances in genomic technologies and the development of more appropriate statistical, mathematical and computational frameworks enable researchers to almost fully describe and measure the diversity of microbial communities in soil, rhizosphere and plant tissues. Under the scaffold of community ecology, we integrate the observed patterns of microbial diversity with current mechanistic understanding of plant-microbe mutualistic and pathogenic interactions, and propose a model in which plant microbial communities are shaped by different ecological forces differentially through the plant life cycle.

The same genomic technologies, applied on natural and reconstructed systems, establish that plant genotype has a small, but significant, effect on the microbial community composition in, on and around plant organs. Despite these advances, technical limitations are still important and only a handful of studies exist where a precise genetic element definitively participates in these interactions.

Studies at the field or ecosystem level are dominated by agricultural settings, examining

[1]Most of the content of this chapter has been published before as a peer-reviewed review (Herrera Paredes and Lebeis, 2016). The text has been been lightly edited and updated, and a few minor mistakes have been amended.

microbial species and communities effects on plant productivity; and conversely, that plant genetics and agricultural practices can potentially impose selective pressures on specific microbes and microbial communities.

Revitalized interest in plant-soil microbial feedbacks requires researchers to systematically pose and evaluate more complex hypotheses with increasingly more realistic microbial settings. Despite the advances reviewed here, most studies focus on one aspect of plant, microbe and soil interactions. Experiments that simultaneously and methodically manipulate multiple components are necessary to establish the ecological principles, and molecular mechanisms, which drive microbially mediated plantsoil interactions. This knowledge will be critical to predict how environmental changes affect microbial and plant diversity, and will guide efforts to improve agricultural and conservation practices.

Microbes possess a profound aptitude for altering their environments with their contributions impacting nutrient cycling in their environment from the local to the global scale (Rousk and Bengtson, 2014). Hence, there is strong evidence that multiple microbial factors affect both plant phenotypes and ultimately genotypes through their impact on the environment (Van Nuland et al., 2016). The microbial metabolism responsible for these changes occurs either directly on the environment or within the context of a host, such as a plant (Bulgarelli et al., 2013; Rousk and Bengtson, 2014). The resulting association between plants and microbial communities is often beneficial for the plant, promoting growth and protecting from stress, which is relevant both in the context of natural ecosystems and agricultural settings. While the microbes that colonize above-ground plant organs (i.e. phyllosphere) might be derived from a variety of sources (Vorholt, 2012) below-ground, root microbiomes likely form from the incredibly diverse microbial communities in the surrounding soil (Bulgarelli et al., 2013). Unfortunately, growing evidence suggests that agricultural practices and climate changes will negatively impact soil biodiversity (Wagg et al., 2014), thereby decreasing the types of microbes available for assembly into both the epiphytic and endophytic microenvironments. Here, we review and discuss the current knowledge of plant-soil microbially mediated

interactions, and the impact of improved genomic technologies on our ability to understand how these relationships impact plant performance, potentially allowing us to sustainably improve plant productivity.

1.1 Plant and rhizosphere microbial diversity throughout the plant life cycle

Plants harbour complex microbial communities inside and on every organ; their assembly depends on, among other factors, the microbial species found in the surroundings. While root and leaf communities are readily distinguishable from the surrounding soil and air communities (Bulgarelli et al., 2012; Lundberg et al., 2012; Maignien et al., 2014), differences in microbial composition and diversity exist between plant organs, developmental stages, plant genotypes and environments (Bulgarelli et al., 2013). Such differences can be explained as the result of ecological processes acting on the microbial communities. Under the theoretical framework of community ecology, the factors influencing the composition and diversity of any community can be classified into four general processes: selection, dispersal, drift and speciation (Vellend, 2010; Costello et al., 2012) (Fig. 1.1. While selection and drift decrease diversity, dispersal and speciation increase diversity, and the relative contributions and interactions between these processes determine the final community assembly. Full understanding of plant-soil microbial interactions requires the knowledge of how each of these processes influences the microbial community of both plant and soil, and how these communities affect the same processes on each other over the course of the plants' life cycle and over generations of plants and microbes within an environment.

For the majority of plants, their life cycle begins with a seed, which must be dispersed from its parent. While seed dispersal is an important ecological process for the plants, its role on microbial dissemination is poorly understood. Seeds carry within them associated microbes from their environment and parent of origin, thereby increasing the microbial diversity in their new environment (Fig. 1.1). Both theory and experimental data predict that the more efficient the vertical transmission of a particular microbe, the stronger the tie of that microbe with plant fitness due to negative selection against virulence (Kover

et al., 1997; Stewart et al., 2005; Pagán et al., 2014). Because of this, it should be expected that the microbes carried by seeds be strongly subject to (ecological) selection by the plant, and therefore more likely beneficial. While no study has systematically determined whether seedborne microbes are enriched in plant beneficial functions, there is long-standing evidence for seed-based dispersal of nodule-forming rhizobia in legume seeds (Ash and Allen, 1948). More recent studies suggest the existence of surprising microbial diversity inside the seeds of maize (Johnston-Monje and Raizada, 2011) and spinach (Lopez-Velasco et al., 2013), although they lack functional tests of those microbes. A single study profiling the fungal and bacterial seed epiphytic communities in several species of *Triticum* and *Brassica* also found a largely conserved set of micro-organisms across both genera, hinting at bacteria-fungus antagonism as an important process in determining the microbial community composition (Links et al., 2014). Besides these observational studies, experimental manipulation of seed epiphytes has shown that bacterial seed coatings can protect against pathogens (Wright et al., 2005; Hartmann et al., 2009) and promote plant growth (Jetiyanon et al., 2008). Because microbial seed epiphytes are thought to have an advantage over soil bacteria during plant colonization, seed coating methods for economically important crops are a major area of research and development with numerous patents being filed ($\sim$4000 results for microbial seed coating on Google patent search), and major investment by biotechnology companies (Smith, 2014). Seeds can also harbour bacterial (Gitaitis and Walcott, 2007), fungal (Biswas et al., 2013; Maruthachalam et al., 2013) and oomycete (Testen et al., 2013) pathogens. While it has been proposed that seed dispersal is a general mechanism to escape the high density of pathogens near parents in natural ecosystems (Harms et al., 2000), in the agricultural world seeds act as important vectors for hundreds of diseases, and most studies point to human activities being the major factor in spreading pathogen-bearing seeds (Elmer, 2001). Understanding how seedborne pathogens interact with microbial communities in the plant and soil is an essential step towards better disease control.

Following seed dispersal, during plant germination, readily dispersed microbes might gain

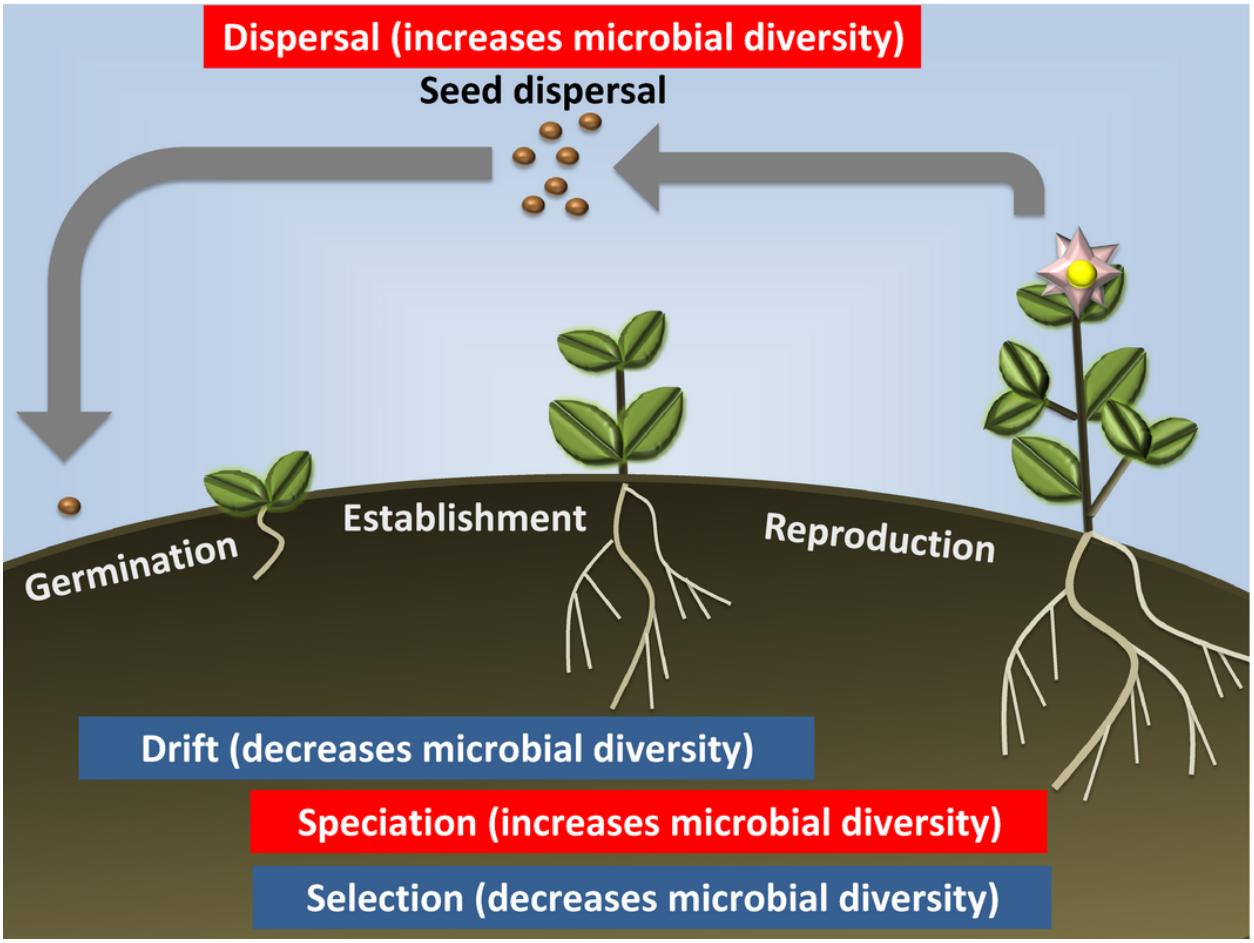


Figure 1.1: **Plant microbiome assembly**. There is evidence for four general ecological processes to occur during the plant life cycle: dispersal, drift, speciation and selection. Microbes hitchhike in and on seeds during dispersal, effectively coupling plant and microbial dissemination and increasing diversity. This process can spread both pathogens and beneficial microbes. During germination and seedling emergence, drift becomes increasingly important; together with selection, they counteract and outweigh the effect of dispersal leading to decreased diversity in plant organs relative to the surrounding soil. The three processes continue to exert the effects in the later stages of plant development, together with microbial speciation, which might occur in any tissue after the initial colonization of communities and coupled with selection, can lead to co-evolution of plants and microbial communities (Van Nuland et al., 2016).

competitive advantage over microbes that attempt to colonize after germination. At the same time, opportunistic microbes from the surrounding soil might gain access to a novel niche as the plant develops. According to this model, early colonization is a highly stochastic process, dominated by dispersal and drift (Fig. 1.1), which leads to 'historically contingent' plant microbial communities where the early colonizers determine the final community, mediated by microbemicrobe interactions, or by plant mechanisms reinforcing the primacy of early colonizers (Costello et al., 2012). Evidence for this model exists in the context of the endophytic compartment of the weedy annual *Arabidopsis thaliana* roots and leaves. Drift may be particularly important given the estimates of a total of 10^5 endophytic bacterial cells per root system (Lundberg et al., 2012) and 10^4 cells cm^2 on the leaf of the same species (Maignien et al., 2014). Given that hundreds of bacterial ribotypes were detected on each organ, these results imply a relatively small population of only tens to hundreds of individuals per ribotype, which greatly favours the influence of drift over other processes, and it is consistent with the observed decrease in microbial diversity with respect to soil (Fig. 1.1). Importantly, the 'historically contingent' model is consistent with the huge individual-to-individual variation in microbial composition found in major sequence-based surveys of the microbiome of roots in diverse species (Bulgarelli et al., 2012; Lundberg et al., 2012; Peiffer et al., 2013; Edwards et al., 2015), and the somewhat smaller but also large variation found in leaves (Redford et al., 2010; Maignien et al., 2014). While there are no experimental or observational data about the microbial communities in plants that follow alternative propagation mechanisms (e.g. rhizomes, spores, stolons, bulbs, tubers, corm or cuttings, as well as horticultural practices such as grafting), the community ecology framework predicts that these plants, which have weaker dispersion, are expected to have reduced microbial diversity (Vellend, 2010). Finally, it is also relevant to define how important the primacy of early colonizers is for plants with annual vs. perennial lifestyles.

While it is well-established that plants can influence the chemical and microbial composition of the rhizosphere, which is the soil area under the root's influence, little is known

about how this effect changes through plant development because most studies focus at later developmental stages, when the root system is firmly established. However, time course experiments in rice determined that the relative abundance of core bacterial taxa from inside the root peaks in the rhizosphere just 3 days after transplantation (Edwards et al., 2015), suggesting that plants may influence the rhizosphere microbial community very early after seedling emergence. Consistently, profiling of the bacterial and fungal communities of seedlings of the Brassicaceae family shortly after emergence shows a decrease in microbial richness, consistent with plant selection and drift happening very early (Barret et al., 2015). Despite these advances, a more systematic evaluation of early time points is still necessary to evaluate the effect of pre-emergence conditions, such as stratification, on microbial communities that colonize very young seedlings. The observation that the composition rhizosphere and rhizoplane bacterial communities are intermediate between that of bulk soil and those that live inside the roots (Bulgarelli et al., 2012; Lundberg et al., 2012; Peiffer et al., 2013; Edwards et al., 2015; Yeoh et al., 2016), has led to the hypothesis that the root microbiome is assembled in a stepwise process where microbes are first recruited to the rhizosphere and then colonize the root (Bulgarelli et al., 2012; Edwards et al., 2015). Under this working model, soil micro-organisms that readily utilize the root exudates would have an advantage in colonizing the roots; however, this model lacks direct empirical testing so far. Another important observation is that soil bacterial composition is the major determinant of the bacterial root microbiome across a variety of plant species (Bulgarelli et al., 2012; Lundberg et al., 2012; Peiffer et al., 2013; Edwards et al., 2015; Yeoh et al., 2016). Interestingly, a similarly strong effect of soil has been reported for phyllosphere bacterial communities (Knief et al., 2010; Zarraonaindia et al., 2015), suggesting that a common environmental pool of microbes exists for both above- and below-ground plant organs.

Sequence-based studies indicate that the relative abundance of bacterial taxa stabilizes quickly, and it is relatively stable in roots (Edwards et al., 2015), but it is unknown whether this steady state is achieved through an isolation of the root microbiome from the surrounding

soil, or through an equilibrium in the exchange rate of microbes between the rhizosphere and the plant. For microbes that are highly abundant or have very efficient dispersal, continual dispersal from the surrounding environment into the plant might counteract the effect of drift, as could selection to maintain them once they are established. While the current standard view is that strong colonizers both invade and persist within the plant host, the turnover rate of the plant microbiota has not been directly measured, and indirect sequence-based methods suggest that it is relatively high in above-ground organs (Redford and Fierer, 2009; Shade et al., 2013). In any case, it is expected that during the 'establishment' phase, when the plant physiology is directed towards increasing plant biomass, the plants would achieve maximum benefit from positive associations. As such, theory predicts that the plant selection over its microbiome is the strongest during this phase (Fig. 1.1). The fact that reproducible enrichment of certain bacterial taxa is commonly found across diverse soils (Bulgarelli et al., 2012); (Lundberg et al., 2012) and in various plant species (Schlaeppi et al., 2014; Yeoh et al., 2016) suggests that plant selection on the microbiome is stronger than ecological drift during this stage, even though it does not completely overtake the 'founders effect' that occurs during seed dispersion (Fig. 1.1).

Most carbon in soils is derived from plants, with individual plants releasing 5-21% of their photosynthetically fixed carbon through the roots (Marschner, 1995), and global carbon release into the rhizosphere in the order of 15002200 kg C ha^{-1} $year^{-1}$ (Kuzyakov and Domanski, 2000). The carbon released is a combination of active secretion of specific root exudates, and passive release of plant debris from both shoots and roots. This process creates a carbon-rich environment in the rhizosphere, while the surrounding bulk soils are considered to be carbon limited (Lambers et al., 2009); as a result, there is a higher density of bacterial cells in this region as compared to the soil, and a distinct bacterial taxonomic profile (Lu et al., 2007; Bulgarelli et al., 2012; Lundberg et al., 2012; Peiffer et al., 2013; Edwards et al., 2015; Yeoh et al., 2016). Furthermore, stable isotope probing data indicate that carbon fixed by the plant via photosynthesis is directly incorporated by specific bacterial taxa in

the rhizosphere (Hernández et al., 2015) and that this assimilation is dependent in close proximity to the root (Lu et al., 2007).

As plants mature, the microbial communities in the root, leaves and rhizosphere may each reach its ecological climax. During this stage, plants reach their maximum photosynthesis rate (Makino et al., 1983) and focus their physiology into the accumulation of biomass. To achieve the maximum growth, plants must have access to enough bioavailable nitrogen and phosphorous. These two elements cannot be readily incorporated by plants from their most abundant sources, resulting in a number of exchange mechanisms between plant and microbes to access each of them. Nitrogen is the most common limiting nutrient in soils, and legumes have evolved a profound interaction with rhizobia (Wang et al., 2012). The exchange of specific molecular signals, plant-secreted flavonoids and rhizobia-secreted Nod factors, initiates a series of molecular responses in the other organism, leading to the formation of specialized organs in the roots called nodules, which accommodate the bacterial symbiont within the plant for nitrogen fixation, in exchange for carbon compounds (Wang et al., 2012). Recently, it has been observed that co-colonization of nodules by a number of other bacteria is under host genetic control (Zgadzaj et al., 2015), though the effect of these microbes on the legumerhizobia symbiosis remains unknown. Other plants can form symbiotic relationships with Actinobacteria or Cyanobacteria through poorly understood molecular mechanisms (Franche et al., 2009). Although the majority of plants cannot form such tight associations with microbes, they might achieve the same success through indirect mechanism, for example through the 'microbial loop', which is a mechanism for plants to exploit predator-prey interactions (reviewed in Bonkowski (2004)). Under this mechanism, the increased density of bacteria in the rhizosphere stimulates the activity of bacterial grazers such as protozoa and nematodes (Clarholm, 1985). Bacterial grazing then contributes with the excretion of one a large portion of the ingested nitrogen as ammonia, which can directly incorporated to plant or nitrified to nitrate by other bacteria before plant incorporation (Bonkowski, 2004; Lambers et al., 2009).

After nitrogen, phosphorous is the most common limiting nutrient in soils (Schachtman et al., 1998). Most of the phosphorous in soil is in insoluble phosphate forms that cannot be used by plants. The vast majority of plants have overcome this limitation by evolving a mycorrhizal interaction. The most prevalent of these plant-fungus interactions is with arbuscular mycorrhizal fungi (AMF), which is estimated to interact with 80% of land plant species (Brundrett, 2009) and is proposed to have played a key role in land colonization by plants (Buscot, 2015). Plants recruit AMFs by secreting compounds, such as strigolactones, that induce spore germination and hyphae formation (Schmitz and Harrison, 2014). The AMF then forms a network of hyphae that is directly connected to the plant root and extends the reach and functional capacity of roots. The fungal partner solubilizes phosphate and then delivers it to the plant root, which in turn provides the fungus with a constant supply of carbon compounds (Smith and Smith, 2012). The AMF hyphae not only extend the root capacity, but may also extend the rhizosphere effect. Interestingly, AMF harbours their own bacterial partners (Naumann et al., 2010; Desirò et al., 2014), although their effect on the root and soil bacterial communities, and on the carbon-phosphorous trade, has not been measured. Besides mycorrhizal fungi, many bacteria can also solubilize phosphate (Rodríguez and Fraga, 1999), and it has been reported that, among cultivable bacteria, there is a higher proportion of phosphate-solubilizing bacteria in bulk soil than in plant tissue (Marasco et al., 2012). However, the importance of this process in the field is poorly understood, and inoculation of soils with phosphate-solubilizing bacteria has produced negligible differences in plant phosphate assimilation (Glick, 2012).

As plants age and enter the reproductive phase, there are substantial changes in metabolism and physiology that redirect carbon flux from the accumulation of biomass, and towards the production of reproductive organs during the sink-to-source transition (Jeong, 2004). However, in the fast-growing annual *A. thaliana*, little difference in root bacterial community was noted at two very different developmental states, before and well after the metabolic switch in carbon allocation (Lundberg et al., 2012). On the other hand, a finer time-scale

demonstrated that different bacterial taxa preferentially colonize the apple flower at different developmental stages (Shade et al., 2013), suggesting that plant development may alter the selective mechanisms driving microbial succession. Further evidence for this has been found in a multiyear study of the leaf microbiome of deciduous trees where leaf age contributes more to community composition than experimental year (Redford and Fierer, 2009). It is unknown how the differences in annual vs. perennial life histories influence the assembly and long-term stability of plant microbiota. To fully elucidate how the order of microbial colonization affects the plant microbiome, it would be necessary to carry out studies with time-series and crossover designs; this type of design has already been used to establish the existence of such 'order effects' in the context of colonization of the mammalian gut (Lee et al., 2013).

Finally, it is important to consider the possibility of co-evolution between soil and plant microbiota. Very little is known about the evolution of host-associated microbial communities, but recently, a neutral model that incorporates microbial acquisition from the environment and from vertical transmission under the Wright-Fisher genealogical model for hosts was developed (Zeng et al., 2015). A prediction from this model is that the least diverse microbiomes evolve from strong vertical transmission, but only a modest level of environmental contribution is required to generate high alpha diversity. While such neutral model is a useful baseline for comparison, more sophisticated models that incorporate non-neutral processes such as selection and speciation are required. It should also be appreciated that from the bacterial perspective, host colonization is a microevolutionary process, and so speciation and the processes behind it, like allopatry and resource partitioning, need to be considered. In this regard, other approaches focus on microbial dynamics and have exploited the generalized Lotka-Volterra system to identify conditions that favour community stability (Stein et al., 2013; Coyte et al., 2015). Ultimately, any theoretical framework that attempts to explain plant-soil microbially mediated feedbacks must incorporate the co-evolution of the soil, rhizosphere and host microbial communities instead of solely examining the host or microbial

perspective (Van Nuland et al., 2016).

1.2 The genomic basis of plantmicrobe interactions

To thrive in the plant tissue, a micro-organism must have the genetic determinants to access and invade at least one plant tissue and then, persist in the presence of a sophisticated immune system and a chemical composition distinct from the surrounding soil. Thus, it is expected that both plant and microbial genomes show evolutionary signatures relating to these interactions. Indeed, studies of *A. thaliana* (Bulgarelli et al., 2012; Lundberg et al., 2012) and maize (Peiffer et al., 2013) have shown a significant, if small, effect of the plant natural genotypes on the root microbiome with a stronger effect reported among barley cultivars (Bulgarelli et al., 2015). Moreover, it has been reported that there is a correlation between the phylogenetic distance and root microbiome dissimilarity in plants of the Brassicaceae (Schlaeppi et al., 2014) and Poaceae (Bouffaud et al., 2014) families. There is also evidence for plant genetic effects on the phyllosphere (i.e. above-ground) community. Poplar fungal leaf microbiome correlates with plant genotype in common garden experiments (Bálint et al., 2013), and a synthetic community approach with *A. thaliana* plants showed differences between accessions and comparison of mutants to wild-type plants pointed at a role for cuticle formation and ethylene signalling in shaping the phyllosphere microbiome (Bodenhausen et al., 2013, 2014) and salicylic acid in root microbiome (Lebeis et al., 2015). Finally, a genome-wide association study in *A. thaliana* of fungal and bacterial leaf microbiome pointed at a number of plant *loci* that affect abundance of specific microbes and species richness; defense was the most common process associated with bacterial abundance but other processes such as cell wall integrity, trichome branching and morphogenesis also affected the microbiome (Horton et al., 2014).

The emerging picture from the majority of studies is that plant *loci* have small and variable effects on the microbiome composition. A limitation of all of these studies is that they rely on profiling of a single marker gene to define the taxonomic composition of the plant microbiome, which means that these studies ignore the possibility that plants select

at the functional, as opposed to taxonomic, level, especially if selection occurs primarily via exudation of compounds that stimulate specific microbial metabolic activities. In fact, it has been shown that bacterial strains that have the same 16S rRNA gene sequence can induce very different plant phenotypes (Blakney and Patten, 2011; Haney et al., 2015; Timm et al., 2015). Importantly, these would be analogous to the observation in that the human gut microbiome has a remarkably stable functional profile despite the huge variation at the taxonomic level (Huttenhower et al., 2012). Equally important would be to extend the study of the role of plant genetic variation on the microbiome beyond the few model organisms that have been used so far. To test this hypothesis, it would be necessary to perform shotgun metagenome sequence on plant-associated microbial communities; however, the complexity of soil and technical difficulties in separating microbial- and plant-derived DNA from plant tissues have so far limited our ability to query the functional content and diversity of plant and rhizosphere microbial communities. Novel computational and experimental methods have been recently developed (Feehery et al., 2013; Howe et al., 2014) that may identify the microbial functions required for plant colonization. Despite these technical limitations, the rhizosphere metagenome has been compared between cucumber and wheat (Ofek-Lalzar et al., 2014), as well as among barley cultivars (Bulgarelli et al., 2015). Each of these studies found a signature of enriched bacterial functions in the rhizosphere although no overlap was seen between studies, possibly due to technical differences. An alternative approach to shotgun metagenomics is comparative genomics, which was used to determine that Pseudomonas isolates from different geographic regions are nearly isogenic to well-characterized beneficial bacteria, raising the possibility that their dispersion has been selected by the plants (Berendsen et al., 2015). Comparative genomics has also been used to investigate the phylogenetic distribution of bacterial genes that confer plant beneficial functions among Proteobacteria. The observed phylogenetic distributions demonstrated that plant beneficial bacteria commonly contain multiple beneficial genes, though there is no core set of plant beneficial genes, suggesting that these genes might be selected in

plant-associated habitats and counterselected elsewhere (Bruto et al., 2014). While most comparative genomics approaches have focused on relatively narrow and well-defined bacterial clades with previously characterized functions, recent efforts to systematically sample the genomic diversity of plant-derived isolates (Bai et al., 2015) allow the differentiation between bacterial functions required to thrive in the plant environment, and bacterial functions that the plant may select because they provide a fitness advantage to the plant.

Studies conducted on the human gut microbiome linking disease states with different bacterial metabolic topologies (Greenblum et al., 2012) suggest that microbe-microbe metabolic exchanges play a key role in structuring host-associated microbial communities. In the context of plants, correlations between bacterial, fungal and oomycete abundance were used to identify the potential keystone microbial species that drive interkingdom community assembly (Agler et al., 2016). Additional experiments are necessary to extend these results to the context of the plant root and rhizosphere, and to demonstrate causality; in particular, microcosm reconstitution experiments with complex, but well-defined, synthetic microbial communities where all the partners are well defined and tractable in isolation harness the power of reductionist science in a realistic setting. Importantly, this approach has successfully dissected the contribution of plant signalling pathways to both leaf and root colonization by bacteria (Bodenhausen et al., 2014; Lebeis et al., 2015). A complementary approach can leverage the extant publicly available bacterial genomes to perform genome-wide metabolic reconstruction. While this approach has not been directly applied to plant-associated communities, metabolic reconstruction and modelling has been used to show a high potential for the emergence of biosynthetic capacity in mixed cultures (Chiu et al., 2014), as well as a large number of potential metabolite exchanges among naturally co-occurring groups of bacteria (Zelezniak et al., 2015). Furthermore, systematic *in vitro* co-culturing of auxotroph pairs has shown a large number of syntrophic interactions, which were supported by genome bacterial genome mining (Mee et al., 2014; Embree et al., 2015). Metabolic modelling approaches depend on fully sequenced genomes and rely heavily on high-quality annotations. Thus,

efforts to expand the set of reference bacterial genomes isolated from plant and rhizosphere samples, such as the study from Bai et al. (2015), are essential building blocks. Improved annotations taking into account the ecological context are also required for modern genomic techniques like transposon insertion sequencing (Goodman et al., 2009) and artificial evolution (Schlötterer et al., 2015). In the long run, these approaches will feed statistical and population genetics models that promise to predict plant phenotypes as outputs of interactions between plants and microbial communities.

1.3 Impacts on plant performance

While pathogenic microbes decrease plant performance, plants also experience positive microbial influences on their productivity by increasing growth or by helping plants to cope with stress (Schnitzer et al., 2011). Hence, some microbes can produce plant growth-promoting phytohormones, such as indole-3-acetic acid (IAA), as well as can mediate acquisition by the plant of nitrogen, phosphate, iron and nitrogen (Knief et al., 2010; Ofek-Lalzar et al., 2014; Sessitsch et al., 2012). Bacteria that perform one or — more commonly (Bruto et al., 2014) — many of these functions in the root are categorized as plant growth-promoting rhizobacteria (PGPR).

Microbes also promote plant performance indirectly by protecting against both abiotic stress and disease (Bulgarelli et al., 2013). In addition to the advantages microbial services provide in low nutrient environments, drought is eased by bacteria producing 1-aminocyclopropane-1-carboxylate (ACC) deaminase (Marasco et al., 2012), which reduces ethylene concentrations under stress conditions, by helping plants during drought stress (Cao et al., 2007). Protective micro-organisms in the roots may also prevent infection via immune priming (Pozo and Azcón-Aguilar, 2007; Zamioudis and Pieterse, 2012). Beneficial root bacteria produce induced systemic resistance (ISR), while AMF can produce mycorrhizal-induced resistance (MIR) (Pozo and Azcón-Aguilar, 2007; Zamioudis and Pieterse, 2012). ISR is achieved via jasmonic acid and ethylene signalling, and it is distinct from another form of systemic resistance, namely systemic acquired resistance (SAR), which is induced by leaf

pathogens and mediated by salicylic acid (Conrath et al., 2006). MIR shares some characteristics with both ISR and SAR, and while the standard view is that fungal stimulation is directly responsible for induced resistance, it has been hypothesized that MIR is a cumulative effect of plant responses to mycorrhizal infection and ISR-inducing rhizobacteria (Pozo and Azcón-Aguilar, 2007; Zamioudis and Pieterse, 2012). Some rhizobacteria are capable of both plant growth-promoting activity and ISR induction. For example, *Pseudomonas fluorescens* strain WCS417 promotes growth mediated by IAA production and ACC deaminase activity, and ISR via jasmonic acid signalling (Schwachtje et al., 2012; Zamioudis et al., 2013).

Differential bacterial colonization of varying plant genotypes can occur at the community level or within a single microbial species. The latter is certainly the case with *P. fluorescens* strains (Haney et al., 2015), in which different ecotypes of *A. thaliana* support different levels of colonization by various strains differing in their ability to promote plant growth and protect against pathogens. Lower colonization did not correlate with higher defense response gene expression, but instead appeared to be related to some other incompatibility. Concordantly, the normal growth promotion and pathogen protection did not occur in ecotypes with the decreased levels of colonization (Haney et al., 2015). More recently, a genetic approach has shown that the plant defense hormone salicylic acid affects the abundance of specific bacterial groups in the root at a high taxonomic level via a combination of direct and indirect effects (Lebeis et al., 2015); importantly, overproduction of salicylic acid leads to the decreased biomass accumulation in plants (Bowling, 1994) Overall, these results suggested the existence of complex fitness trade-offs where the result of the plant-bacteria interaction is determined by the specific combination of plant accession, bacterial strain and plant pathogen in the environment.

Influence over plant growth may not be influenced by individual microbes, but may also be a community-level phenotype. Artificial selection experiments achieved increased plant biomass by repeatedly selecting soil microbial communities (Swenson et al., 2000; Panke-Buisse et al., 2015). As our understanding of plant-microbe partnerships improves, co-evolutionary

hypotheses between plants and microbial environments become evident; in particular, it is important to understand how plant domestication has impacted the ability of plants to form microbial partnerships. Because plant domestication leads to a loss of diversity of the *loci* under selection, and those adjacent to them, a possible consequence is the loss of traits that were not directly under artificial selection; for this reason, it has been hypothesized that domestication has reduced the ability of plants to form beneficial associations with rhizosphere microbes (Pérez-Jaramillo et al., 2016). Indeed, recent studies have found that there are specific, but not overlapping, differences between wild and domesticated root microbiomes of both lettuce and barley (Bulgarelli et al., 2015; Cardinale et al., 2015). Specifically, compared to wild barley, domesticated barley grown in a common soil had increased relative abundance of the bacterial classes Alphaproteobacteria and Betaproteobacteria (Bulgarelli et al., 2015), which contain a number of taxa known to affect plant health, such as rhizobia. The mechanisms behind these changes might involve the microbial genes found in the core set of root micro-organisms. Thus, using shotgun metagenome sequencing of barley rhizosphere communities, it was discovered that bacterial genes related to their interactions with both plant and phage were under positive selection, promoting secretion (e.g. type 3 secretion systems), nutrient acquisition (e.g. siderophores) and stress tolerance (e.g. detoxification) (Bulgarelli et al., 2015). These results are strikingly similar to those from a metagenomic study performed on rice rhizospheres (Sessitsch et al., 2012), as well as anecdotal evidence for genes found in individual PGPR Pseudomonas strains (Berendsen et al., 2015). Together, these observations suggest that plant beneficial traits are repeatedly selected by the plants and/or indirectly by farmers and breeders during domestication.

While numerous agricultural practices could provide selective pressures that lead to differential plant microbiomes between wild and domestic crops, recent studies have highlighted that simply growing plants in monoculture instead of mixed fields significantly contributes to microbiome composition, significantly decreasing microbial biodiversity (Zuppinger-Dingley et al., 2014). Conversely, higher microbial diversity is correlated with increased plant height

and leaf area (Zuppinger-Dingley et al., 2014). Negative impacts of plant monoculture in fields on microbial biodiversity might be influenced by an accumulation of plant-specific beneficial and pathogenic microbes. While no studies have directly demonstrated whether pathogens or beneficial microbes accumulate more rapidly, a recent study with tobacco grown in a native soil demonstrated the accumulation of both within a decade of field establishment (Santhanam et al., 2015). It is possible that diversity plays a similar role in maintaining a healthy plant microbiome, but systematically controlling and varying diversity in microcosm reconstitution experiments is required to fully distinguish between cause and effect. Thus, conspecific fields have decreased microbial diversity with a correlating increase in diseased plants (Schnitzer et al., 2011). Indeed, with increasing plant diversity from 1 to 15 species, there is a decrease in non-mycorrhizal infection, while beneficial mycorrhizal infection remains constant (Schnitzer et al., 2011). Together these indicate that higher microbial species diversity decreases the plant-pathogen interactions leading to improved plant growth.

1.4 Conclusion

From an ecological perspective, the health of a community can be viewed as its ability to withstand and recover from perturbations, and low bacterial diversity in the mammalian gut has been associated with susceptibility to perturbation (Virgin and Todd, 2011) and disease (Turnbaugh et al., 2009b). Recent studies have begun to paint a picture for how the dynamics of plant microbiomes are controlled and impacted by various factors. It is vital that we understand these processes in order to effectively implement them potentially in management and agricultural practices.

CHAPTER 2

Defining the core Arabidopsis thaliana root microbiome[1]

Land plants associate with a root microbiota distinct from the complex microbial community present in surrounding soil. The microbiota colonizing the rhizosphere (immediately surroundingthe root) and the endophytic compartment (within the root) contribute to plant growth, productivity, carbon sequestration and phytoremediation (Rodriguez et al., 2008; De Deyn et al., 2008; van der Lelie et al., 2009). Colonization of the root occurs despite a sophisticated plant immune system (Jones and Dangl, 2006; Dodds and Rathjen, 2010), suggesting finely tuned discrimination of mutualists and commensals from pathogens. Genetic principles governing the derivation of host-specific endophyte communities from soil communities are poorly understood. Here we report the pyrosequencing of the bacterial 16S ribosomal RNA gene of more than 600 *Arabidopsis thaliana* plants to test the hypotheses that the root rhizosphere and endophytic compartment microbiota of plants grown under controlled conditions in natural soils are sufficiently dependent on the host to remain consistent across different soil types and developmental stages, and sufficiently dependent on host genotype to vary between inbred *Arabidopsis* accessions. We describe different bacterial communities in two geochemically distinct bulk soils and in rhizosphere and endophytic compartments prepared from roots grown in these soils. The communities in each compartment are strongly influenced

[1]Most of the content of this chaper has been published before as a peer-reviewed article (Lundberg et al., 2012). The text has been been lightly edited and re-arranged to facilitate reading. The figure order has been changed to match the updated text order. Section and subsection headers have been added for easier navigation. Several minor mistakes have been amended. Numerous supplementary files were made available online at the time of publication, and are not included here; they will be referred to as Supplementary Table or Supplementary Dataset and can be obtained at `http://www.nature.com/nature/journal/v488/n7409/full/nature11237.html#supplementary-information`.

by soil type. Endophytic compartments from both soils feature overlapping, low-complexity communities that are markedly enriched in Actinobacteria and specific families from other phyla, notably Proteobacteria. Some bacteria vary quantitatively between plants of different developmental stage and genotype. Our rigorous definition of an endophytic compartment microbiome should facilitate controlled dissection of plantmicrobe interactions derived from complex soil communities.

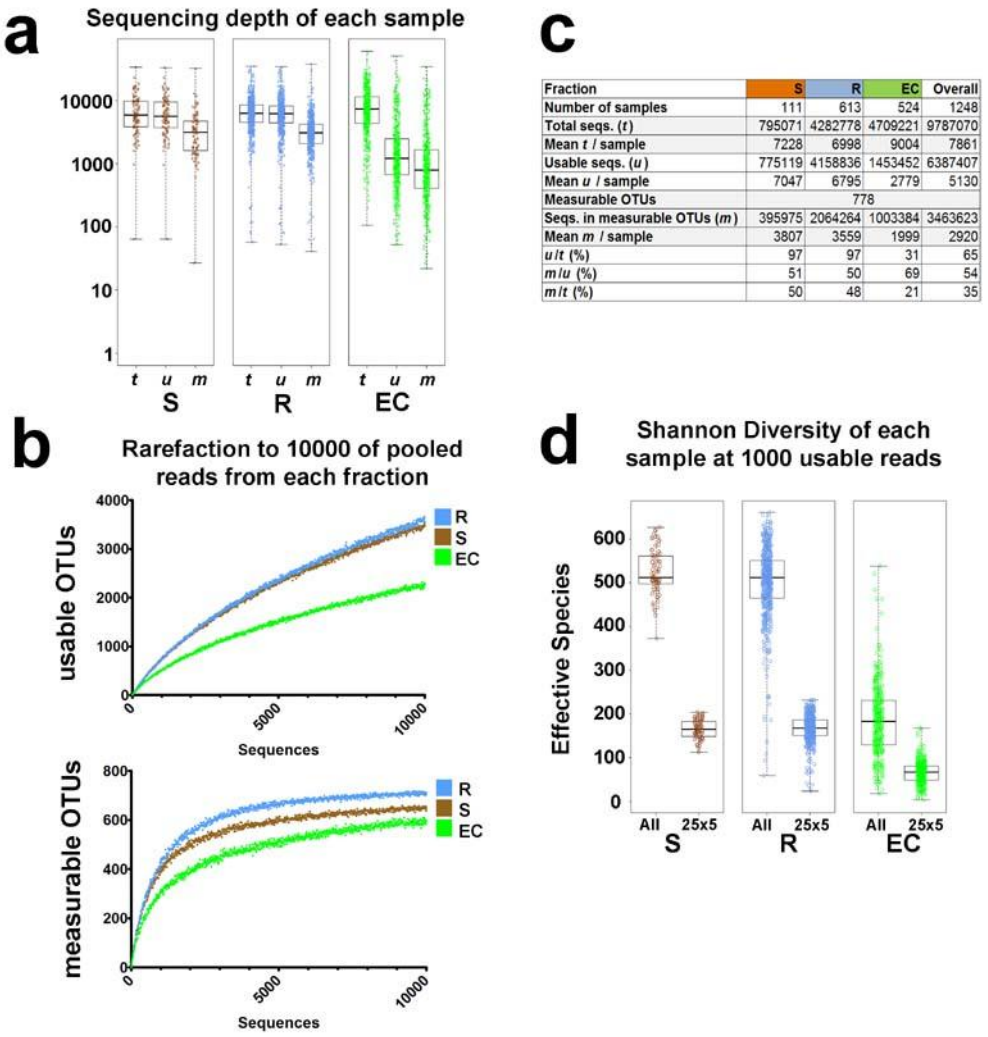
Roots influence the rhizosphere by altering soil pH, soil structure, oxygen availability, antimicrobial concentration, and quorum-sensing mimicry, and by providing an energy source of dead root material and carbon-rich exudates (Marschner et al., 1986; Dennis et al., 2010). The microbiota inhabiting this niche can both benefit and undermine plant health; shifting this balance is of agronomic interest. Mutualistic microbes may provide the plant with physiologically accessible nutrients and phytohormones that improve plant growth, may suppress phytopathogens or may help plants withstand heat, salt and drought (Mendes et al., 2011; Firáková et al., 2007). The rhizosphere community is a subset of soil microbes that are subsequently filtered via niche utilization attributes and interactions with the host to inhabit the endophytic compartment (EC) (Schulz et al., 2006). Although a variety of microbes may enter and become transient endophytes, those consistently found inside roots are candidate symbionts or stealthy pathogens (Schulz et al., 2006; Hallmann et al., 1997). Notably, *Arabidopsis* and other Brassicaceae are not well colonized by arbuscular mycorrhizal fungi, implying that other microorganisms may fill this niche.

Microbial community structure differs across plant species (Redford et al., 2010; Hardoim et al., 2008), and there are reports of host-genotype-dependent differences in patterns of microbial associations (Inceolu et al., 2010; Inceolu et al., 2011). However, the divergent methods used in those studies relied on small sample sizes and low-resolution phylotyping techniques potentially confounded by off-target sequences and chimaeric amplicons. We developed a robust experimental system to sample repeatedly the root microbiome using high-throughput sequencing. Our results confirm many of the general conclusions from earlier

studies and, because of controlled experimental design and the power of deep sequencing, provide a key step towards the definition of this microbiomes functional capacity and the host genes that potentially contribute to microbial association phenotypes. Such plant genes would constitute major agronomic targets.

We used 454 pyrosequencing to sequence 16S ribosomal RNA (rRNA) gene amplicons for DNA prepared from eight diverse, inbred *A. thaliana* accessions. Plants were grown fromsurface-sterile seeds in climate-controlled conditions in two diverse soils, respectively termed Mason Farm and Clayton (Supplementary Table 1; detailed in methods 2.1.2). For each soil, we assayed multiple individuals from each *A. thaliana* accession grown from sterile seeds in both soils across independent full-factorial biological replicates, in which all genotypes and bulk soils (pots without a plant) for a given soil type were grown in parallel (Supplementary Table 2). We isolated separate rhizosphere and EC fractions from individual plant root systems (Fig. 2.13 and Supplementary Table 2). We established 1114F and 1392R as our primer pair (methods 2.1.9 and 2.1.10; Fig. 2.14). Using an otupipe-based pipeline (`http://drive5.com/otupipe/`), we grouped sequences into 97%-identical operational taxonomic units (OTUs), reduced noise and removed chimaeras (methods 2.1.11). We determined technical reproducibility thresholds to conclude that OTUs defined by ≥ 25 reads in ≥ 5 samples (hereafter *25x5*) are individually *measurable* OTUs (Benson et al., 2010; Gottel et al., 2011) (Figs 2.18 and 2.15; methods 2.1.11). All data reported here are from one run of our otupipe-based pipeline (Fig. 2.16 and Supplementary Database 1).

Excluding additional control samples, we ribotyped 1,248 samples comprising 111 bulk soil, 613 rhizosphere and 524 EC samples, generating 9,787,070 high-quality reads (Figs. 2.16 and 2.1ac). After removing plant-sequence-derived OTUs, we obtained a table of *usable* OTU read counts per sample containing 6,387,407 reads distributed across 18,783 OTUs. We normalized this table of usable reads by rarefying to 1,000 reads per sample (Supplementary Database 2a) or, alternatively, by dividing the reads per OTU in a sample by the sum of *usable* reads in that sample, resulting in a table of relative abundances (frequencies) (Supplementary



Fraction	S	R	EC	Overall
Number of samples	111	613	524	1248
Total seqs. (*t*)	795071	4282778	4709221	9787070
Mean *t* / sample	7228	6998	9004	7861
Usable seqs. (*u*)	775119	4158836	1453452	6387407
Mean *u* / sample	7047	6795	2779	5130
Measurable OTUs	778			
Seqs. in measurable OTUs (*m*)	395975	2064264	1003384	3463623
Mean *m* / sample	3807	3559	1999	2920
u/*t* (%)	97	97	31	65
m/*u* (%)	51	50	69	54
m/*t* (%)	50	48	21	35

Figure 2.1: **Sequencing statistics and quality**. **a)** Sequencing depth per sample in reads for the three sample fractions S, R, and EC. Each dot represents a single plant or soil sample. Within each fraction, the total (t), *usable* (u), and *measurable* (m) read counts are shown for all samples. The box plots contain the 1st and 3rd quartiles, split by the median; whiskers extend to include the farthest points. **b)** Rarefaction curves to 10,000 sequences for cumulative reads from S, R, and EC fractions considering all usable OTUs (top) and only measurable OTUs (bottom). **c)** Table summarizing the total and usable *usable* reads per sample fraction, as well as the number and proprtion of total and *usable* reads that fall within *measurable* OTUs. d) Shannon diversity of individual samples from each fraction, calculated from the rarefaction-normalized table, before (left) and after (right) applying the *measurable* OTU threshold.

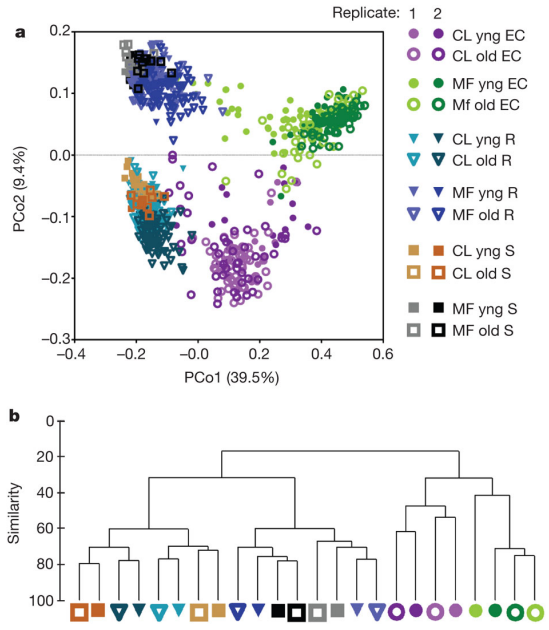


Figure 2.2: **Sample fraction and soil type drive the microbial composition of root-associated endophyte communities**. **a)** Principal Coordinate Analysis of pairwise, normalized, weighted UniFrac distances between samples based on rarefaction to 1,000 reads in *usable* OTUs. CL, Clayton; MF, Mason Farm; R, rhizosphere; S, soil. **b)**, Hierarchical clustering (group-avegage linkage) of the $\log_2$-transformed rarefied counts from the *measurable* OTUs. Based on the pairwise BrayCurtis dissimilarity.

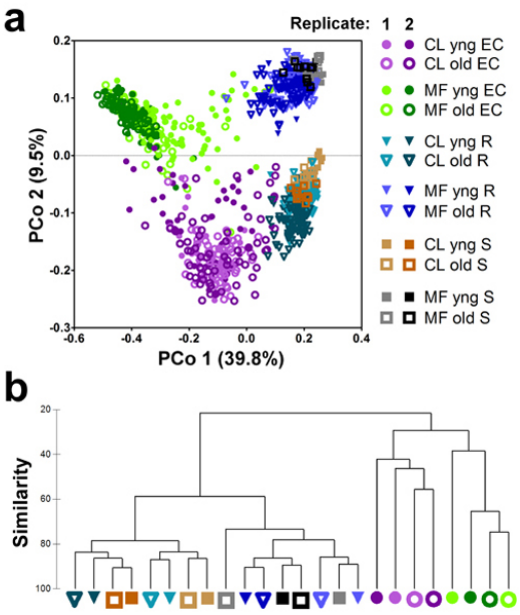


Figure 2.3: **Sample fraction and soil type drive the microbial composition of root-associated endophyte communities**. **a)** Principal Coordinate Analysis of pairwise, normalized, weighted UniFrac distances between samples based on relative abundances of *usable* OTUs. CL, Clayton; MF, Mason Farm; R, rhizosphere; S, soil. **b)**, Hierarchical clustering (group-avegage linkage) of the $\log_2$-transformed relative abundances from the *measurable* OTUs. Based on the pairwise BrayCurtis dissimilarity.

Database 2b). Using the *25x5* threshold, we defined 778 *measurable* OTUs representing 54% (3,463,632) of the usable reads (Fig. 2.1c and Supplementary Table 3). The diversity of the 778 measurable OTUs in soil, rhizosphere and EC fractions showed expected relative trends when compared with the diversity by fraction of all usable OTUs (Fig. 2.1d). We display parallel analyses of the rarefaction-normalized and frequency-normalized data, while in the text we use the numbers from the rarefied data.

We used principal coordinate analysis on pairwise, normalized, weighted UniFrac distances between all samples, considering all *usable* OTUs, to identify the main factors driving community composition (Figs. 2.2a and 2.3a). The first principal coordinate (PCo1) revealed

that the two bulk soils and their associated rhizospheres were differentiated from the respective EC fractions. Soil type was the main factor in the second component (PCo2). This pattern was recapitulated by hierarchical clustering of pairwise BrayCurtis dissimilarities considering only measurable OTUs (Figs. 2.2b and 2.3b). Samples harvested at different developmental stages clustered together, indicating that this variable does not have a major effect on overall community composition (Figs. 2.2 and 2.3; yng versus old, where yng refers to the time of appearance of an inflorescence meristem and old refers to fruiting plants with greater than 50% senescent leaves). Additional control samples from the reference genotype Col-0 harvested from four independent digs of Mason Farm soil underscored the reproducibility of these bacterial community profiles (Fig. 2.4). Together, these data demonstrate that the interaction of diverse soil communities with plants determines the assembly of the rhizosphere, leading to winnowed ECs, that the ECs from at least these two diverse soils are very different from the starting soil communities and that there is little difference in communities over host developmental time.

We fitted a general linear mixed model (GLMM) to samples from each set of plant fractions (rhizosphere or EC), plus the bulk soil controls, to identify measurable OTUs whose abundances differ significantly between plant and bulk soil as a result of soil type, developmental stage, fraction and genotype (methods 2.1.12 and Supplementary Database 3). This approach allowed us to quantify the contribution from each variable to the community composition (Supplementary Table 4). Controlling for sequencing plate effects, plant fraction is the most important factor; its effect is strongest for the EC, consistent with our UniFrac and BrayCurtis analyses. Soil type is less important, followed by experiment, developmental stage and, finally, genotype, which had a small but consistent effect.

Hierarchical clustering based on abundances from the 256 OTUs identified by the GLMM to differentiate rhizosphere and EC from soil recapitulated the separation of EC from soil and rhizosphere (Figs. 2.5A and 2.6a, left; compare with Figs. 2.2 and 2.3). Of these, 164 OTUs were enriched in EC samples (Figs. 2.5Ba and 2.6ba; dark and light red bars), defining an

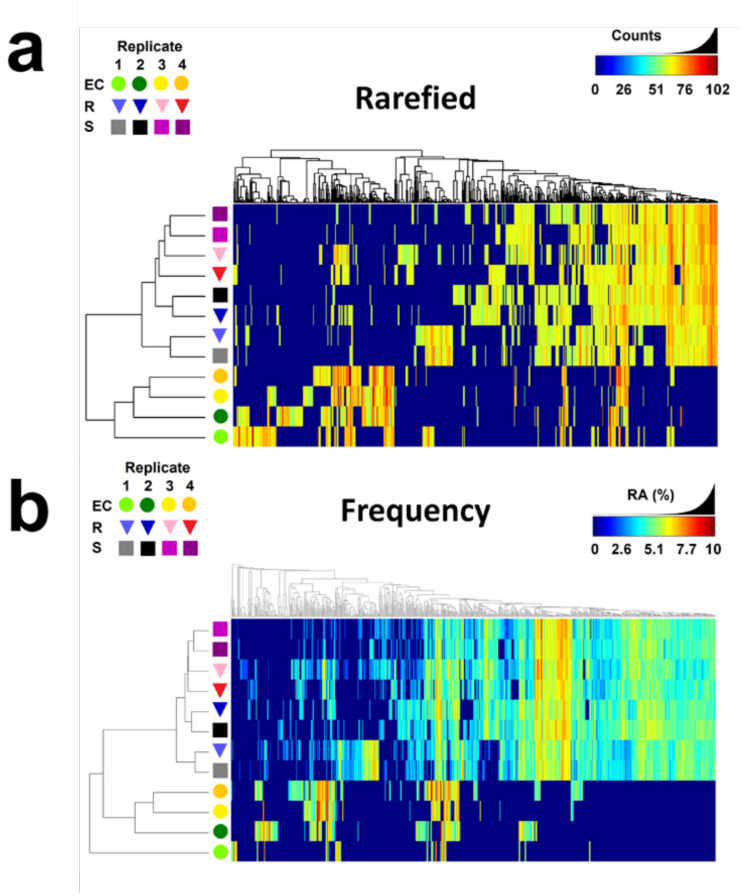


Figure 2.4: **OTUs identified from four independent biological replicates are reproducible**. Heat map displaying the reproducibility between four independent replicates at the yng developmental stage of bulk soil (squares), Col-0 R samples (triangles), and Col-0 EC samples (circles). Each symbol represents the median of six or more samples. All data were $\log_2$-transformed for visualization. The quantities in the color key represent the original (untransformed) counts (top) and frequencies (bottom) for each color. OTUs that had a median of 0 in all Col-0 and soil groups shown and were removed from the display.

A. thaliana 'EC microbiome'. Of these 164, 97 were enriched in EC samples from both soil types (Figs. 2.5Ba and 2.6ba; dark red bars), potentially representing a core EC microbiome. By contrast, 67 of these 164 were enriched in EC to a greater extent in one soil than the other (light red bars in Figs. 2.5Ba and 2.6ba; and gold and brown bars in Figs. 2.5Bb and 2.6bb). Importantly, 32 OTUs were depleted in EC samples (Figs. 2.5Ba and 2.6ba; blue bars). Some OTUs exhibited rhizosphere enrichment; these significantly overlapped the EC-enriched OTUs ($P < 10^{-16}$; one-sided hypergeometric test) and also sometimes had a soil-type component (Figs. 2.5Bc-d and 2.6bc-d). Only a few rhizosphere-specific enrichments

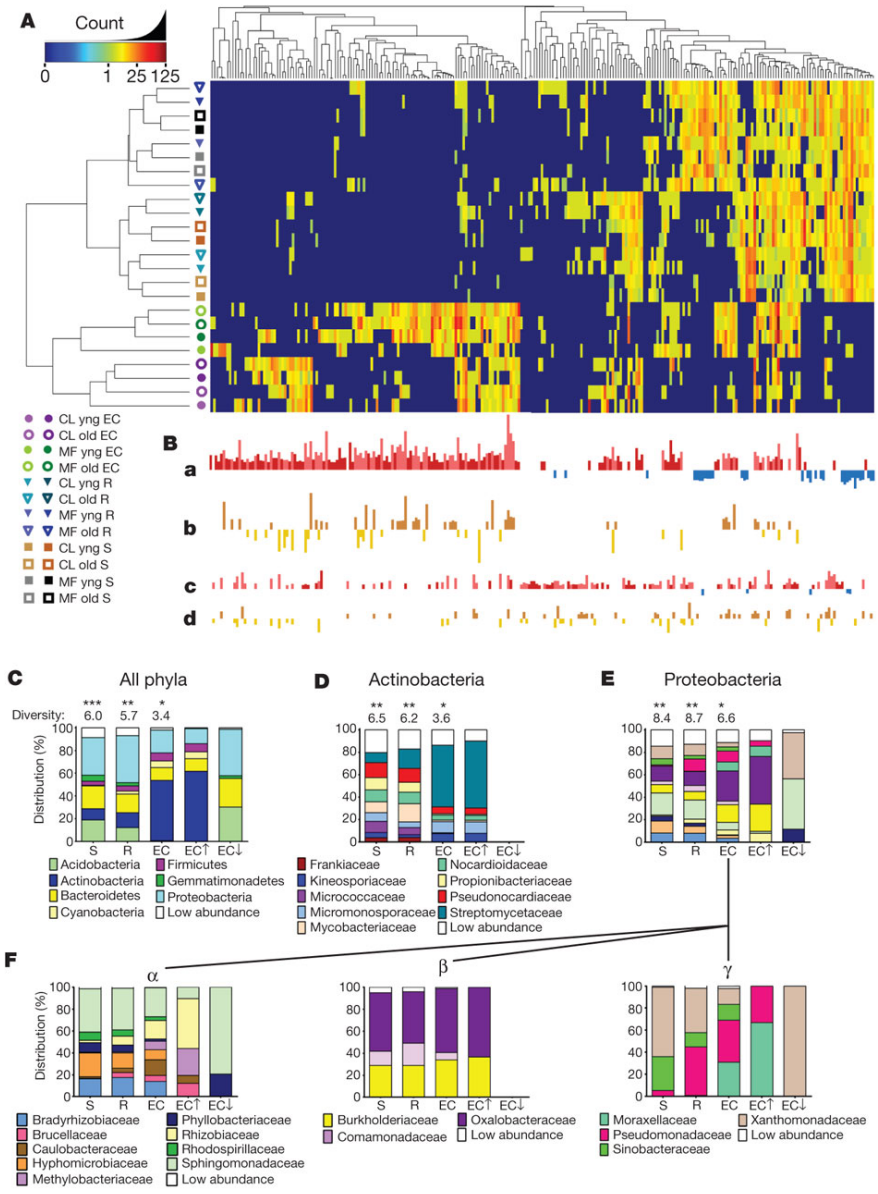


Figure 2.5: **OTUs that differentiate the EC and rhizosphere from soil**. **A** Heat map showing rarefied abundances from rhizosphere- and EC-differentiating OTUs. Different hues of the same colour correspond to different replicates as in Fig. 2.2. **B** Strength of GLMM predictions. **a**, OTUs predicted as EC enriched (red, up) or EC depleted (blue, down). OTUs with consisten behsviout in both soils are shown in darker hues. **b** OTUs that achieve higher abundance in the EC of plants in Mason Farm (brown, up) or Clayton (gold, down) than on the other soil. **c** OTUs predicted as rhizosphere enriched (as in a). **d** OTUs higher in rhizosphere in one soil type (as in b). **C** Phyla distribution of *measurable* OTUs compared with phyla of EC OTUs enriched (EC↑) or depleted (EC↓) relative to soil. Shannon diversity is given above each bar. A different number of asterisks represents a significant difference ($P < 0.05$, weighted ANOVA; methods 2.1.19 and Supplementary Table 5). **D** Family distribution from the phylum Actinobacteria. **E** Family distribution from the phylum Proteobacteria. **F** Family distribution from classes: Alphaproteobacteria (α), Betaproteobacteria (β) and Gammaproteobacteria (γ).

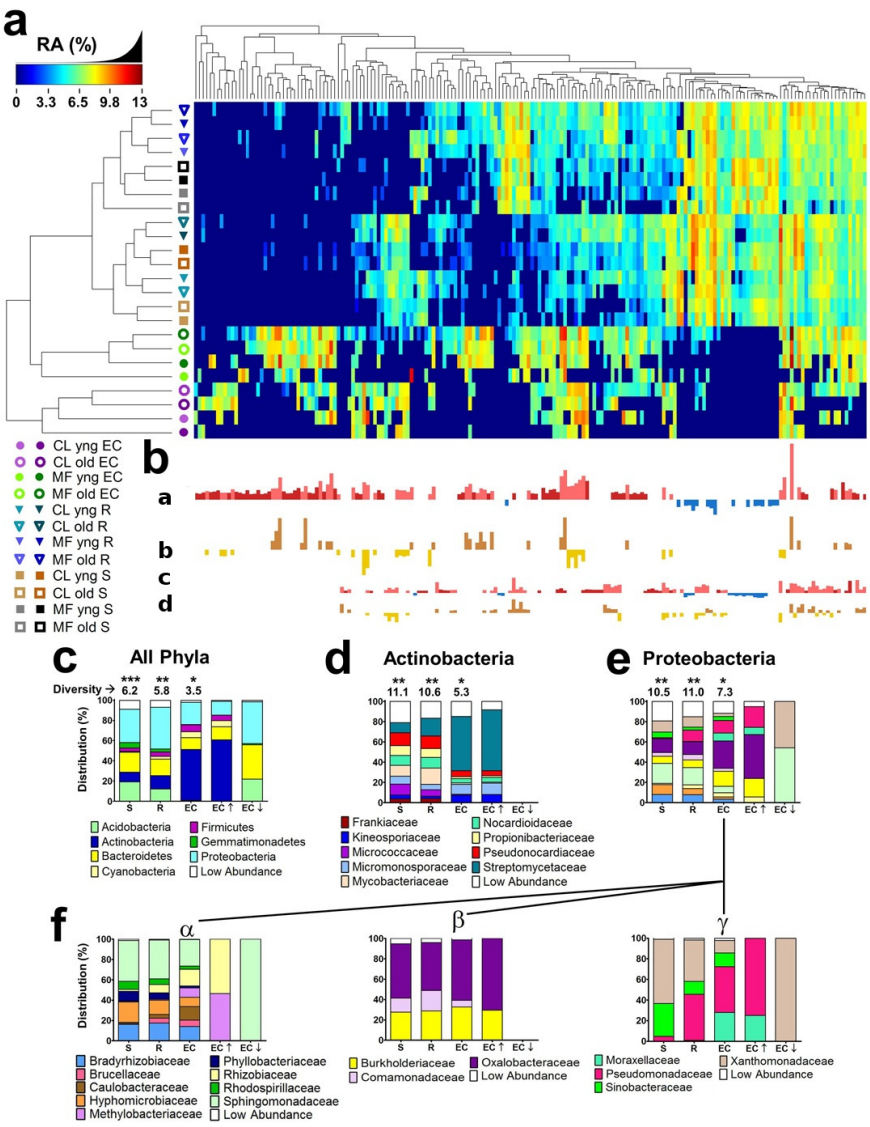


Figure 2.6: **OTUs that differentiate the EC and rhizosphere from soil**. **A** Heat map showing relative abundances from rhizosphere- and EC-differentiating OTUs. Different hues of the same colour correspond to different replicates as in Fig. 2.2. **B** Strength of GLMM predictions. **a**, OTUs predicted as EC enriched (red, up) or EC depleted (blue, down). OTUs with consisten behsviout in both soils are shown in darker hues. **b** OTUs that achieve higher abundance in the EC of plants in Mason Farm (brown, up) or Clayton (gold, down) than on the other soil. **c** OTUs predicted as rhizosphere enriched (as in a). **d** OTUs higher in rhizosphere in one soil type (as in b). **C** Phyla distribution of *measurable* OTUs compared with phyla of EC OTUs enriched (EC↑) or depleted (EC↓) relative to soil. Shannon diversity is given above each bar. A different number of asterisks represents a significant difference ($P < 0.05$, weighted ANOVA; methods 2.1.19 and Supplementary Table 5). **D** Family distribution from the phylum Actinobacteria. **E** Family distribution from the phylum Proteobacteria. **F** Family distribution from classes: Alphaproteobacteria (α), Betaproteobacteria (β) and Gammaproteobacteria (γ).

were not also enriched in the EC (Supplementary Table 3). Hence, the *A. thaliana* EC microbiome is enriched for both a shared set of OTUs commonly assembled across two replicates from two diverse soils, and a set of OTUs that are assembled from each soil.

We assessed taxonomic distributions, first those of the 778 measurable OTUs in soil, rhizosphere and EC fractions, and then those of the 256 EC-enriched and 32 EC-depleted OTUs (Figs. 2.5C and 2.6c, and Supplementary Table 3). Measurable OTUs were distributed across seven dominant phyla and contained,5070% of the usable reads in all fractions (Fig. 2.10c). Phyla distribution of the EC-enriched OTUs reflected that of the entire EC. Conversely, the phyla distribution of the EC-depleted OTUs typically resembled that of the rhizosphere fraction (Figs. 2.5C and 2.6c). The lower Shannon diversity (Figs. 2.5C and 2.6c, numbers above bars) of the EC fraction is consistent with enrichment for a subset of dominant phyla. Specifically, the EC microbiome was dominated by Actinobacteria, Proteobacteria and Firmicutes, and was depleted of Acidobacteria, Gemmatimonadetes and Verrucomicrobia, when soil types were considered either together or separately (Figs. 2.5C, 2.6c and 2.19, and Supplementary Table 5). Lower-order taxonomic analysis (Figs. 2.5D and 2.6d) demonstrated that enrichment of a low-diversity Actinobacteria community in the EC was driven by a subset of families, predominantly Streptomycetaceae.

Other phyla, such as Proteobacteria, were represented by both EC enrichments and EC depletions at the family level (Fig. 2.5E and 2.6e). Strikingly, two alphaproteobacterial families, Rhizobiaceae and Methylobacteriaceae, and two gammaproteobacterial families, Pseudomonadaceae and Moraxellaceae, dominated the EC population in their respective classes (Figs. 2.5F and 2.6f, α and γ). Equally striking was the EC redistribution of particular alpha- and gammaproteobacterial families that were common in soil and rhizosphere (Figs. 2.5F and 2.6f)

Specific OTUs, three from the family Streptomycetaceae and one from the order Sphingobacteriales, demonstrate the robustness of EC enrichments (Figs. 2.7ad and 2.8a-d). A few OTUs were either significantly enriched in rhizosphere but not in the EC (Figs. 2.7e-f and

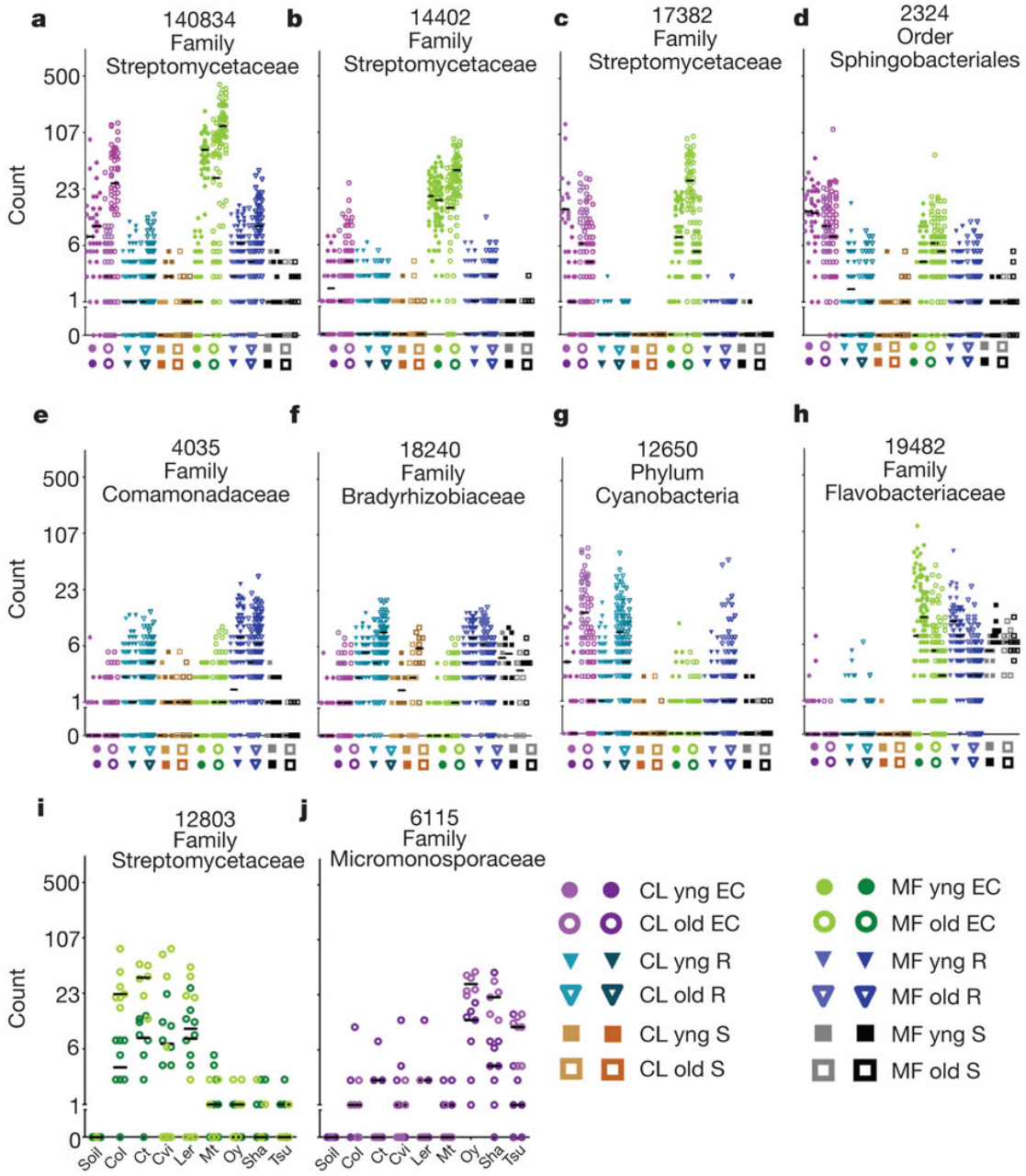


Figure 2.7: **Dot plots of notable OTUs**. Rarefied counts for each OTU are shown in a $\log_2$-scale. **ah** Abundances by sample group. Biological replicates in the same column have different hues. The median of each replicate is shown with a horizontal black bar. **i-j** Abundances by *Arabidopsis* accession. Each OTU in the figure has model predictions in several categories Supplementary Table 3).

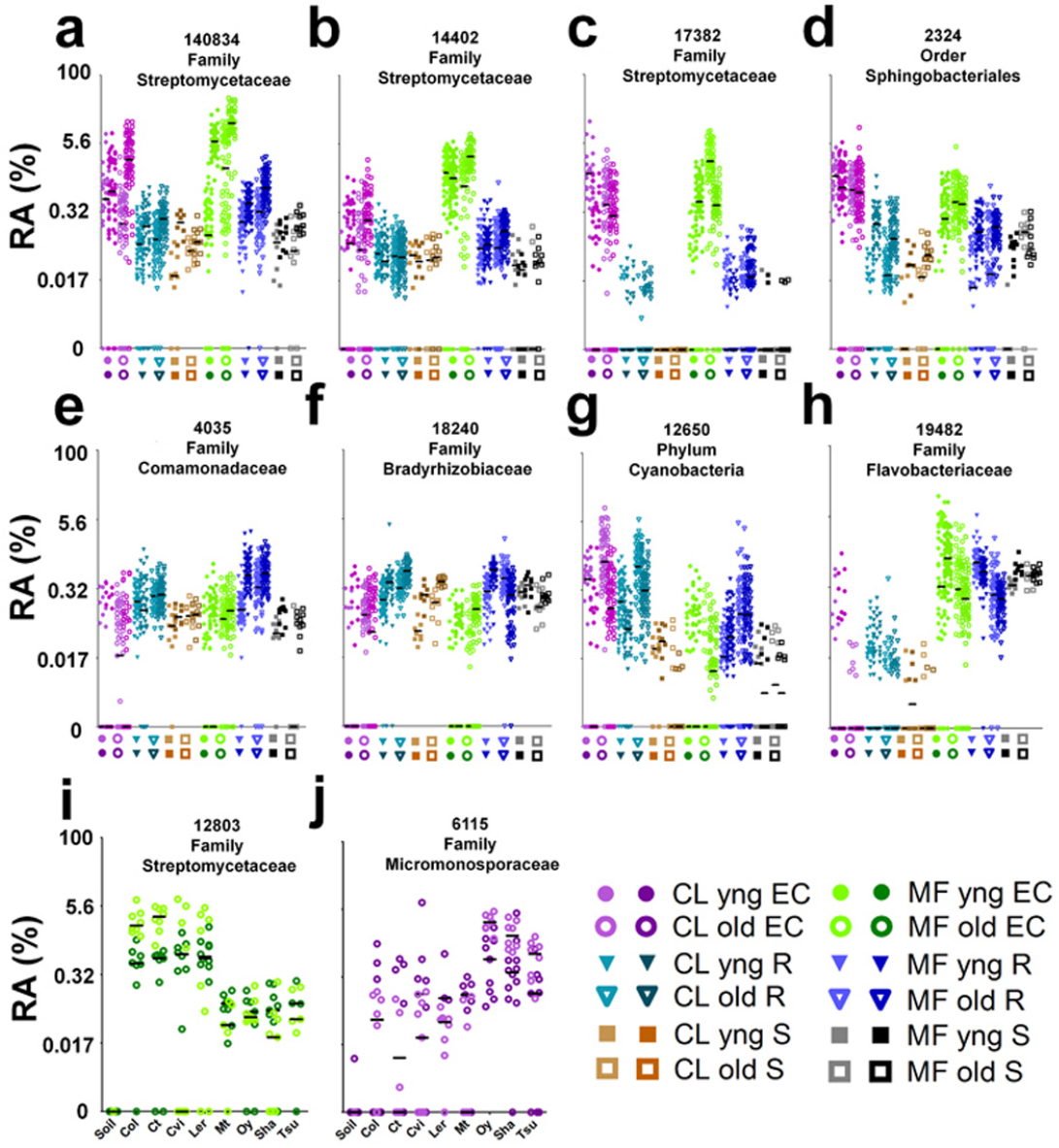


Figure 2.8: **Dot plots of notable OTUs**. Relative abundances for each OTU are shown in a $\log_2$-scale. **ah** Abundances by sample group. Biological replicates in the same column have different hues. The median of each replicate is shown with a horizontal black bar. **i-j** Abundances by *Arabidopsis* accession. Each OTU in the figure has model predictions in several categories Supplementary Table 3).

2.8e-f, and Supplementary Table 3), or were associated with one of the two developmental stages (Figs. 2.7g-h and 2.8g-h, and Supplementary Table 3). Data in Figs. 2.5, 2.6, Fig. 2.7 and 2.8, and Supplementary Table 3 demonstrate that entire taxa at various levels are enriched in or depleted from the EC microbiome. Additionally, rhizosphere taxa capable of colonizing the root vicinity are nonetheless prevented from colonizing the EC.

Several OTUs differentiated inbred *A. thaliana* accessions. Genotype-dependent enrichments and depletions were significant but weak (Supplementary Tables 5 and 3). To identify accession-dependent effects specific to a soil type or a developmental stage, we fitted a partial GLMM that modelled each genotype against bulk soil for each experiment or developmental stage group, and tested the models predictions with a non-parametric KruskalWallis test corrected for multiple testing (methods 2.1.13). We considered only those significant accession-dependent effects that were present in the same direction in both biological replicates. We further required that these OTUs have a consistent prediction in the full GLMM, which narrowed the field to 12 OTUs (or 27 with frequency-normalized data; Supplementary Table 3). In Figs. 2.7 and 2.8, we display relative abundances of two such OTUs, one for each soil type, both Actinobacteria (Figs. 2.7i-j and 2.8i-j). That these enrichments were detected by the full GLMM (which accounts for plate effects due to 454 sequencing), and were sequenced over several plates (Fig. 2.9) supports a true genotype effect. Thus, a small subset of the EC microbiome is likely to be quantitatively influenced by host-genotype-dependent fine-tuning in specific soil environments. This could allow compensatory contributions of the EC microbiome and host genome variation to overall metagenome function.

Because the rhizoplane is stripped during preparation of EC fractions, we confirmed the presence of live bacteria on roots using catalysed reporter deposition and fluorescence in situ hybridization (CARDFISH) to whole Col-0 root segments (Eickhorst and Tippkötter, 2008). Eubacteria were common on unsonicated roots (Fig. 2.10a). Actinobacteria detected with probe HGC69a were visible on the surface of roots grown in Mason Farm soil, and co-localized with a subset of the eubacterial signals using double CARDFISH (Fig. 2.10b),

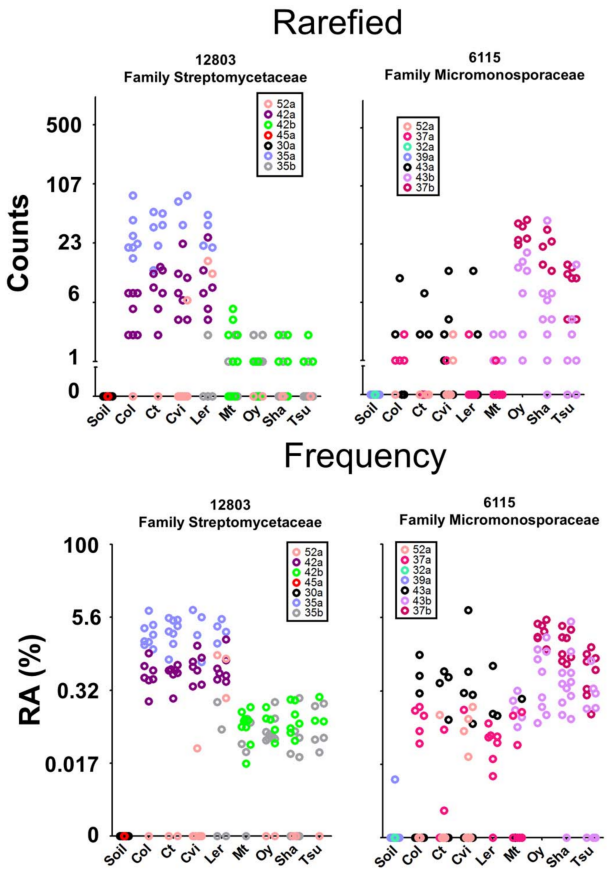


Figure 2.9: **Genotype-variable OTUs colored by sequence plate**. Displays the data from Figs 2.7i-j and 2.8i-j colored by sequence plate according to the legend within each plot. (Note: 'a' and 'b' in our plate naming scheme do not represent different regions of the same plate. All 454 regions were modeled independently in the Full GLMM).

suggesting that their enrichment in EC fractions either comes from, or egresses through, the rhizoplane. Similarly, we confirmed the rare presence on the rhizoplane of Bradyrhizobiaceae (Fig. 2.11c), a family with members defined by the GLMM as more abundant in Mason Farm rhizosphere than Mason Farm EC (Figs. 2.7f and 2.8f). We enumerated the relative number of CARDFISH signals on a set of filters made from equal amounts of material harvested in the same way as were the samples processed for pyrotag sequencing (Fig. 2.11a-b). We confirmed that Actinobacteria were found in higher abundance, and that Bradyrhizobiaceae were present in lower abundances, in EC samples than in the bulk soil and rhizosphere samples. We also noted that emerging lateral roots were typically heavily colonized by a

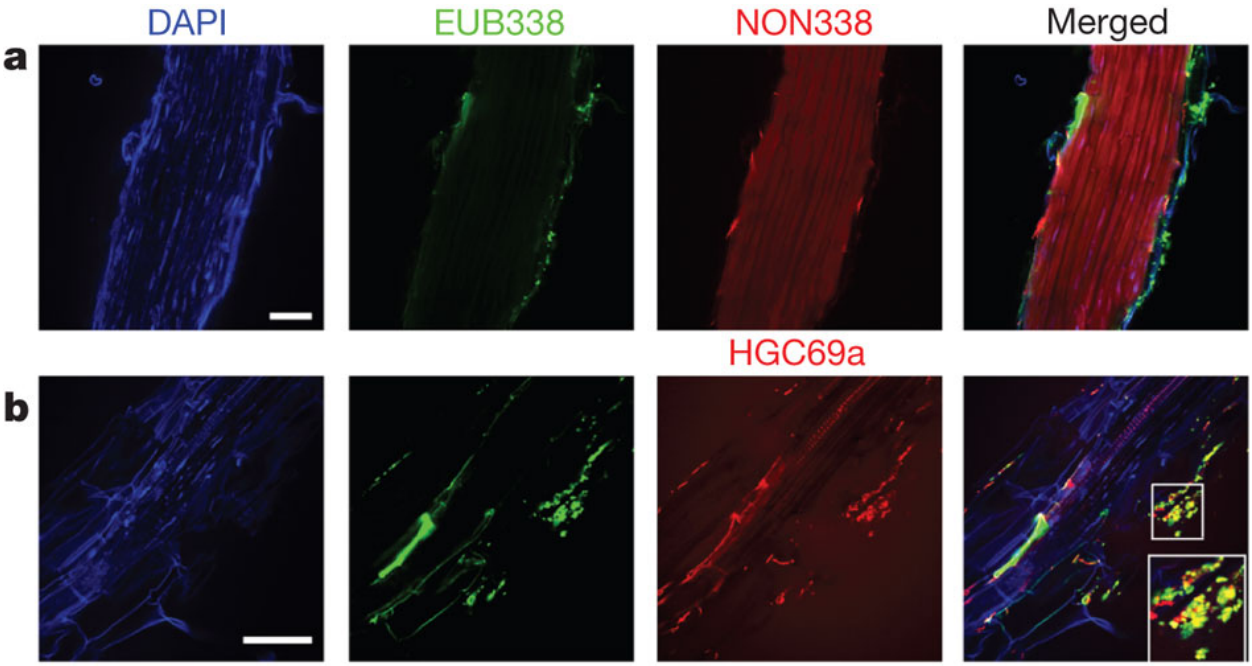


Figure 2.10: **CARD-FISH confirmation of Actinobacteria on roots**. A single set of Mason Farm yng Col-0 roots were fixed and stained using CARD-FISH. DAPI, 4',6-diamidino-2-phenylindole. Double CARD-FISH was applied using the EUB338 eubacterial probe (green) and either the NON338 probe **(a)**, which is the nonsense negative control of EUB338, or the HGC69a Actinobacteria probe **(b)**. Inset, twofold enlargement of boxed region. Scale bars, 50μm.

variety of bacteria (Fig. 2.11d) consistent with previous observations (Chi et al., 2005). These results are PCR independent support for our sequencing methods.

We present a reduced-complexity, robust experimental platform with which to study root microbiota. Our data, and similar conclusions presented in a companion publication (Bulgarelli et al., 2012) using a similar platform, provide the deepest analysis available regarding the principles of root microbiome assembly for any plant species. Remarkably, our conclusions are very similar to those in the work by Bulgarelli et al. (2012) and we identify phyla and family level enrichments in the EC fraction that largely overlap with those reported by Bulgarelli et al. (2012). We note three main differences between our study and that of Bulgarelli et al. (2012): different soils from a different continent, a different primer pair and a different portion of root harvested (top 3cm by Bulgarelli et al. (2012); whole root here).

A subset of the soil bacterial population is typically enriched in rhizosphere samples

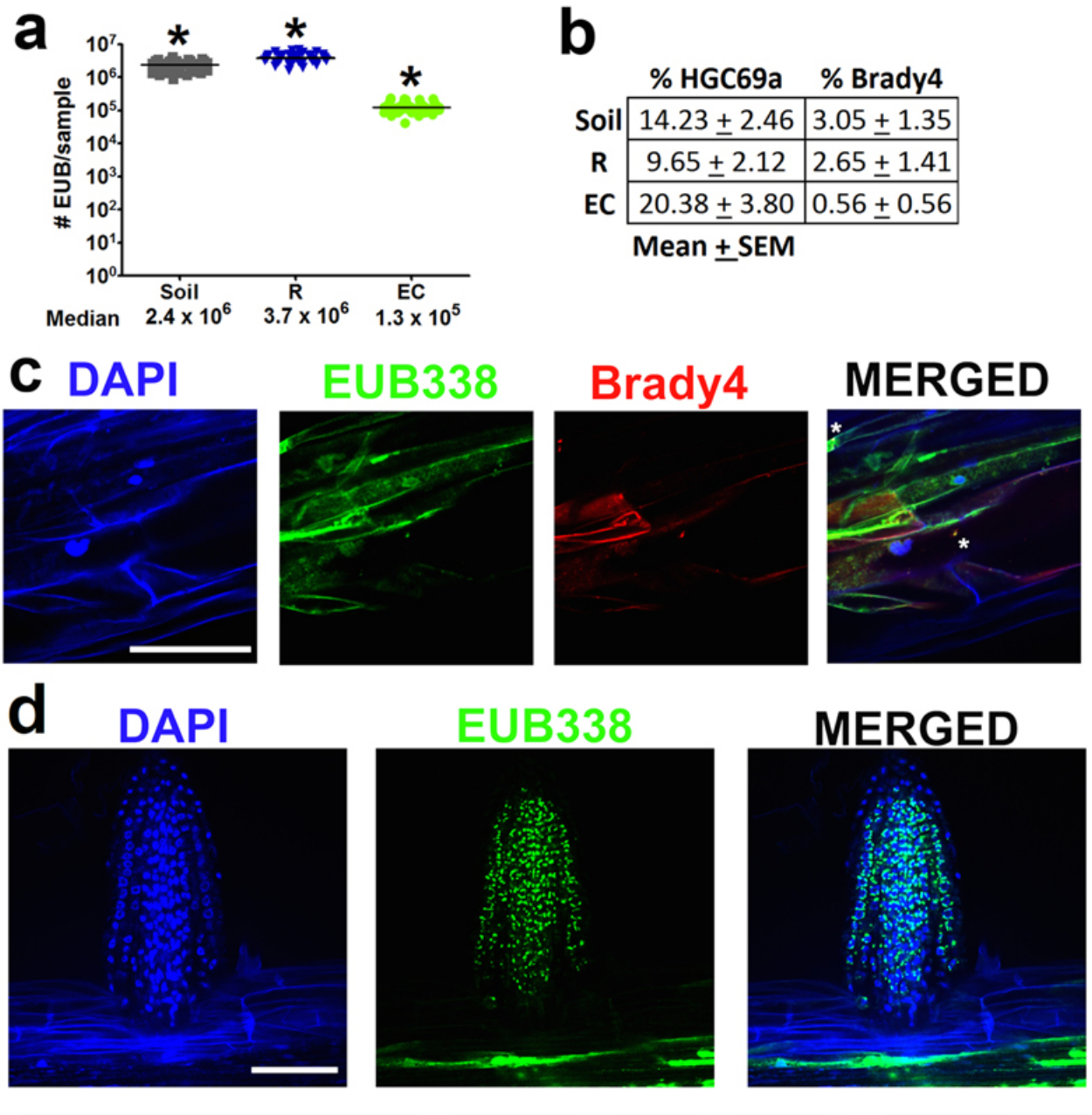


Figure 2.11: **Quantification of microbes in the three sample fractions using CARD-FISH**. Four sets of Col-0 roots were pooled, processed, diluted, and put onto filters. **a** Number of bacteria present in each sample estimated by co-staining with EUB338 and DAPI (methods 2.1.21). Sample sizes are: bulk soil (n=40), rhizosphere (n=39), and endophytic compartment (n=40). Asterisks indicates statistical significance at $p < 1x10^{-16}$ (ANOVA with post-hoc TukeyHSD) between each of the sample groups. **b** Double CARD-FISH estimates of Actinobacteria (HGC69a) and Bradyrhizobiaceae (Brady4) re;atove abundances in differengt fracations. Sample sizes are: bulk soil (n=10), rhizosphere (n=10), and endophytic compartment (n=10). **c** Double CARD-FISH of the EUB338, eubacterial probe (green) and the Brady4, Bradyrhizobiaceae probe (red), counterstained with DAPI (asterisks indicate signals that are positive in all 3 channels). **d** Newly forming lateral roots and root tips were found commonly to be heavily colonized. Scale bars represent 50 microns

(Dennis et al., 2010). Thus, a diverse bacterial community can surround the root surface and thrive there, recruited by biophysical and/or host-derived metabolic cues. We demonstrate that the *A. thaliana* microbiome undergoes dramatic loss of diversity as the spatial level of plantmicrobe 'intimacy' further increases from the external rhizosphere to the intercellular EC. Both common and soil-type specific OTUs are established inside roots grown in diverse soils. A small number of bacterial taxa, particularly the Actinobacteria family Streptomycetaceae, and several Proteobacteria families, are highly enriched in the EC. Actinobacteria are well known for production of antimicrobial secondary metabolites (Firáková et al., 2007), and many proteobacterial families contain plant-growth-promoting members. Conversely, several taxa (Acidobacteria, Verrucomicrobia and Gemmatimonadetes, and various proteobacterial families) that are common in soil and rhizosphere are depleted from the EC. This depletion suggests that these taxa are either actively excluded by the host immune system, outcompeted by more successful EC colonizers or metabolically unable to colonize the EC niche. Our identification of a limited-diversity EC facilitates detailed characterization of the isolates comprising the core *A. thaliana* microbiome, which could facilitate the design of community-based plant probiotics.

Within the EC, we identified rare cases of quantitative variation in the enrichment of specific bacteria at two developmental stages or by different host genotypes, consistent with rare genotype-dependent associations noted by Bulgarelli et al. (2012). The former result suggests that the EC microbiome is robust to the sourcesink differences across these two developmental stages, which may be related to the relatively high frequency of putative saprophytes defined in Bulgarelli et al. (2012). The latter result suggests that host genetic variation can drive either differential recruitment of beneficial microbes and/or differential exclusion. A limited-diversity EC microbiome with common features suggests similar host needs across *A. thaliana*, potentially extending to other plant taxa. These are probably fulfilled by contributions from a limited number of bacterial taxa across diverse soils. The identification of genotype-specific endophyte associations in particular soils may signal interactions that

meet environment-specific host needs, balancing contributions of EC microbiome and host genome variation to overall metagenome function. These two generalities suggest that the *A. thaliana* root microbiome might assemble by core ecological principles similar to those shaping the mammalian microbiome, in which core phylum level enterotypes provide broad metabolic potential combined with modest levels of host-genotype-dependent associations that individualize the metagenome (Arumugam et al., 2011; Spor et al., 2011). Isolation and characterization of the microbes that define host-genotype-dependent associations, and characterization beyond the 16S gene, should be particularly instructive in unravelling the molecular rules contributing to endophytic colonization and persistence.

2.1 Methods

2.1.1 General strategy

Seed sterility was verified by plating and deep-sequencing of homogenates from sterile seedlings (Fig. 2.12; methods 2.1.3). We established seedling growth, harvesting and DNA preparation pipelines as detailed in the specific sections below. We defined the bacterial community within each soil, and the community associated with plant roots across a number of controlled experimental variables: soil type, plant sample fraction, plant age and plant genotype. For plant age, we harvested roots from two developmental stages: at the formation of an inflorescence meristem (yng) and during fruiting when $\geq$50% of the rosette leaves were senescent (old). The former represents plants at the peak of photosynthetic conversion to carbon, whereas the latter represents a stage well after the sourcesink shift has occurred, marking the change in carbon allocation from vegetal to reproductive utilization (Masclaux et al., 2000). We prepared two microbial sample fractions from each individual plant: a rhizosphere (bacteria contained in the layer of soil covering the outer surface of the root system that could be washed from roots in a buffer/detergent solution), and EC (bacteria from within the plant root system after sonication-based removal of the rhizoplane; Fig. 2.13). We also collected control soil samples (soil treated in parallel, but without a plant grown in it).

2.1.2 Soil collection and analysis

For each full-factorial experiment, the top 8inches of earth were collected with a shovel and transported to the lab in closed plastic containers at room temperature from two collection sites. The first collection site, Mason Farm, is managed by the North Carolina Botanical Garden and is free of pesticide use and heavy human traffic and is located in Chapel Hill, North Carolina, USA (+35°53'30.40", -79°1'5.37"). The second collection site is the Central Crops Research Station in Clayton, North Carolina, USA (+35°39'59.22", 78°29'35.69") and is also free of pesticide use. Visible weeds, twigs, worms, insects and so on were removed with gloves, and the soil was then crushed with an aluminium mallet to a fine consistency and sifted through a sterile 2mm sieve. Because sieved soil from Mason Farm drained poorly and test plants grown in it suffered from hypoxia, we adopted the practice of mixing sterile (autoclaved) playground sand into both Mason Farm (MF) and Clayton (CL) soils at a soil:sand volume ratio of 2:1. Soil micronutrient analysis was performed on pure and 2:1 mixed soils by the University of Wisconsin soil testing labs.

2.1.3 Seed sterilization and germination

All seeds were surface-sterilized by a treatment of 1min in 70% ethanol with 0.1% Triton-X100, followed by 12min in 10% A-1 bleach with 0.1% Triton-X100, followed by three washes in sterile distilled water. Seeds were spread on 0.5% agar containing half-strength Murashige & Skoog (MS) vitamins and 1% sucrose. Seeds were stratified in the dark at 4°C for one week, then germinated at 24°C under 18h of light for one week. Seed coat sterility was confirmed by lack of visible contamination on MS plates during germination, and also by absence of visible contamination after plating some of the whole seeds on KB, 1/10-strength LB and 1/10-strength '869' bacterial growth media.

To address whether there were seed-borne microbes that might survive surface sterilization, one-week-old seedlings were taken from sterile MS plates and homogenized by aseptic bead beating under non-bacteriolytic conditions (three 3-mm glass balls per 2-ml tube, with 300-μl PBS, using a FastPrep from MP Bio at speed 4.0ms^1 for 10s). The homogenate was streaked

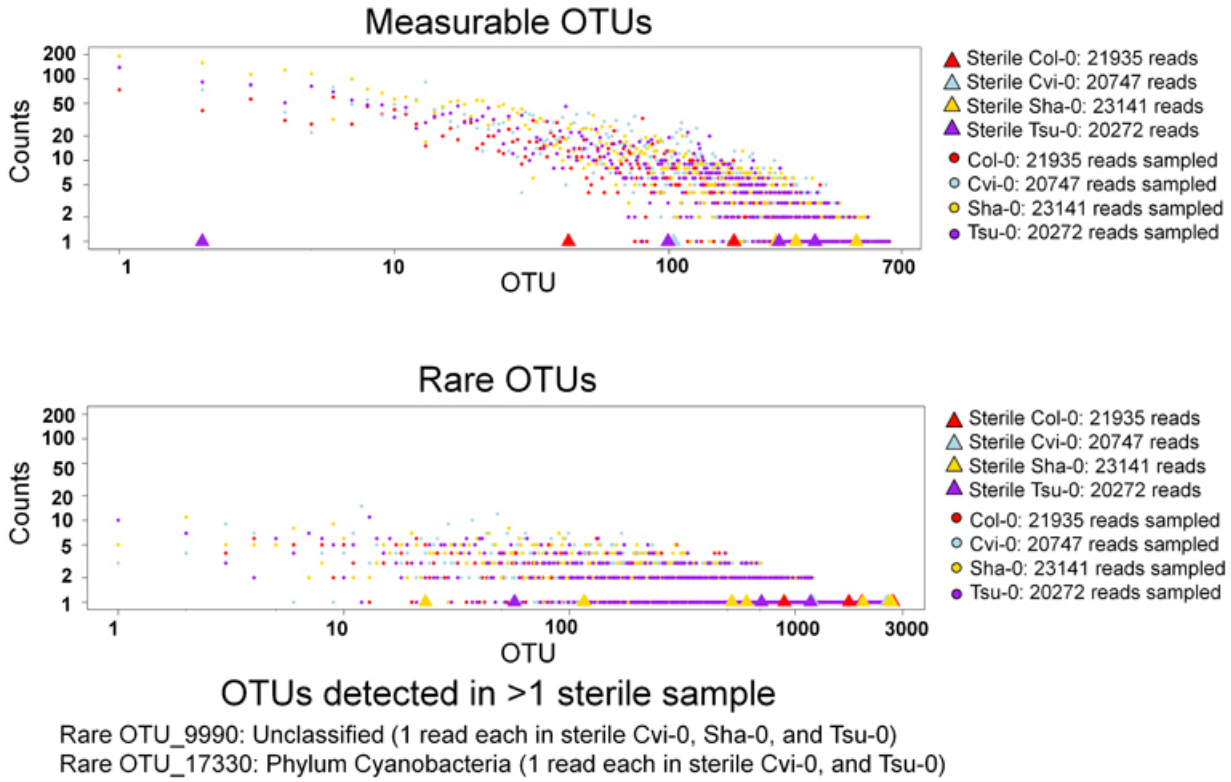


Figure 2.12: **Pyrosequencing of sterile seedlings as compared to vs. non-sterile EC samples**. Abundance of OTUs found in sterile (triangles) and non-sterile (circles) samples. Each position on the X axis represents an OTU in the full dataset (*measurable* OTUs on top, rare OTUs on bottom) and the position on the Y axis represents the number of sequence reads found in that OTU. Both axes are shown in log scale.

onto 1/10-strength LB, 1/10-strength '869' and KB media. No colonies were observed. To detect potential unculturable microbes, we pyrosequenced 16S amplicons from the same homogenates using bacteriolytic DNA preps from the genotypes Col-0, Cvi-0, Sha-0 and Tsu-0 (Fig. 2.12). Each accession was individually barcoded and sequenced with 1114F and 1392R, yielding 21,935, 20,747, 23,141 and 20,272 reads, respectively. A matching number of total reads was sampled from each accession using pooled data from the full experimental data set for comparative analysis. Thus, 86,095 high-quality reads were obtained from both non-sterile plants and sterile plants, the majority of which were chloroplast sequences. Far more non-plant reads were obtained from the non-sterile plants (19,093 of 86,095, or 22%) vs. sterile plants (34 of 86,095, or 0.04%), a difference approaching three orders of magnitude.

The 34 reads from non-sterile plants were members of 31 OTUs (triangles some overlap on the log-scale axis). No OTU in a sterile plant sample was represented by more than one read, and only two OTUs were shared by more than one of the accessions; both of these shared OTUs were not in the measurable set, and had poor taxonomic classification. 11 of these 31 OTUs were not represented in the non-sterile samples. Furthermore, by including extra unused barcodes in our mapping files, or by sequencing sterile water in excess, we have been able to occasionally 'detect' single representatives of OTUs in our dataset, demonstrating that technical noise can cause singletons (data not shown). While we cannot rule out that unculturable microbes survive surface sterilization and exist at extremely low abundance, we have no evidence that such microbes exist in A. thaliana roots.

2.1.4 Seedling growth

One-week-old healthy seedlings were aseptically transplanted from MS plates to sterile (autoclaved) 2.5-inch-square pots filled with either MF or CL soil, with one seedling per pot. Seedlings were transferred by lifting from underneath the cotyledon leaves using open tweezers; no pressure was applied to the hypocotyl. Some pots were designated 'bulk soil' and were not given a plant. All pots, including bulk soil controls, were always watered from the top with a shower of distilled water (non-sterile) as an accessible proxy for rain water that avoids chlorine and other tapwater additives. Pots were spatially randomized and placed in growth chambers providing short days of 8h light (8001,000lx) at 21°C and 16h dark at 18°C. The use of short days was to help synchronize flowering time between *A. thaliana* genotypes and to facilitate robust rosette and root growth. After harvesting the floral transition developmental stage, remaining plants and bulk soils were moved from the growth chamber to 16 h days in the greenhouse to promote a more synchronized flowering and senescence for the senescent developmental stage.

2.1.5 Harvesting

Each plant was killed and harvested at one of two developmental time points: (1) at the floral transition and (2) after fruiting when senescence is well underway. We considered the floral transition to have begun when the shoot apical meristem was first apparent in

five or more plants. Cvi-0, Sha-0 and Ct-1 occasionally flowered one to two weeks earlier under our conditions than the other A. thaliana genotypes. The senescence harvest began when five or more plants showed 50% or more yellow and/or brown rosette leaves (Levey and Wingler, 2005); this occurred approximately four to five weeks after transfer to the greenhouse. Senescence occurred in the same order as bolting (flowering).

Our maximum harvesting and processing capacity was 30 plants per day, meaning that each harvesting period for each full-factorial biological replicate (90 pots) lasted between one and two weeks. On each harvest day, we strove to represent all genotypes and at least one bulk soil to avoid potential confounding harvesting artefacts with genotype effects. Because we harvested as many pots each day as time allowed, we did not always harvest in multiples of our genotype number and did not have equal representation of each genotype on each harvest day.

The harvesting scheme is visualized in Fig. 2.13a-c. Using gloves and a flame sterilized work surface, plants are overturned, pots are removed, and soil is crumbled/brushed away leaving $\leq$1 mm rhizosphere soil on roots. The aboveground plant organs were aseptically removed. Loose soil was manually removed from the roots by kneading and shaking with sterile gloves (sprayed with 70% EtOH) and by patting roots with a sterile (flamed) metal spatulathis neighbouring soil fell to the sterile (flamed) work surface. We followed the established convention of defining rhizosphere soil as extending up to 1mm from the root surface (van Elsas et al., 1988) and we removed loose soil on all root surfaces until remaining aggregates were within this range. Roots were placed in a clean and sterile 50-ml tube containing 25ml phosphate buffer (per litre: 6.33g of $NaH_2PO_4*H_2O$, 16.5g of $Na_2HPO_4*7H_2O$, 200μl Silwet L-77). Tubes were vortexed at maximum speed for 15s, which released most of the rhizosphere soil from the roots and turned the water turbid. The turbid solution was then filtered through a 100 μm nylon mesh cell strainer into a new sterile 50 ml tube to remove broken plant parts and large sediment. The roots were transferred from the empty tube to a new sterile 50 ml tube with 25 ml sterile phosphate buffer, and the turbid filtrate was centrifuged for 15min at

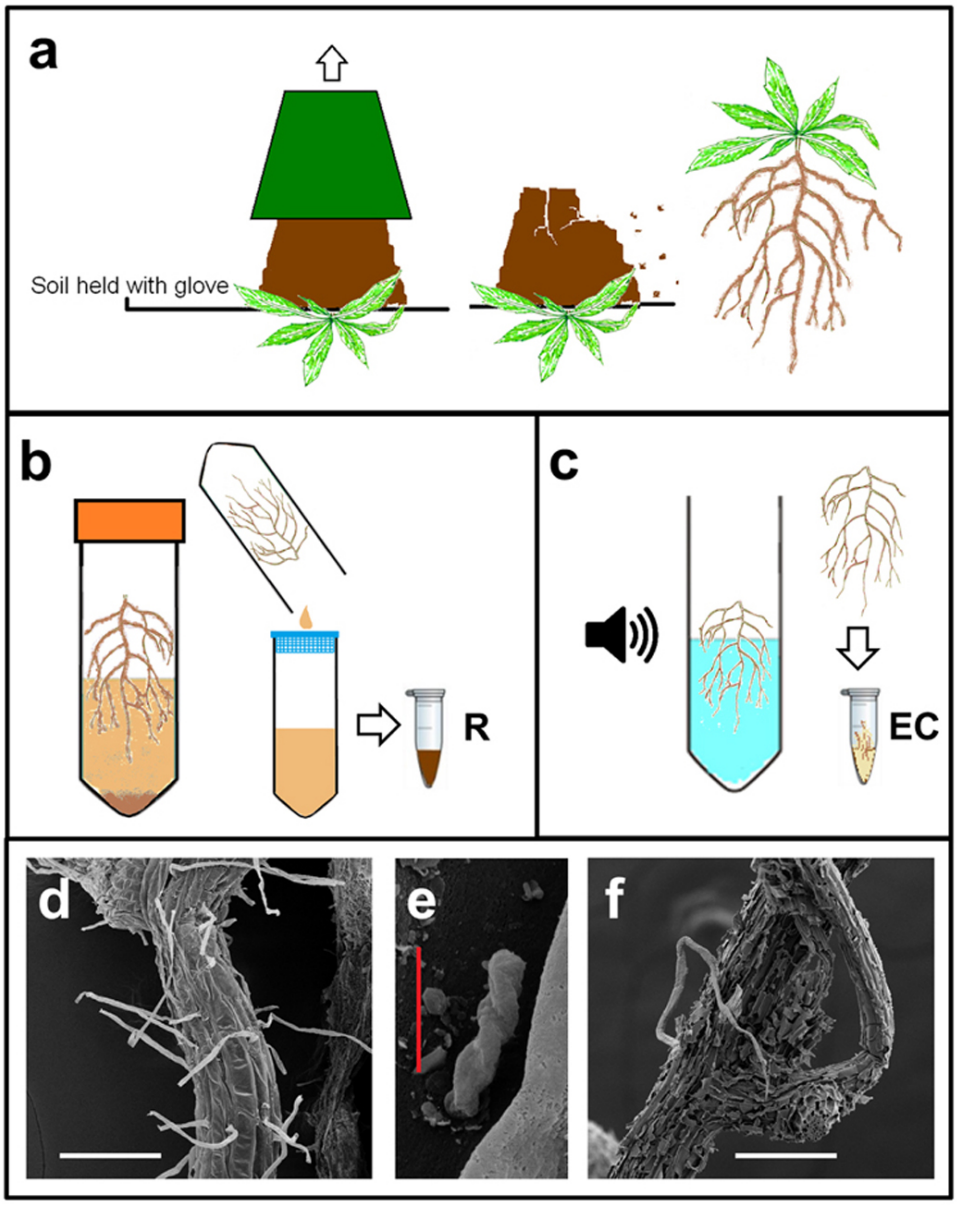


Figure 2.13: **Harvesting scheme.a** Plants are overturned, pots and soil are removed, leaving $\leq$1 mm rhizosphere soil on roots. **b** The above-ground parts are cut away and rhizosphere soil is rinsed from roots; the rinse is pelleted and becomes the rhizosphere R fraction. **c** Roots are sonicated. The surface-cleaned roots are then snap frozen and lyophilized to become the EC fraction. **d** SEM showing intact root surface after rhizosphere soil has been removed, but prior to sonication. Scale = 100 microns. **e** SEM showing a root-surface bacterium on root shown in d. Scale = 1 micron. **f** SEM showing the disruptive clearing of nearly the entire root surface after sonication. Scale = 100 microns.

3,200g to form a pellet containing fine sediment and microorganisms.

Most of the supernatant was removed and the loose pellets were resuspended and transferred to 1.5 ml microfuge tubes, which were then spun at 10,000g for 5min to form tight pellets, from which all supernatant was removed. These rhizosphere pellets, averaging 250mg, were flash-frozen in liquid nitrogen and stored at 80C until processing. The root systems, while in the 25ml of new buffer, were cleaned of remaining debris with sterile tweezers and transferred to new sterile buffer tubes until the buffer was clear after vortexing (without major sediment on the tube bottom). The roots were then sonicated in a Diagenode Bioruptor at low frequency for 5min (five 30 s bursts followed by five 30 s rests). The sonication further disrupted tiny soil aggregates and attached microbes, cleaning the root exterior. We opted for physical removal of surface microbes by sonication instead of killing them with bleach because sequencing measures DNA; at lower concentrations, bleach kills microbes without necessarily destroying the DNA. Although an extended bleach treatment would also destroy unwanted DNA, it could also enter roots and destroy DNA of interest.

After sonication, the roots were snap-frozen, freeze-dried to remove ice and then stored at 80°C until processing. Our rhizosphere and EC fractions were collected using time-practical protocols designed to partition sequencing-quality DNA and may differ slightly from classic definitions of these fractions that rely on partitioning culturable bacteria. We note that sonication may leave some rhizoplane microbes behind, especially if they are in a microniche shielded from the ultrasound. Such artefacts may cause our collected fractions to differ from theoretical definitions.

2.1.6 DNA extraction

To extract DNA, the samples were resuspended in a lysis buffer and microbial cells were mechanically lysed through bead beating. For all bulk soil and rhizosphere data, bead beating and purification were performed with the MoBio PowerSoil kit (SDS/mechanical lysis) because of its unmatched ability to remove humics and other PCR inhibitors in our soil. EC DNA from Arabidopsis experiments was prepared with the MP Bio Fast DNA Spin

Kit for soil (also a SDS/mechanical lysis) because the more intense bead-beating protocol and lysis matrix gave improved lysis of whole roots and higher DNA yield, and soil PCR inhibitors were less of a problem with these samples. Our procedure yielded around 1μg of DNA per rhizosphere sample, and more total DNA for EC samples (although a significant portion of EC DNA sequenced was of host origin). Although MoBio Powersoil and MP Bio Fast DNA use highly similar bead-beating/mechanical lysis methods, we developed a custom method of sample pre-homogenization that allowed us to prepare some EC samples using the MoBio kit. A comparison of Col-0 fractions soil, rhizosphere and EC across four soil digs of MF, where EC was prepared using MoBio in two digs and MP Bio in the other two digs, shows that although we cannot rule out a slight kit effect, both kits produce highly similar clustering separating EC from rhizosphere and soil fractions (Fig 2.4, replicates 3 and 4). DNA quantity was assessed with the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) and a plate fluorospectrometer.

2.1.7 PCR

For each 1114F-barcoded 1392R primer set, PCR reactions with ∼10ng of template were performed in triplicate along with a negative control to reveal contamination. The PCR program used was 95°C for 3min followed by 30 cycles each of 95°C for 30s, 55°C for 45s and 72°C for 1min, followed by 72°C for 10min and then cooling to 16°C. We first verified that the no-template control did not contain DNA via gel electrophoresis, and then pooled the three replicate PCR products and quantified DNA from each pool with PicoGreen (Invitrogen). Pooled PCR products from 3048 barcoded samples were then combined in equimolar ratios into a master DNA pool, which was cleaned with Mo-Bio UltraClean PCR Clean-Up kit before submission for standard JGI pyrosequencing using a half-plate of Roche 454-FLX with titanium reagents.

2.1.8 454 pyrotag sequencing

To identify organisms present in each sample, 454 sequencing of the SSU rRNA genes was performed. For 454 sequencing, the SSU rRNA genes present in each sample were amplified with the primers 1114F and 1392R containing the 454 adaptors (Engelbrektson et al., 2010).

Each sample was assigned a reverse primer with a unique 5 bp barcode, allowing 3048 samples to be pooled per half-plate. In preparation for sequencing, working aliquots of the master pool were immobilized on beads and amplified by emulsion PCR, the emulsion was broken with isopropanol, DNA-carrying beads were enriched and the enriched beads were loaded on the instrument for sequencing. During the emPCR protocol, we reduced the amplification primer amount from 460mul in the standard protocol to 58mul per emulsion cup. This is the same amount of primer used for the paired-end emPCR protocol. One-and-three-quarter million beads were loaded in each plate region (reduced from 2,000,000 beads per region in the standard protocol). A detailed standard protocol is available on request.

2.1.9 Primer test and technical reproducibility

We first tested three sets of broad-specificity 16S rRNA 5' primers [REFS, WRONG IN PAPER] (Fig. 2.14a-b) and established technical reproducibility metrics. We used 13 samples chosen from each of the three sample fractions (soil, rhizosphere and EC) and both soil types (MF and CL) (Fig. 2.14c). Each sample was amplified individually with each of the forward primers (804F, which broadly targets bacteria and archaea; 926F, a universal primer; and 1114F, which broadly targets bacteria), paired with the barcoded universal reverse primer (1392R) and sequenced twice to measure technical reproducibility. We identified bacteria by grouping highly similar (97% identity) sequences into OTUs (methods 2.1.11). We chose 1114F for our experiments, on the basis of its broad coverage of the bacterial domain (Lane, 1991) and higher usable data yield (Figs. 2.14f-i and 2.15).

To assess possible bias introduced by amplification for pyrotagging, we compared the taxonomic distribution of a metagenome library created without amplification with a corresponding pyrotag dataset. Both datasets are from Col-0 Mason Farm young samples. 16S rDNA reads from this metagenome library (One HiSeq lane; more than 400 million 150 bp paired-end reads) were extracted by alignment against the 16S Silva database (release 106). Aligned reads were then assigned a taxonomy using an RDP training set built with the Greengenes reference database (version: May 9 th 2011). This allowed classification of 57,663

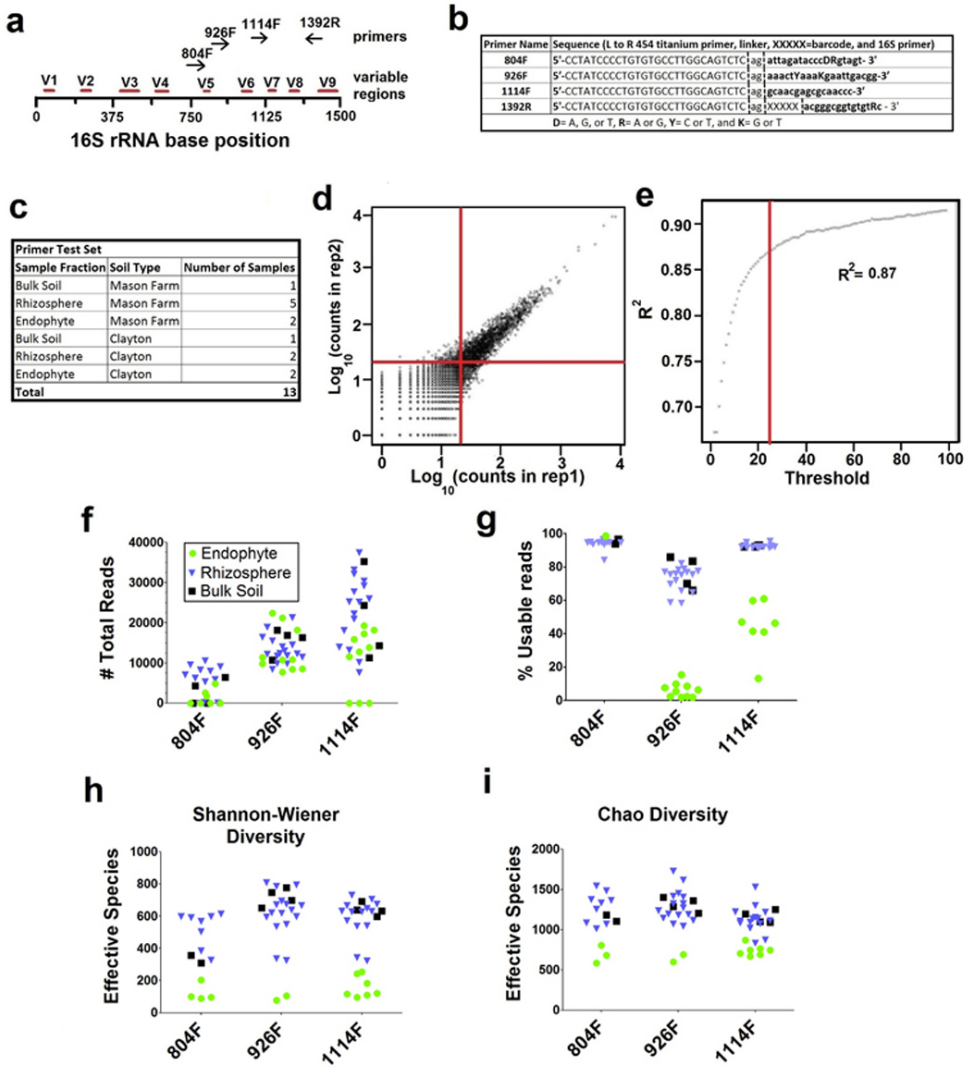


Figure 2.14: **Primer test and technical reproducibility.a** Position on the 16S gene of each of the primers tested. **b** Sequence of each primer used. **c** Composition of the 13 samples tested. **d** Log10 transformation of raw reads per OTU for one independent replicate (x-axis) vs. the other (y-axis), where both replicates were PCR-amplified and sequenced from the same sample (axes are in $\log_{10}$-scale). The intersection of the red lines shows where an OTU with 25 reads in each replicates would lie. e) Progressive drop-out analysis displaying the R^2 correlation of the data in d as OTUs with low read numbers are discarded. Red line indicates the correlation when OTUs under 25 reads are removed. In f-i green circles are EC samples, blue triangles are R samples, and black squares are bulk soil samples. **f** Total reads from different forward primers. **g** Percent of the *usable* reads from f which are not identified as plant or chimeric OTUs. **h** Shannon-Weiner species diversity of 1000 *usable* reads. *i* Chao1 diversity of 1000 usable reads from each sample.

16S reads from the metagenome sample using a bootstrap threshold ¿=0.50. There is an excellent overall correlation between the relative abundance of pyrotags and metagenome 16S rDNA reads across the major phyla represented in the datasets. Only two major classes, Thaumarchaeota and Planctomycea, were not amplified by the 1114F-1392R primers (Fig. 2.15). Slightly higher abundance of Actinobacteria and Betaproteobacteria was observed in pyrotag data than in metagenome 16S reads.

This was investigated further. For those classes in which underrepresentation in the pyrotag data are observed (red class names in Fig. 2.15), we used in silico PCR analyses using the Greengenes database as template and our pyrotags primer pair, allowing a maximum of 2 mismatches, to investigate at which taxonomic level the under-representation would be discerned (Fig. 2.15, righ side). We show that Thaumarchaeota (class) and Planctomycea (class) may be misrepresented in our pyrotag data. Since the Greengenes database contains many sequences amplified with the 1392R primer and therefore lacks this primer's sequence, we removed all sequences shorter than 6,449 (in absolute position) in our reference database to minimize false negative rate (i.e. sequences not amplifying because they are not long enough to match the 1392R primer sequence)

We identified bacteria present by grouping highly similar (97% identity) sequences into OTUs using a standard QIIME (quantitative insights into microbial ecology)-based pipeline (Caporaso et al., 2010) with default settings; thus, this stand-alone test consists of a different set of OTUs than those described in this work. The primer test samples are included in our submitted data and are found on 454 half-plates 26b and 27a. The progressive drop-out analysis, displaying the coefficient of determination (R^2) of the least-squares regression between the two technical replicates as low-abundance OTUs are sequentially discarded, was calculated using the software R with a custom script available at `http://labs.bio.unc.edu/Dangl/Resources/scripts_Lundberg_et_al_2012.htm`.

2.1.10 Primer specificity sequence

- 804F prokaryote: 5'-agattagatacccdrgtagt-3'.

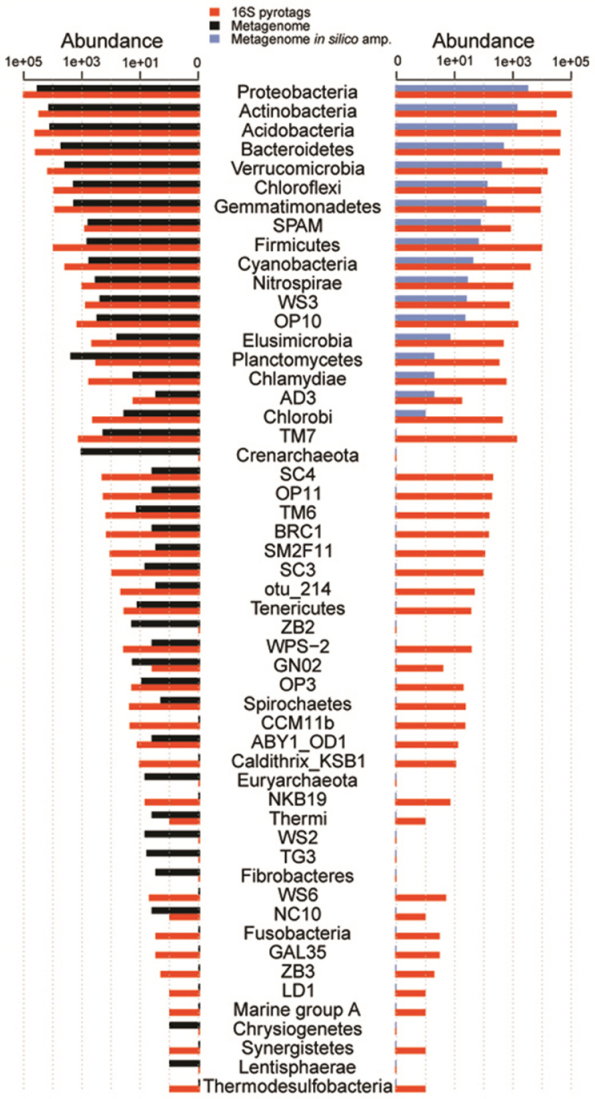


Figure 2.15: **Test for PCR bias in pyrotagging**. Taxonomic abundance comparison between 16S pyrotagging and metagenome (left), and 16S pyrotagging and *in silico* metagenome amplification (right).

- 926F universal: 5'-actcaaaggaattgacgg-3'.
- 1114F bacteria: 5'-gcaacgagcgcaaccc-3'.
- 1392R barcoded universal: 5'-XXXXXacgggcggtgtgtrc-3'.

2.1.11 Sequence processing pipeline and assignment of OTUs

As each 454 plate was sequenced, raw reads from individual plates were immediately run through PYROTAGGER (Kunin and Hugenholtz, 2010) to diagnose plate quality so that plates could be re-queued if necessary. Plates with a reasonable number of long, high-quality raw reads with matching barcodes were used in the final analysis of OTU picking and taxonomy assignment. Using QIIME-1.4.0 (Caporaso et al., 2010), short reads were removed and the remaining reads were trimmed to 220bp, and low-quality reads were removed from the analysis using default quality settings (`http://qiime.org/scripts/split_libraries.html`). These high-quality sequences were clustered into OTUs using a custom script derived from otupipe (`http://drive5.com/otupipe`). The three main steps used from otupipe include (1) de-replicating sequences to reduce the size of the data set and the run time of clustering analysis, (2) de-noising sequences by forming clusters of 97% identity and representing these with the consensus sequence, and (3) forming OTUs by clustering de-noised consensus sequences at 97% identity.

The consensus sequence of sequences in each OTU was used as a representative sequence. Each representative sequence was assigned a taxonomy by two methods: (1) using the RDP classifier (Sul et al., 2011) trained on the 4 February 2011 Greengenes reference sequences and (2) by assigning the Greengenes (DeSantis et al., 2006) taxonomy of the best BLAST hit within a combined database including the complete Greengenes 16S database and 18S *A. thaliana* sequences from NCBI. By the BLAST-based method, sequences without a hit below the E-value threshold of 0.001 are considered unclassified.

For taxonomy-supervised classification, reads that passed default QIIME quality thresholds (but that were not clustered into OTUs) were trimmed to 220bp and were classified via RDP

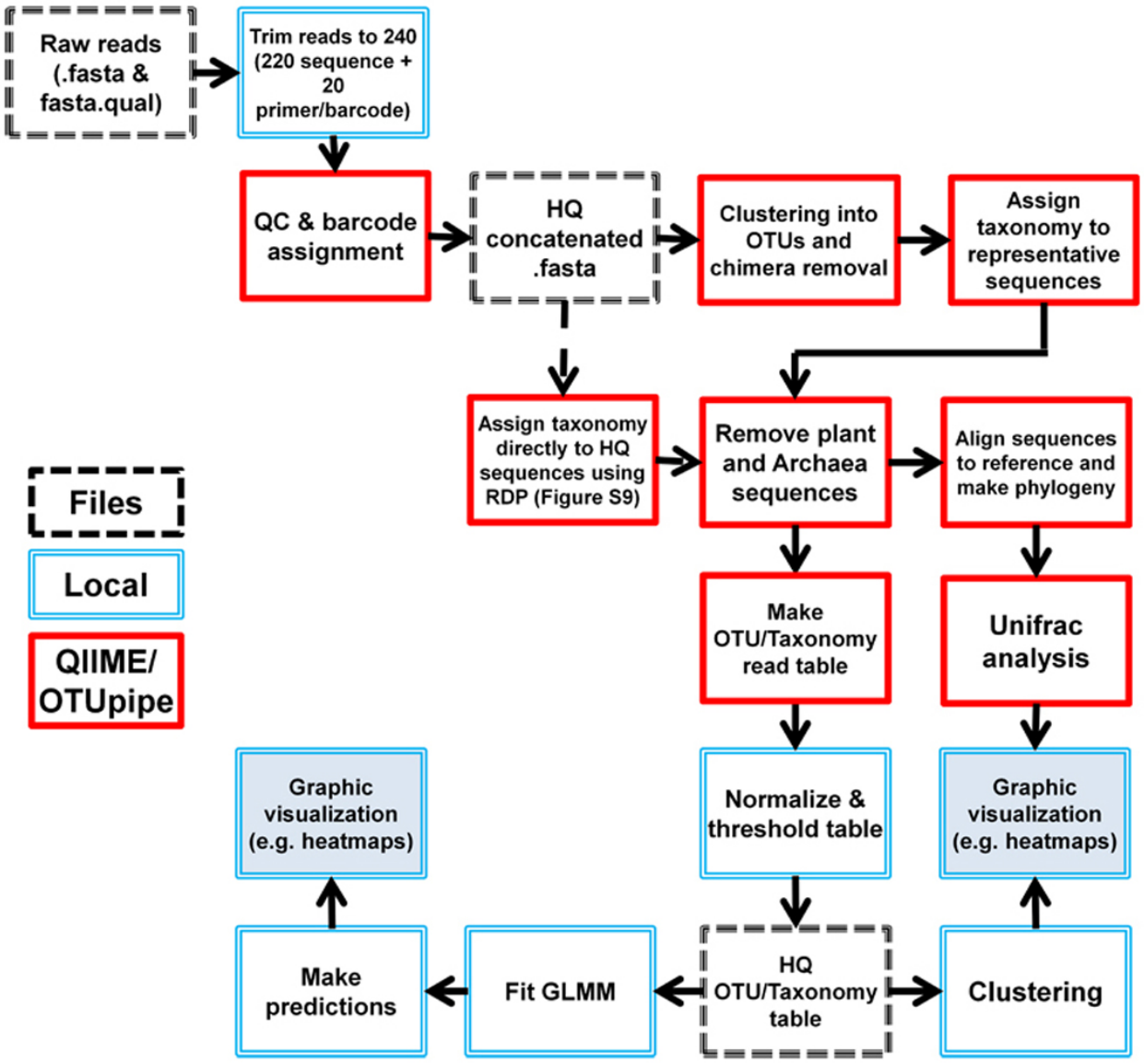


Figure 2.16: **Informatics pipeline**. Order of events. Broken-line black-line boxes represent files. Blue double-line boxes describe events that occur locally using custom scripts. Red boxes describe events that are implemented through QIIME/OTUpipe.

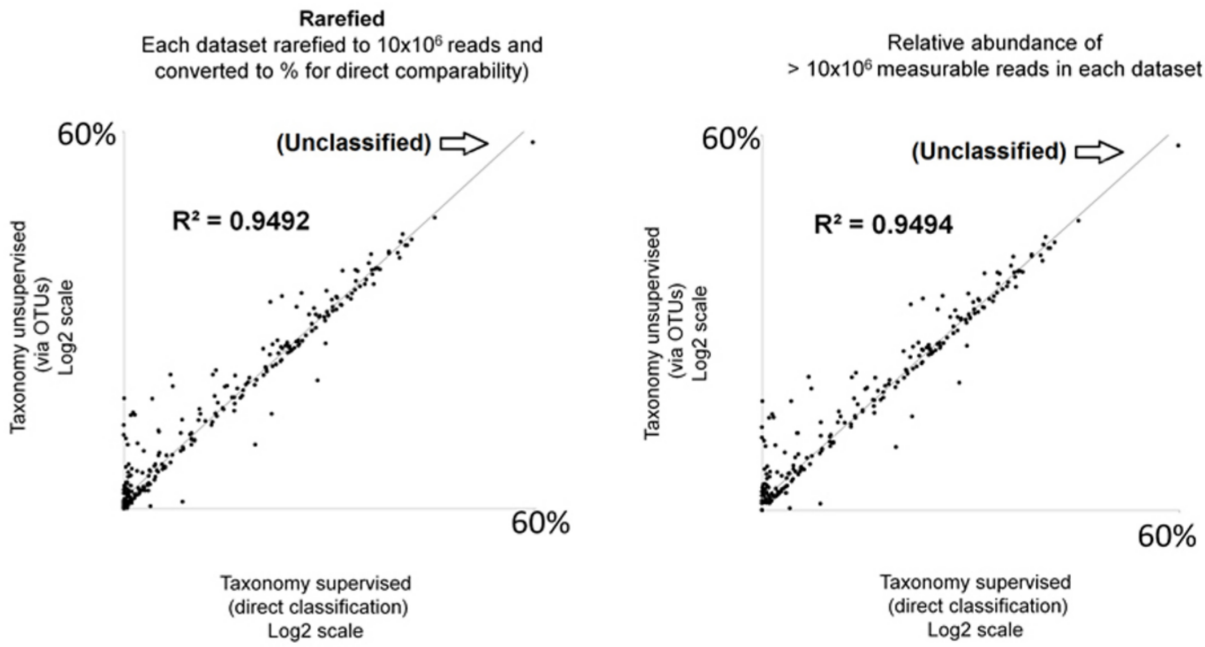


Figure 2.17: **16S taxonomy classification at the family level is robust to method**. Correlation between taxonomy-supervised (X-axis) and taxonomy-unsupervised (Y-axis) family-level abundances, for both the rarefied (left) and relative abundance (right) normalized data.

against Greengenes (Feb. 4 2011 version) training set to get family-level taxonomy. The abundance of each family was compared to the abundance of that family when the family assignments were assigned after the taxonomy-unsupervised grouping of reads into OTUs. Fig. 2.17 shows the total reads from non-chloroplast families from both taxonomy-supervised (X-axis) and taxonomy-unsupervised (Y-axis), for both the rarefied and relative abundance normalized data. The scatterplots thus show the high correlation at the family level for supervised and unsupervised taxonomy assignment. The dataset used for this figure included extra samples not described here, and was clustered as a single .fasta using the default QIIME implementation of Uclust (Kunin and Hugenholtz, 2010).

Once OTUs were assigned a taxonomy, all OTUs annotated as chloroplasts, Viridiplantae or Archaea by any of the methods were removed from the OTU table, resulting in the set of *usable* OTUs.

We pooled usable reads from each bulk soil and rarefied to 200,000 reads per soil; this was

permuted 100 times. We observed a median of 9,709 OTUs in MF soil and 9,897 OTUs in CL soil. Rarefaction curves to 200,000 reads in each bulk soil (not shown) indicated that, even at 200,000 reads, we were not capturing the entire community in either soil. Consequently, the total number of OTUs we report for our bulk soils may be lower than that found in some reports aimed at finding the true microbial diversity in soils.

A handful of samples had been sequenced more than once, over more than one 454 half-plate (for example to increase the read depth from problematic samples). These duplicated samples were pooled into a single sample by adding the unnormalized counts in the OTU table, and the resulting column was renamed to reflect the pooling that took place. Next any sample that had fewer than 50 *usable* reads was discarded, resulting in the unnormalized *usable* OTU table. At this point, both a frequency table and a rarefied table (1,000 *usable* reads per sample) were created as alternative normalization techniques.

The frequency table was made from the unnormalized *usable* OTU table by dividing the number of reads for each OTU in a given sample by the total number of reads in that sample and multiplying by 100, and repeating this across all samples.

We also created a rarefied table; because some samples, particularly samples from the EC, had fewer than 1,000 usable reads in the unnormalized usable OTU table, counts from independent samples sharing the same soil type, genotype, fraction, age and experiment were pooled to make groups of at least 1,000 reads, and the sample names were changed to reflect the pooling that had taken place (Rarefaction_MappingFile in Supplementary Database 1). Then all samples were rarefied to 1,000 counts using the rrarefy() function in the vegan package of R (Oksanen et al., 2014).

We present both methods because each has advantages and limitations. The advantage of the frequency table is that it keeps each individual plant separate, contains more individual samples and uses all of the data, but this comes at the cost of increased granularity in the normalized relative abundance percentages for some of the samples with fewer reads, causing problems with direct comparability. The major advantage of the rarefied table is

that comparisons are not biased by sampling depth and all read counts have equal weight, but this comes at the cost of reduced sample number and samples that mix information from several replicated individuals because we needed to pool some of our samples to meet our rarefaction threshold, and also at the cost of higher overall granularity because we discarded many reads from more deeply sequenced samples.

Because the majority of OTUs were represented by a very small number of reads and these OTUs were not technically reproducible (2.14d-e), both the rarefaction-normalized and the frequency-normalized OTU tables were thresholded to generate *measurable* OTUs for the majority of analyses (the major exception being the UniFrac analysis in Fig. 1: weighted UniFrac distance is robust to rare OTUs). An OTU was deemed measurable if and only if there were $\geq$25 reads in $geq5$ samples in the unnormalized usable OTU table. As described in the text and Fig. 2.14, this threshold was derived from the fact that the correlation between abundance in the same OTU in technical replicates improved greatly as OTUs approached an abundance of 25 reads, and from the fact that although contamination might create an OTU at this abundance once, the probability of an OTU being spurious decreases greatly if it occurs at a measurable level in several (we chose $geq5$) independent samples.

2.1.12 Detection of differentially enriched OTUs by the GLMM

The OTU abundances were analysed with a GLMM to estimate the effect of the different variables on each measurable OTU. The lme4 R package (Bates, 2010) was used to fit the model. The abundance of each OTU on each sample (y_{ij}) was $\log_2$-transformed and modelled as a function of the abundance of the same OTU in bulk soil samples (std_check) as a fixed effect, and plant genotype (b_1), sample type (plant or bulk soil, b_2), plant developmental stage (b_3), soil type (b_4), sequencing half-plate (b_5) and biological replicate (b_6) were modelled as random effects. The full model is specified by:

$y_{ij} = \beta * std.check + b_{1ij} + b_{2ij} + b_{3ij} + b_{4ij} + b_{5ij} + b_{6ij} + e_{ij}$

where e_{ij} is the residual error and std_check was calculated as the mean abundance of each OTU in all the bulk soil samples from each combination of experiment and developmental

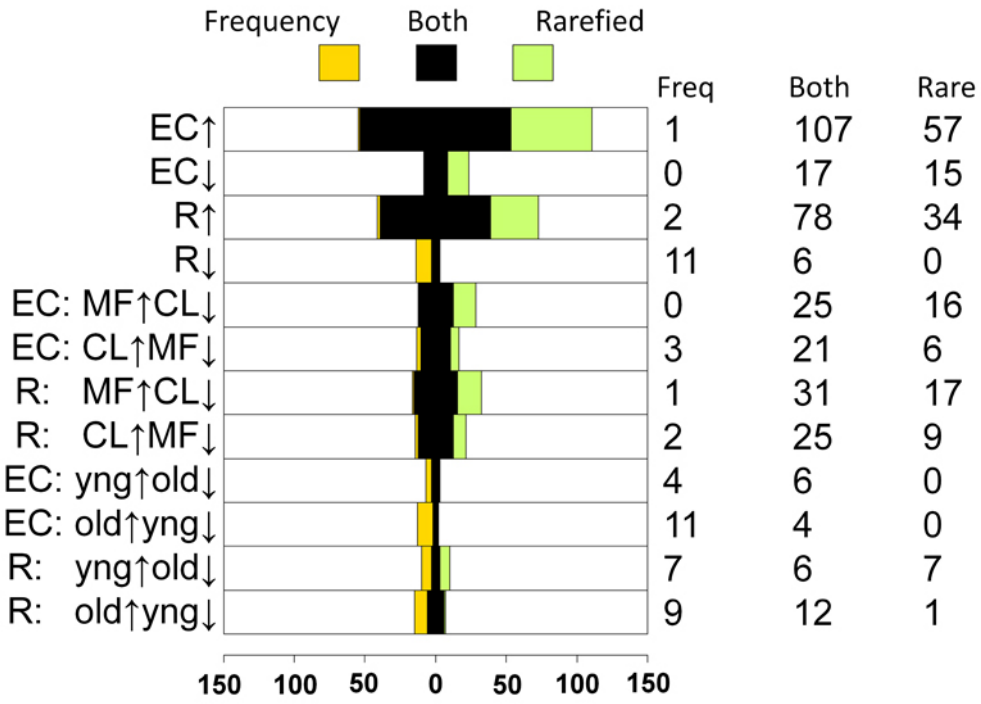


Figure 2.18: **Overlap of GLMM predictions between rarefaction-normalized and frequency-normalized OTU tables**. The number of OTUs predicted by the full GLMM in each category that are unique to the frequency table is shown in orange. The number of OTUs predicted by the full GLMM in each category that are unique to the rarefied table are shown in green. The number of OTUs that were shared predictions in the two tables is shown in black.

stage.

There were not enough paired samples of rhizosphere and EC from the same individual plant to model the effect of both fractions directly. Instead, the abundance table was split into EC and rhizosphere samples, and the effect of each fraction with respect to bulk soil controls was estimated. The same model specification was used independently on both fractions, and for both the frequency and the rarefied tables (methods 2.1.11). The percentage of total variance explained by each random variable on the OTU abundances is reported in Supplementary Table 5.

For each level of the random effects, the conditional mode and 95% prediction interval were estimated by Markov chain Monte Carlo sampling from the fitted model. A specific level is considered to have an effect on an OTU if the prediction interval of its conditional mode

does not include zero. OTUs detected this way are reported in Supplementary Database 3.

2.1.13 Partial GLMM

There were not enough samples to estimate all the interaction effect between all variables without drastically reducing the size of the data set and our statistical power (Supplementary Table 2). To assess specific interactions of the genotype effect with other variables, a constrained version of the previously defined GLMM was used that employed only the fixed effect (std_check) and the random effects for plant genotype (b_1) and sample type (b_2). Samples were split into groups of the same experiment, developmental stage and fraction (thus, all the other variables from the full model are tested within each group), and the model was fitted and analysed in the same way as the full GLMM. A non-parametric KruskalWallis test was used to verify independently the predictions of the partial GLMM for significance, where p-values were corrected to q-values using th$q - value > 0.05$ were discarded as insignificant. The intersection of the significant genotype predictions between both biological replicates of each condition was calculated. The intersection analysis from the partial GLMM is displayed in Supplementary Table 3.

2.1.14 Scanning electron microscopy sample preparation

Arabidopsis roots were fixed in 2% paraformaldehyde, 2.5% glutaraldehyde and 0.15M sodium phosphate buffer, pH 7.4. The samples were dehydrated using a gradual ethanol series (30%, 50%, 75%, 100%, 100%) and dried in a Samdri-795 supercritical dryer using carbon dioxide as the transitional solvent (Tousimis Research Corporation). Roots were mounted on aluminium planchets with double-sided carbon adhesive and coated with 10nm of goldpalladium alloy (60:40 Au:Pd, Hummer X Sputter Coater, Anatech USA). Images were made using a Zeiss Supra 25 FESEM operating at 5kV and a working distance of 5mm, and with a 10 μm aperture (Carl Zeiss SMT Inc.), at the Microscopy Services Laboratory, Pathology and Laboratory Medicine, UNC at Chapel Hill.

2.1.15 Log_2 transformation

All $\log_2$ transformations on OTU tables followed the formula $log_2(1000x + 1)$, where x is the rarefied read counts (or frequency) per OTU.

2.1.16 Heat maps

Heat maps were constructed using custom scripts and the function `heatmap.2()` from the R package gplots (Warnes et al., 2016). For better visualization, all data was $\log_2$-transformed (methods 2.1.15. Hierarchical clustering of rows and columns in the heat maps is based on BrayCurtis dissimilarities and uses group-average linkage.

2.1.17 Diversity

The Shannon diversity index and the non-parametric Chao1 diversity were calculated with the vegan package in R (Oksanen et al., 2014). The exponential function was applied to the Shannon diversity index to calculate the true Shannon diversity (effective number of species).

2.1.18 Rarefaction curves

Rarefaction curves were made with custom scripts that sampled each sample fraction only once at each read depth. To reveal the variance in sampling, no attempt was made to smooth the curves by taking the average of repeated samplings.

2.1.19 Taxonomy histograms and statistics

Taxonomy histograms were created using custom scripts and visualized in GraphPad PRISM version 5.0 for Windows (Motulsky, 2003) (GraphPad Software, Inc.; `http://www.graphpad.com`). The 'low-abundance' category was created to help remove visual clutter, and contained any taxonomic group that did not reach at least 5% in any one fraction. The Shannon diversity index was calculated as described in section 2.1.17. Differences in distribution at varying taxonomic levels, and differences in Shannon diversity between soil, rhizosphere and EC fractions, were tested by weighted analysis of variance (to account for differing numbers of soil, rhizosphere and EC samples), invoking the central limit theorem (>60 samples in each group in all tests for both frequency-normalized and rarefaction-normalized tests). For more details about tests, see additional notation in Supplementary Table 5.

2.1.20 Sample clustering using UniFrac

A phylogenetic tree was built with the representative sequence for each OTU and the pairwise, normalized, weighted UniFrac distance (Lozupone and Knight, 2005). For UniFrac,

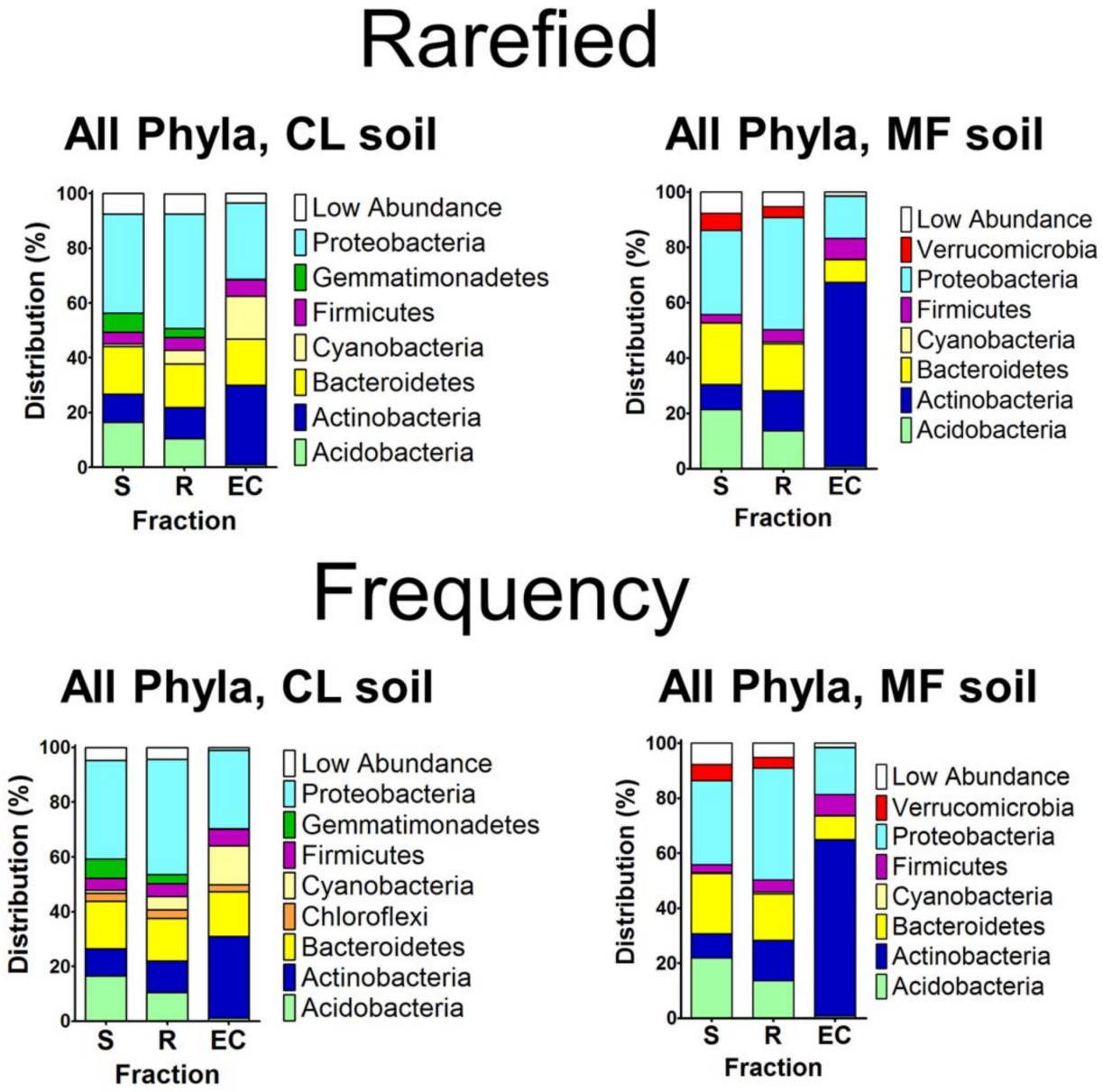


Figure 2.19: **Phyla in each sample fraction by soil type**. Histogram displaying the distribution of the phyla present in the 778 measurable OTUs in soil (S), rhizosphere (R) and endophytic compartments (EC) with each soil type, MF and CL, considered independently. Rarefaction-normalized on top; frequency-normalized on bottom. Accompanying statistics on the distributions are in Supplementary Table ST5.

representative sequences from all non-plant OTUs, including those that did not meet the *255* sample threshold, were considered. UniFrac distances between samples are based on the fraction of branch length that is unique to each sample in a shared phylogenetic tree composed of OTU representative sequences from all samples. Thus, samples containing OTUs of highly divergent sequences will be more distant from each other, because the OTUs comprising each sample will occupy different major branches on the shared phylogenetic tree of OTUs, whereas samples containing highly similar OTUs will share these major branches. In weighted UniFrac, the branch length unique to each sample is multiplied by the frequency at which that OTU occurs in the sample. Thus, weighted UniFrac can detect differences between two samples that have the same set of OTUs that differ quantitatively between the samples.

Principal coordinate analysis was performed using pairwise, normalized, weighted UniFrac distances between all samples on the unthresholded but normalized OTU tables, and the first two principal coordinates of UniFrac were visualized with GraphPad PRISM version 5.0 for Windows.

2.1.21 CARD-FISH application to roots

We applied a modified protocol described previously (Eickhorst and Tippkötter, 2008). Briefly, several root systems from a bolting Col-0 grown in MF were fixed using 4% formaldehyde in PBS at 4°C for 3h, washed twice in PBS and stored in 1:1 PBS:molecular-grade ethanol at 20°C. Treatments with lysozyme solution (1h at 37°C, 10mgml^{1}; Fluka) and achromopeptidase (30min at 37°C, 60Uml^{1}; Sigma) were sequentially used for prokaryotic cell-wall permeabilization. Endogenous peroxidases were inactivated with methanol treatment amended by 0.15% H_2O_2 at room temperature for 30min and washed again. Probes targeting either the 16S or the 23S rRNA (EUB338 (5'-GCTGCCTCCCGTAGGAGT-3', 35% formamide), NON338 (5'-ACTCCTACGGGAGGCAGC-3', 30% formamide), HGC69a (5'-TATAGTTACCACCGCCGT-3', 25% formamide) and Brady4 (5'-CGTCATTATCTTCCCGCACA-3', 30% formamide)) were defined using probeBase (Loy et al., 2007) (`http://www.microbial-`

ecology.net/default.asp), labelled with enzyme horseradish peroxidase on the 5' end (Invitrogen), diluted in hybridization buffer (final concentration of 0.19ngml[1]) with each probe's optimum formamide concentration, and hybridized at 35°C for 2h. Unbound probes were washed away from samples in wash buffer (NaCl content adjusted according to the formamide concentration in the hybridization buffer) at 37°C for 30min. Fluorescently labelled tyramide was used for signal amplification, and samples were washed before mounting on glass slides.

For double CARD-FISH, a subset of samples went through a second round of the protocol, starting at the peroxidase inhibition with a second variety of fluorescently labelled tyramide used to be able to distinguish the signals from each probe. Roots were mounted on glass slides using Vectashield with DAPI (Vector Laboratories, catalogue no. H-1200) for mounting solution, and sealed with nail polish for storage. All microscopy images were made on a confocal laser scanning microscope (Zeiss LSM 710 META) located in the Biology Department at UNC. The Brady4 probe, which has not been used for this application previously, was tested on filters of cultured Bradyrhizobiaceae and three negative control cultured strains to determine the most specific formamide concentration in the hybridization buffer.

For application of samples onto filters, bulk MF soil, rhizosphere and EC samples from four sets of Col-0 roots were pooled and harvested in the way described above before DNA extraction. Samples were then fixed as described above and passed through a 10 μm filter. The concentrations of plant material were made equal and samples were sonicated in a water bath for 5min. The sample suspension was further diluted to 1:500 in water and applied to a 25-mm polycarbonate filter with a pore size of 0.2μm (Millipore) using a vacuum microfiltration assembly. Filters were embedded in 0.2%, low-melting-point agarose and dried, and CARD-FISH was applied as described above. For quantification of bacteria, filters were visualized on a Nikon Eclipse E800 epifluorescence microscope. Positive EUB338 probe signals that co-localized with a DAPI signal were counted as Eubacteria. Positive Actinobacteria or Bradyrhizobiaceae signals were counted as positive when the HGC69a or Brady4 probe co-localized with both EUB338 and the DAPI signal. To estimate the number of active

bacteria present per sample, the number of EUB positive signals co-localizing with a DAPI signal was counted and the number of EUB positive signals per sample was calculated.

2.1.22 Sample naming in OTU tables

All sample names in OTU tables are in the following form: [soil type].[genotype].[sample number][fraction].[age].[experiment]_[plate]. For example, M21.Col.6E.old.M1_2b should be interpreted as [soil type] = M21 = Mason Farm 2:1, [genotype] = Col = Col-0, [sample number] = 6, [fraction] = E = endophyte compartment, [age] = old, [experiment] = M1 = Mason Farm replicate 1, [plate] = 2b.

CHAPTER 3

A reduced complexity platform for a complex system[1]

It is well established that plants assemble a distinct microbiome in and around the root (Lundberg et al., 2012; Bulgarelli et al., 2012; Schlaeppi et al., 2014) (Chapter 2), and in above ground organs (Bodenhausen et al., 2013; Horton et al., 2014; Maignien et al., 2014). At the same time, there is evidence from the Brassicaceae and Poaceae families that host phylogenetic distance correlates with microbiome composition differences across species (Schlaeppi et al., 2014; Bouffaud et al., 2014). Evidence indicates that the within-species root microbiome differences are statistically significant but small in magnitude across a variety of species: bacterial community profiles in and around the roots of *A. thaliana* wild accessions in natural soil showed that only a handful taxa displayed genotype-dependent differences (Bulgarelli et al., 2012; Lundberg et al., 2012); similarly, another study found that differences between accessions were restricted to a subset of Pseudomonadaceae bacteria (Haney et al., 2015); among other species, barley rhizosphere microbial communities showed taxonomic and functional differences that might be related to domestication and explained ∼5% of the microbiome variation (Bulgarelli et al., 2015); and maize rhizospheres of 27 modern inbreds across sites exhibited small proportion of heritable variation in total bacterial diversity across

[1]The contents of this chapter has not been peer reviewed. This chapter describes the work performed to develop, implemente and establish the synthetic community approach in the dangl lab, as well as its application to novel questions. Besides myself (Sur Herrera Paredes), multiple people in Jeff Dangl's group and will be recognized with authorship when some or all of this work is published. People that contributed include but are not limited to: PhD student Derek Lundberg, and undergraduate students/research technicians Meredith McDonald and Surojit Biswas. The specific contributions are as follow: SHP, DL and JD designed the experiments. SHP, DL, SB and MM performed the experiments and collected samples. SHP, DL and SB obtained the sequencing data. SHP, DL and JD analyzed the data. SHP wrote the manuscript with input from JD.

fields, and substantially more heritable variation between replicates of the inbreds within each field (Peiffer et al., 2013).

The small genotype-dependent root microbiome differences between natural accessions is in stark contrast with the differences observed in the above-ground (phyllosphere) microbiome. Field surveys of tree phyllosphere bacterial communities has revealed a stronger effect of tree species than sampling site or time (Redford and Fierer, 2009; Laforest-Lapointe et al., 2016). At the same time, a field study of *A. thaliana* wild accessions, showed sufficient genotype-dependent patterns to perform Genome Wide Association (GWA), and identified plant loci related to defense, cell wall integrity, trichome/cuticle synthesis and morphogenesis as relevant determinants of bacterial and fungal community assembly (Horton et al., 2014). Another large-scale field experiment in *Boechera stricta* (Brassicaceae) grown in multiple sites through its natural range, simultaneously profiled leaf and root bacterial communities and found a strong signature of host control on the leaf microbiome that was absent in roots (Wagner et al., 2016).

The difference in genotypic signatures between rhizosphere and phyllosphere might indicate that bacterial communities in and around the root are more dependent on microbe-microbe competition and microbial adaptation to the host-associated environment. Alternatively, it might also mean that the host selection occurs at a level that is beyond the resolution of typical microbiome profiling methods, which typically target a single marker gene and thus miss the microbial genomic context. Previous work has showed that strains of the same *Pseudomonas fluorescens* ribotype can differentially associate with *A. thaliana* accessions with consequences for plant fitness (Haney et al., 2015). Full metagenomic sequence could potentially overcome this problem by providing a full taxonomic and functional picture of the root microbiome; however, significant experimental and analytical challenges limit the utility of this approach. For instance, there is no high throughput method to physically separate bacterial and plant host DNA prior to library preparation, meaning that almost all the sequences recovered derive from the host. At the same time, metagenomic assembly

SBS4

MEDIA		GENOTYPE Col-0	Cvi-0	Oy-0	Sha-0	*C. rubella*	Bd21
¼ MS	Soil						
	N	14	5	7	6	5	5
	Root	14	5	7	4	4	5
1/25 MS	Soil						
	N	11					
	Root	8					
Johnson	Soil						
	N						
	Root						
LowN	Soil						
	N	14					
	Root	13					
LowS	Soil						
	N	11					
	Root	11					
LowP	Soil						
	N	11					
	Root	11					
Johnson LowP	Soil						
	N						
	Root						

SBS5

MEDIA		GENOTYPE Col-0	Cvi-0	Oy-0	Sha-0	*C. rubella*	Bd21
¼ MS	Soil	10					
	N	13					
	Root	11					
1/25 MS	Soil	3					
	N	11	5	4	5	4	5
	Root	9	5	4	5	4	5
Johnson	Soil						
	N	15					
	Root	14					
LowN	Soil	10					
	N	5					
	Root	4					
LowS	Soil	2					
	N	11					
	Root	10					
LowP	Soil	8					
	N	7					
	Root	7					
Johnson LowP	Soil						
	N	15					
	Root	14					

Figure 3.1: **Experimental design and sample number**. We tested a number of different hosts with various degrees of genotypic divergence in one media, and one host in several media in two independent experiment (left and right). Between the first and second experiment we changed to a more nutrient limiting environment to see if the more stressful conditions would reveal stronger genotypic differences. We harvested roots and neighboring soils (N) in both experiments. In the second experiment (SBS5, right) we also harvested unplanted pots (soil). We also added Johnoson media and a phosphate dropdown on this media (Johnson LowP) to determine its similarity to results on MS media. Numbers indicate number of samples that passed all quality control steps and were used for final analysis.

of complex environments is an open bioinformatics problem, with state of the art methods typically only assembling ∼10% of the data. We decided to take an approach based on microcosm reconstitution, by inoculating seedlings — growing in a calcined clay substrate — with a well-defined but complex synthetic bacterial community (SynCom), while varying either the nutritional composition of the soil, or plant host (Fig. 3.1). This approach allowed us to disentangle changes in bacterial community composition that are due to microbial adaptation to abiotic changes in the environment, and changes that are due to the action of the plant-host and are accessible to natural selection.

3.1 Robust re-colonization of *A. thaliana* roots across nutritional conditions

We first asked whether there is qualitative differences between the communities that assemble in the bulk soils, neighboring soils and roots when media is diluted, or specific

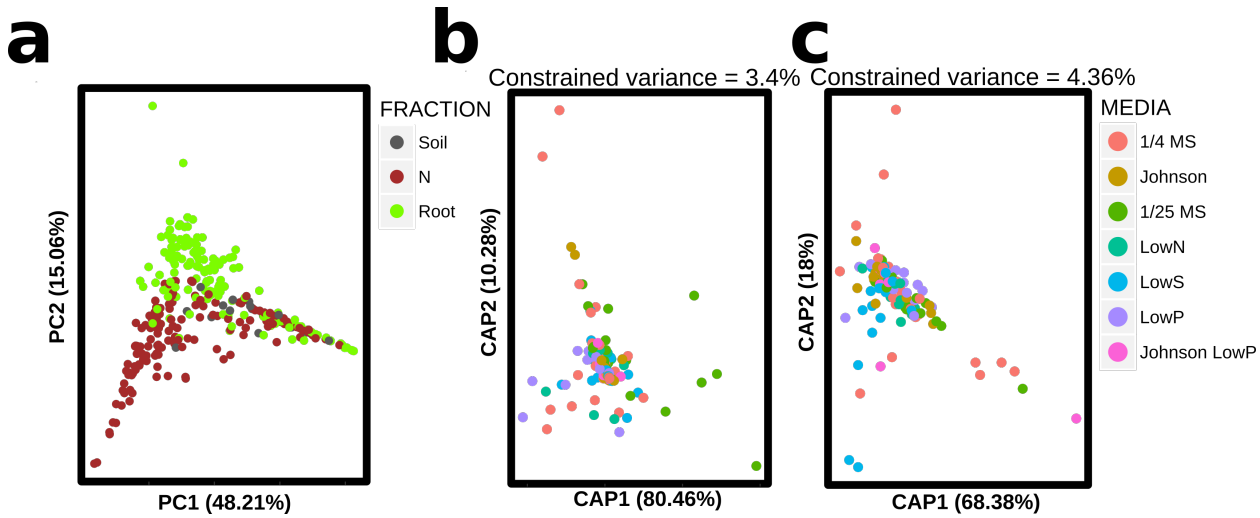


Figure 3.2: **Nutritional composition alters root colonization**. **a** PCA of all samples. **b** Canonical analysis of principal coordinates (CAP) of bacterial communities in neighboring soil (N) samples. Constrained variance indicated on top is the effect on community variation that is attributable to media in the neighboring soil, after conditioning for other variables. **c** CAP analysis of bacterial communities from root samples. Constrained variance indicated on top is the effect on community variation that is attributable to media in the plant roots, after conditioning for other variables.

nutrients (Nitrogen, Phosphorous, Sulfur) are dropped down. We found a surprising level of consistency in the communities that assemble in all conditions (Figs. 3.2 and 3.3). While ordination methods showed a separation between soil and root samples (Fig. 3.2a), and a slightly stronger effect of media on community composition inside the root (4.3% of the variance) than in the neighboring soil (3.4% of the variance) (Fig. 3.2b-c), we found a surprisingly consistent level of bacterial colonization across all media types (Fig. 3.3) (Fig. 3.3; section 3.7.7).

We then asked whether there are specific quantitative differences in colonization between different sample fractions (soil, neighboring soil and roots) and nutrient compositions. We identified those differences with a regularized logistic regression model (section 3.7.8). We found a similar number of differences between sample fractions (28 strains), nutritional composition (23 strains) and the interactions between the two variables (30 strains), but the effect of the interactions tended to be stronger (Fig. 3.4). Most of the differences due to sample fraction where characterized by a higher abundance in neighboring soils than in bulk

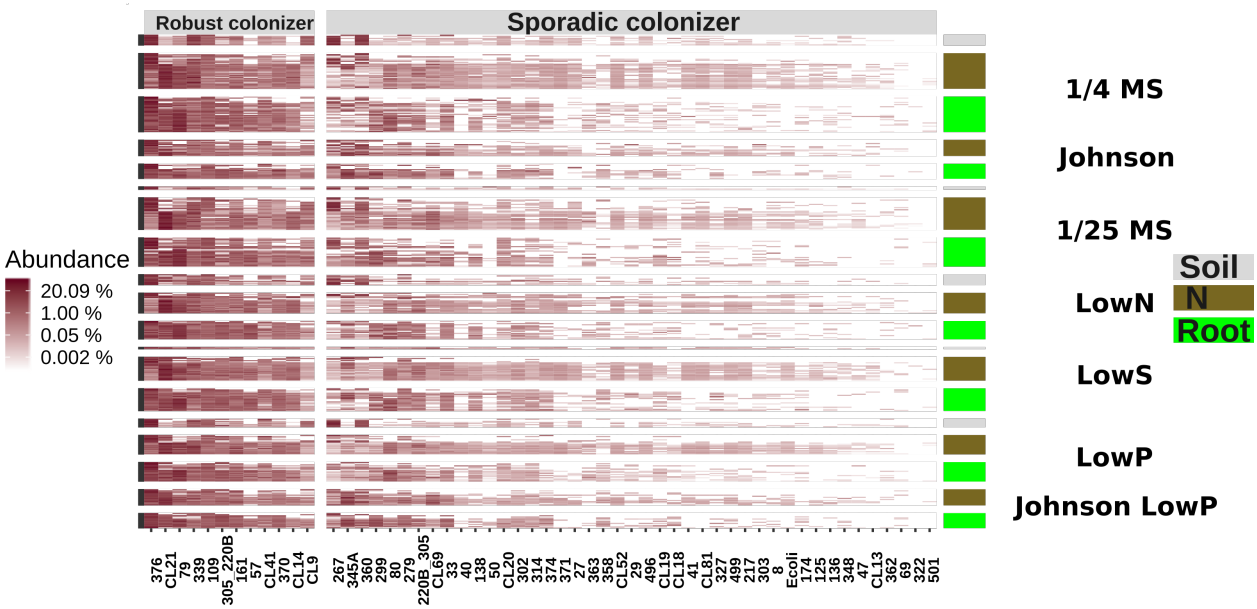


Figure 3.3: **Root microbiome in varying nutrient compositions**. Heatmap showing bacterial percent abundances in different nutritional compositions, and sample fractions. Individual strains are shown as columns, and robust and sporadic colonizers are indicated in the top. Each row is an independent sample, with media composition indicated by the colored panels on the right. Within each block (combination of media and fraction), samples are sorted by independent experiment.

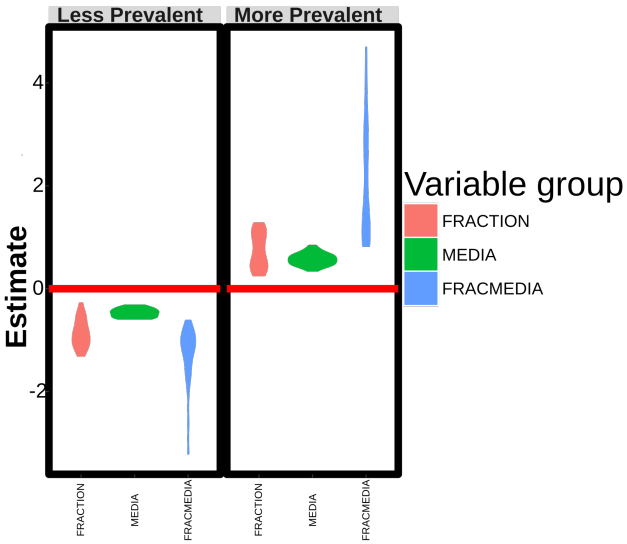


Figure 3.4: **Media has a small effect on bacteria presence/absence but a larger effect inside the root**. Violin plots showing the effect sizes of different variables on the presence/absence of individual strains. Y-axis shows the coefficient estimate for the corresponding variable, and can be interpreted as the log fold-change in abundance between a samples in different fractions, media or combination of fraction and media (FRACMEDIA). Only significant (q-value < 0.05) effects are plotted.

soils or in the root. The two strongest examples are shown in Fig. 3.5a-b. Only five out of twenty eight strains that showed fraction level differences were characterized by a higher abundance in the root than in the neighboring soil, and their effect tended to be weaker. Fig. 3.5c shows prevalence patterns for strain 299, which is more prevalent in roots of all media than in the corresponding general soil.

Soil nutritional composition had the fewer (23 strains) and in general weaker effect over microbial prevalence (Fig. 3.4), and the majority of their effect was due to differences in the low S (LowS) media (16 strains, including top 2), followed by the highly diluted media (1/25 MS, 10 strains) and the low P (LowP) media (8 strains). Low S mostly increased the prevalence of specific strains (only 1/16 showed decreased prevalence in this media), but the effect of media is comparable in magnitude with the weaker fraction level differences (Fig. 3.5d). An example of this is strain CL21, which has a higher prevalence in LowS samples than in other media, and is the most strongly affected strain by any single media. Interestingly, this increase in prevalence of multiple strains caused by low S levels is not

caused by an overall change in bacterial richness, since this parameter showed no variation between nutritional compositions (Fig. 3.5).

Different sample fractions represent different micro-environments available to bacterial strains. It is interesting to ask whether samples from specific combinations of fraction and nutritional composition, lead to different colonization patterns. We found that 30 strains showed prevalence differences in 67 specific cases. We found a fairly even distribution of significant prevalence difference (21 cases in bulk soil, 21 cases in neighboring soils and 25 cases in roots), though the strongest effects tended to be in bulk soil samples (all of the top 18 cases, while all 21 bulk soil cases are in the top 30 cases). For example, strain CL52 is less abundant than expected in the roots of plants grown in LowS, 1/4 MS and 1/25 MS, but shows no significant differences in other cases (Fig. 3.5f). Another case is strain 327, which is absent from roots grown on LowP media, despite being quite prevalent in the neighboring soils of the same media, and in the roots of plants grown in other media (Fig. 3.5e).

3.2 Robust re-colonization of roots across host phylogenetic distance

We also asked whether divergent hosts showed similar bacterial colonization patters. We evaluated the presence/absence profiles of bacterial isolates in the roots and neighboring soil of four *A. thaliana* accessions (Col-0, Cvi-0, Oy-0 and Sha-0), a related Brassicaceae species (*Capsella rubella*) and a highly divergent monocot from the Poaceae family (*Brachypodium distachyon* Bd21). Principal component analysis and constrained ordination showed similar patterns to those observed across nutritional composition, with root and neighboring soils forming separate groups and a slightly stronger effect of host genetics inside the root (5.46% of variance) than in the surrounding neighboring soils (2.29% of the variance) (Fig. 3.12; section 3.7.6). Notably, no big differences were observe between *A. thaliana* and the other two plant species tested, despite the large evolutionary distance between them. In fact, the variation due to genotype of those two species falls within the *A. thaliana* variation, and isolates that were able to colonize *A. thaliana* Col-0 roots were —in general— also able to colonize the other hosts (Fig. 3.7).

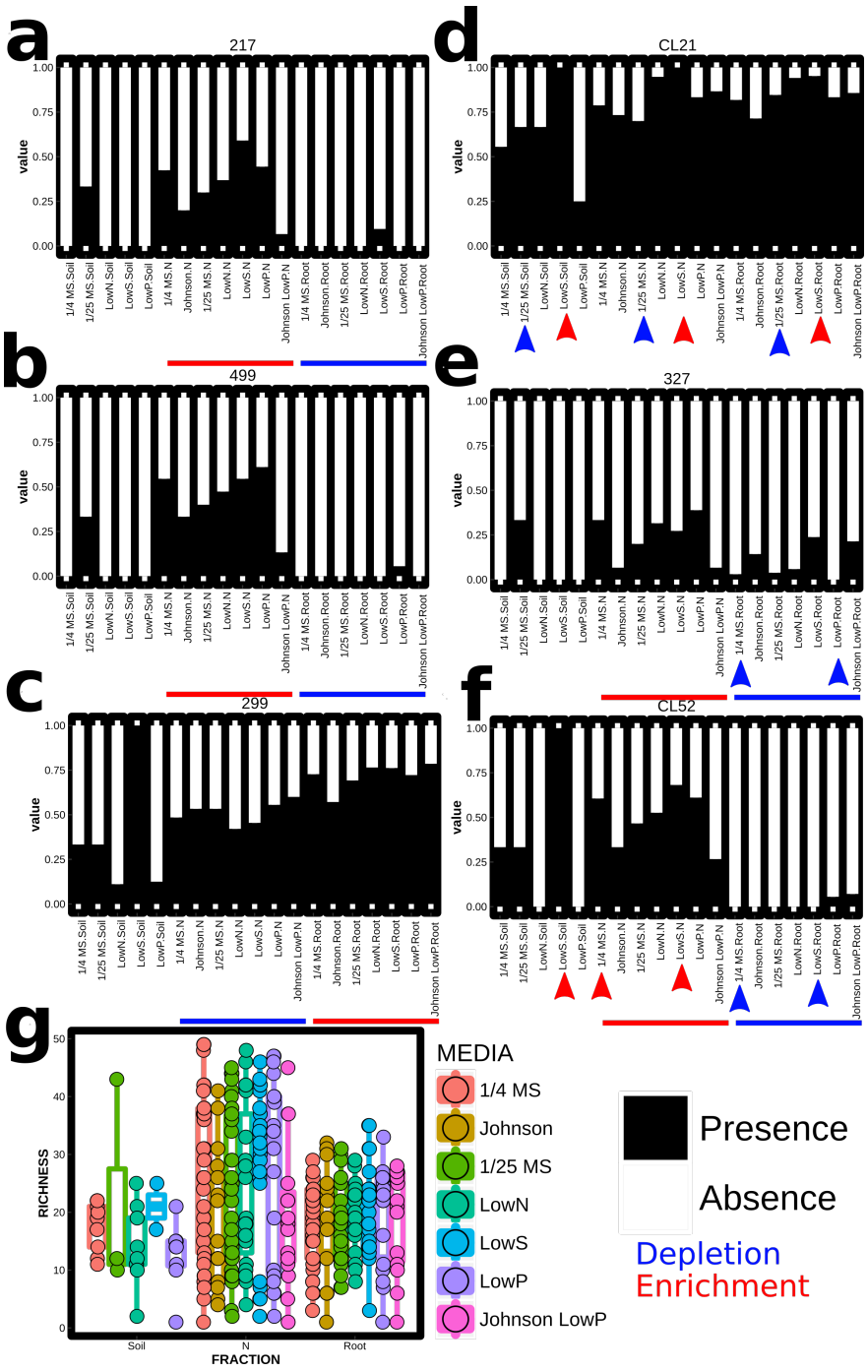


Figure 3.5: **Examples of isolates that show presence/absence differences in different fractions and conditions**. In **a-f**, each bar represents the samples from a given media and fraction, and the black portion of the bar represents the proportion of samples from that group in which a particular strain was present. Red and blue bars below each plot show fraction differences, while arrows indicate differences between specific groups of samples. **g** Shows bacterial richness (number of different strains found in a sample) in the same groups of samples.

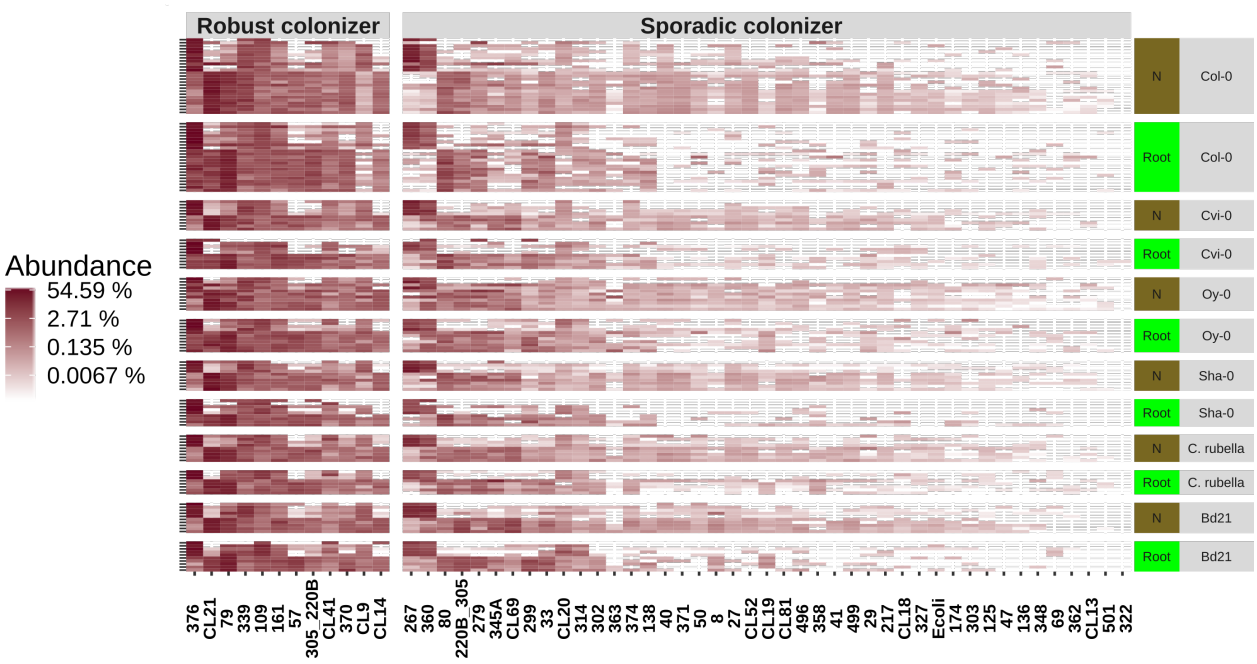


Figure 3.6: **Root microbiome in different hosts**. Heatmap showing bacterial percent abundances in different nutritional compositions, and sample fractions. Robust and sporadic colonizers are indicated in the top. Samples are sorted by experiment within each block.

Differences between neighboring soil and root samples (19 strains) were largely consistent with the ones observed between nutritional compositions (data not shown), and were similar in magnitude to those of host genotype, but slightly weaker than those specific for a combination of fraction and genotype (Fig. 3.13). Only a handful of genotypic differences were detected that were consistent across sample fraction (9 strains), the majority of which involved Col-0 (6 strains) though they accumulated towards the weakest effects. We also looked for prevalence differences between host genotypes that were specific to a particular fraction. As expected the majority of these differences (14/21) were between root samples and involved a total of 16 strains.

One of the strongest effects involved the divergent monocot host *B. distachyon* Bd21. Strain 8 had higher prevalence neighboring soils than in roots, but was more prevalent in and around *B. distachyon* roots (Fig. 3.7a). Strain 29, on the other hand, showed no differences between sample fractions, but was systematically absent from Sha-0 samples regardless of the fraction (Fig. 3.7b). Strain 496 was enriched in the neighboring soils with respect to the

root, but is still found in roots, except for those of Col-0 and Oy-0 (Fig. 3.7c).

Overall, we found a small but significant effect of both genotype and media in the presence/absence of bacteria in the roots and neighboring soils. These results suggest that bacterial colonization of plant roots is highly robust, and is mostly mediated by bacterial competition as different strains try to take advantage of the available niches. Interestingly, the strongest differences were not found on either the highly diluted media (1/25 MS), or the highly divergent host (*B. distachyon*); instead, *drop-down* of a specific nutrient (*i.e.* S) had the most differences in bacterial prevalence; and *A. thaliana* accession Oy-0 harbored a more rich bacterial community than other hosts (Fig. 3.7d), despite affecting a similar number of strains.

3.3 Specific changes in the root microbiome under different nutritional conditions

Besides differences in bacterial colonization, it is also possible that there are quantitative differences in bacterial relative abundance under different nutritional conditions. We used a Zero-Inflated Negative Binomial (ZINB) model to identify such differences (Lebeis et al., 2015) (section 3.7.9). We found 25 instances where there are significant relative abundances differences that can be attributed to an interaction between sample fraction and media (Table 3.1). These differences involved 11 strains and all media types, but strain CL21 and CL14 were the most sensitive to this parameter combination with seven and four instances, respectively, in which they were affected (Fig. 3.8). These two strains were relatively rare in bulk soil samples but are highly enriched in both neighboring soil and root samples across all media; but we were able to detect quantitative variation between the media types in specific fractions (Table 3.1). This is in stark contrast to what we observed in terms of presence/absence variation (section 3.1), where most of the differences between media types occurred among low prevalence isolates (Fig. 3.5e-f), with the notable exception of strain CL21 which showed variation in both prevalence (presence/absence; Fig. 3.5d) and relative abundance (Fig. 3.8; Table 3.1).

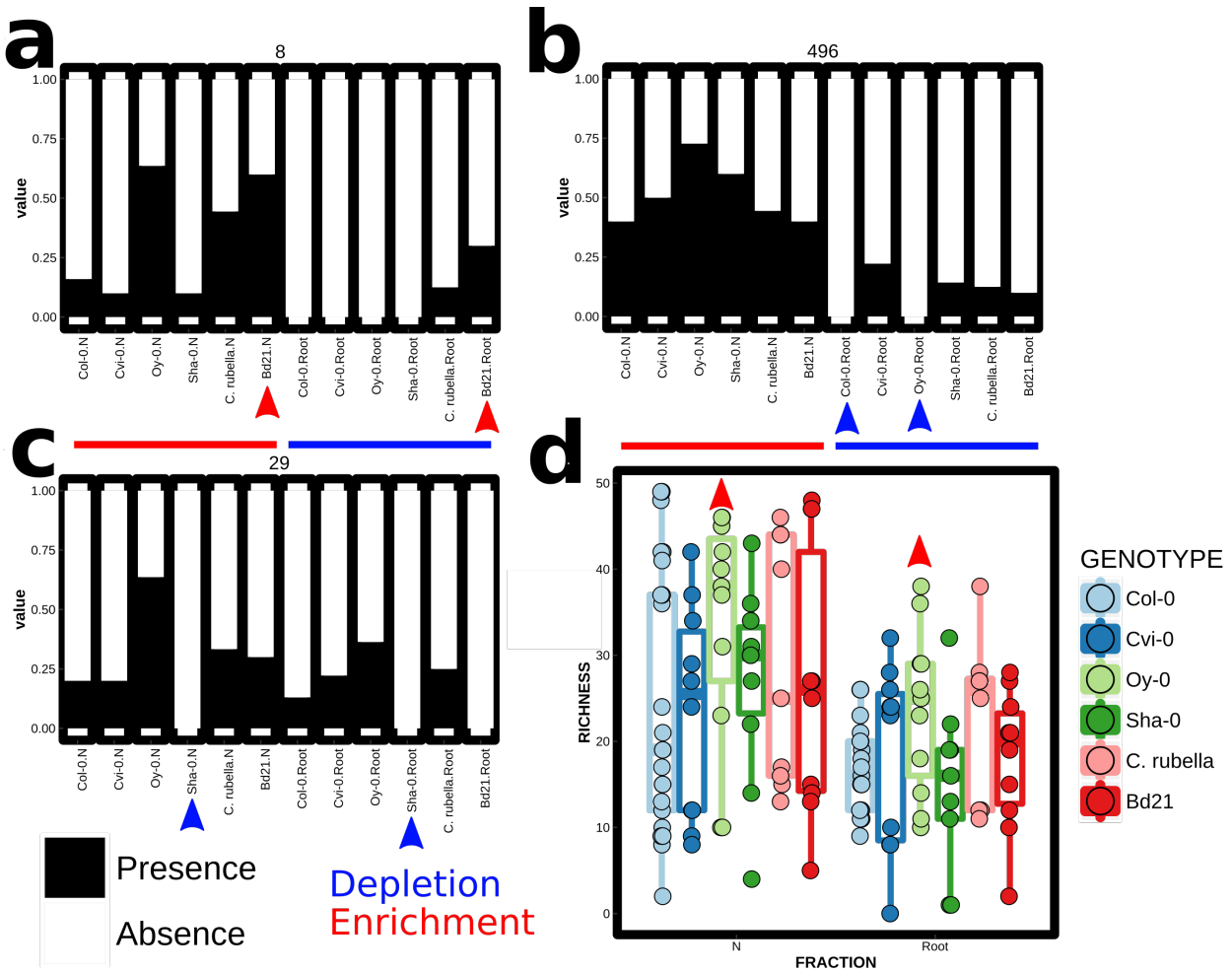


Figure 3.7: **Presence/Absence variation of isolates between hosts**. **a** Strain 8 is mostly excluded from roots, but it is present more often than expected in neighboring soils and roots of *B. distachyon* Bd21. **b** Strain 496 is present at a lower rate in the roots than in the soils, but it is absent from the roots of Col-0 and Oy-0. **c** Strain 29 is equally prevalent between neighboring soils and roots, but it is absent from both fractions of Sha-0 samples. **d** Bacterial richness (number of different strains detected in each sample) of samples of different fractions and genotypes. The only significant difference (after controlling for batch effects) is an increased richness in Oy-0 samples of both fractions with respect to the other genotypes (q-value < 0.05).

Taxon	**Variable**	**Estimate**	***q*-value**
CL21	LowN.Root	-6.408931	8.454272E-16
CL21	LowS.Root	-7.067537	1.329227E-09
CL14	LowN.Root	-5.774673	4.23655E-09
50	1/25 MS.Root	-8.837079	3.348696E-08
CL14	LowN.N	-5.289415	1.655905E-05
161	LowN.N	-3.251344	3.726209E-05
CL41	LowP.Root	-3.537116	0.000119682
CL14	LowS.Root	-4.907614	0.000119682
161	Johnson.Root	3.43132	0.000119682
345A	LowN.N	-5.570872	0.0003216634
267	LowS.Root	6.860572	0.0003895009
CL21	1/4 MS.N	3.142551	0.001023724
CL14	LowS.N	-4.760557	0.001350852
CL21	LowP.Root	-3.057111	0.001716946
CL9	LowN.Root	2.934064	0.002077016
161	1/4 MS.N	-2.409676	0.002077016
40	1/25 MS.Root	-5.556394	0.002293188
279	LowN.Root	-3.517204	0.002687698
CL21	Johnson.Root	-3.417778	0.003287743
CL21	Johnson LowP.Root	-3.296168	0.003745
CL21	LowN.N	-2.961913	0.003745
345A	LowS.N	-6.04034	0.006375684
371	LowS.N	-3.841529	0.006447609
345A	LowN.Root	-3.693198	0.007363134
40	LowN.Root	-4.643262	0.009988183

Table 3.1: Isolates that show quantitative variation in different media and sample fraction.

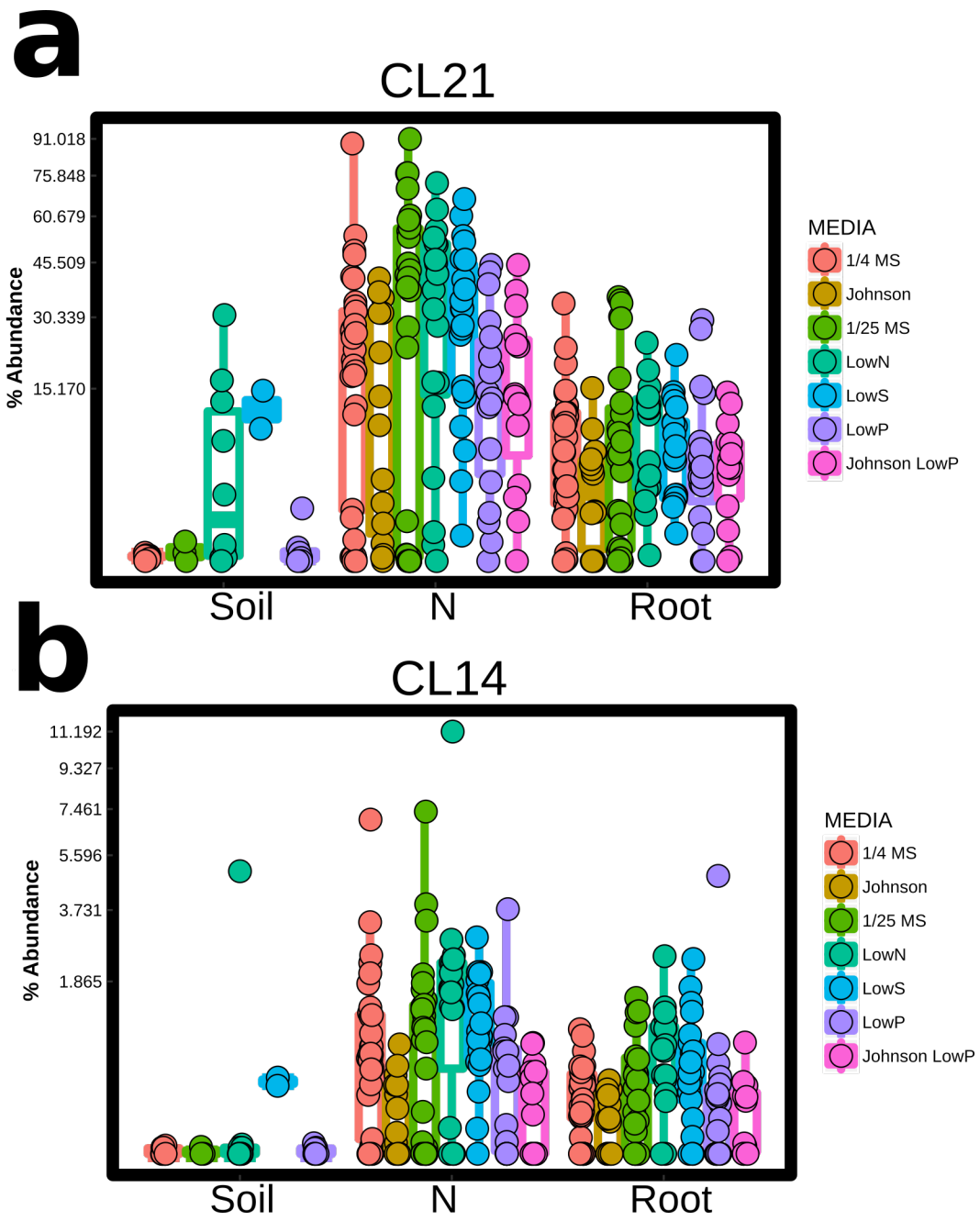


Figure 3.8: **Isolates sensitive to media and sample fraction**. Isolate CL21 (a) and CL14 (b) are both more abundant in neighboring soils and roots than in bulk soil samples, and show quantitative variation in relative abundances in different media. Results of statistical tests are in Table 3.1

Together, our results suggest that changes the nutritional conditions in the soil, lead to changes in the root environment and microbial community, that allow sporadic access to strains that would be normally excluded in the root (section 3.1; Fig. 3.5). Our observations can be explained by the fact that nutritionally challenged plants down-regulate defense (Castrillo et al., 2017; Yamada et al., 2016), and the observation that hypo-immune plants can be colonized by bacteria that are normally not able to do so (Lebeis et al., 2015). On the other hand, isolates that were highly abundant in the root microbiome under full nutritional conditions, continued to successfully colonize plant roots in other conditions, but at variable relative abundances (section 3.3; Fig. 3.5). This might be a result of a combination of bacteria-bacteria competition caused by the changes in colonization of low-prevalence members, and physiological changes in the plant host.

3.4 Specific changes in the root microbiome under different host genotypes

We also found a number of significant relative abundance differences due to plant host genotype. Interestingly, we found 13 instances in which the difference was consistently present in the neighboring soil and root samples, suggesting that the host either alters the surrounding environment strongly enough for it to mirror the root, or that bacterial colonization of the root provides a competitive advantage to those isolates, by maintaining a population that can expand into the neighboring soil. Consistent with host phylogeny, the largest number of differences from Col-0 were found with *B. distachyon* Bd21, with 5 instance; intriguingly the same number of differences from Col-0 were found with *A. thaliana* accession Oy-0, while another accession (Cvi-0) and a related Brassicaceae *(C. rubella*) showed only two and one differences, respectively. Figure 3.9a-b shows examples of two strains that with consistent differential abundances between a pair of hosts in both neighboring soils and root samples.

We have shown that the effect of host genotype is stronger in the root than in the neighboring soil (Fig. 3.12; section 3.2). Thus, we asked whether there are bacterial relative abundance differences that are specific to the root. We found nine such instances involving eight strains. We found the relative effect of Oy-0 to be even stronger than when looking for

Taxon	Variable	Estimate	*q*-value
371	Oy-0.Root	3.953721	4.214462E-06
109	Bd21.Root	1.367001	0.0001777078
27	Oy-0.Root	2.053737	0.0003340636
371	*C. Rubella*.Root	3.507121	0.0003340636
41	Oy-0.Root	3.340118	0.0003340636
CL52	Oy-0.Root	2.377527	0.001058683
CL69	*C. Rubella*.Root	3.508078	0.004553106
217	Oy-0.Root	1.936684	0.004871109
40	Oy-0.Root	2.167675	0.004871109

Table 3.2: Significant relative abundance differences for specific bacterial between hosts.

differences that were consistent across fractions. Six of the nine *genotype by fraction* specific differences involved Oy-0, two *C. rubella* and one *B. distachyon* Bd21 (Table 3.2). Two examples are showin in Fig. 3.9c-d. Overall, we observed that Oy-0 is the most distinct of the four *A. thaliana* accessions tested, while the variation of related Brassicaceae *C. rubella*, clearly falls within the *A. thaliana* variation, consistent with a previous report on natural soils (Schlaeppi et al., 2014). Moreover, the root microbiome of the highly divergent monocot grass *B. distachyon* also falls within this range of variation.

A common feature among the genotype specific differences in the root is that they involved low abundance strains (7/8 are sporadic colonizers; Fig. 3.6), were depleted in the root with respect to the neighboring soils, but were robustly enriched in the roots of a specific host (Fig. 3.9c-d). Genotype-dependent enrichment of low abundance isolates was also observed in at least some cases of isolates enriched in both fractions from a particular host (Fig. 3.9a-c). The predominance of enrichments observed in Oy-0 roots (Table 3.2) explain the increased bacterial richness that we measured in that accession (Fig. 3.7d). In summary, the enrichment of normally depleted or low abundance strains and the qualitative-like enrichment in specific hosts, points to a parallelism with *gene for gene* interactions which are responsible for disease resistance in plants (Flor, 1971), and suggest a simple underlying genetic architecture that could be amenable to genetic mapping.

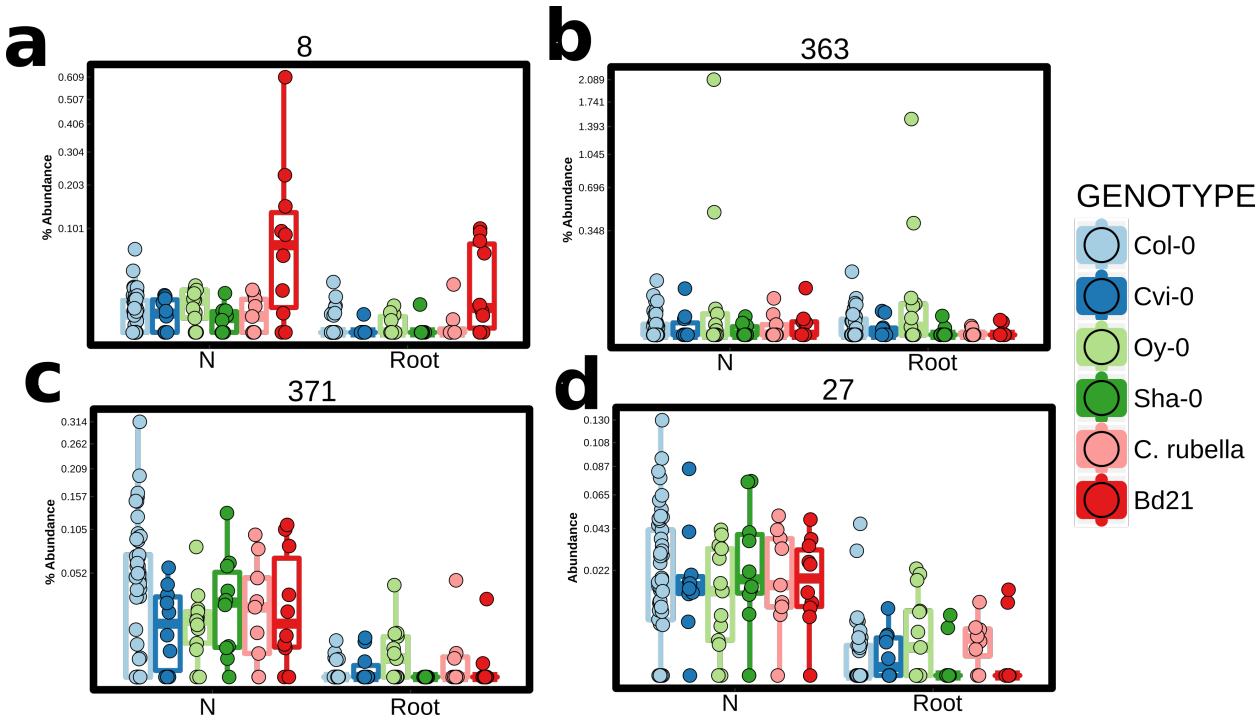


Figure 3.9: **Isolates enriched in specific hosts**. a) Strain 8 is more abundant in the neighboring soil than in the plant root, and it is more abundant in the neighboring soil and root of *B. distachyon* Bd21 than in *A. thaliana* Col-0. b) Strain 363 is equally abundant between fractions, but it is more abundant than expected in the neighboring soil and root of Oy-0 than in Col-0. c) Strain 371 is less abundant in roots than in neighboring soil, but it is enriched in the roots of *C. rubella* and Oy-0. d) Strain 27 is less abundant in roots than in neighboring soil, but it is enriched in the roots of Oy-0.

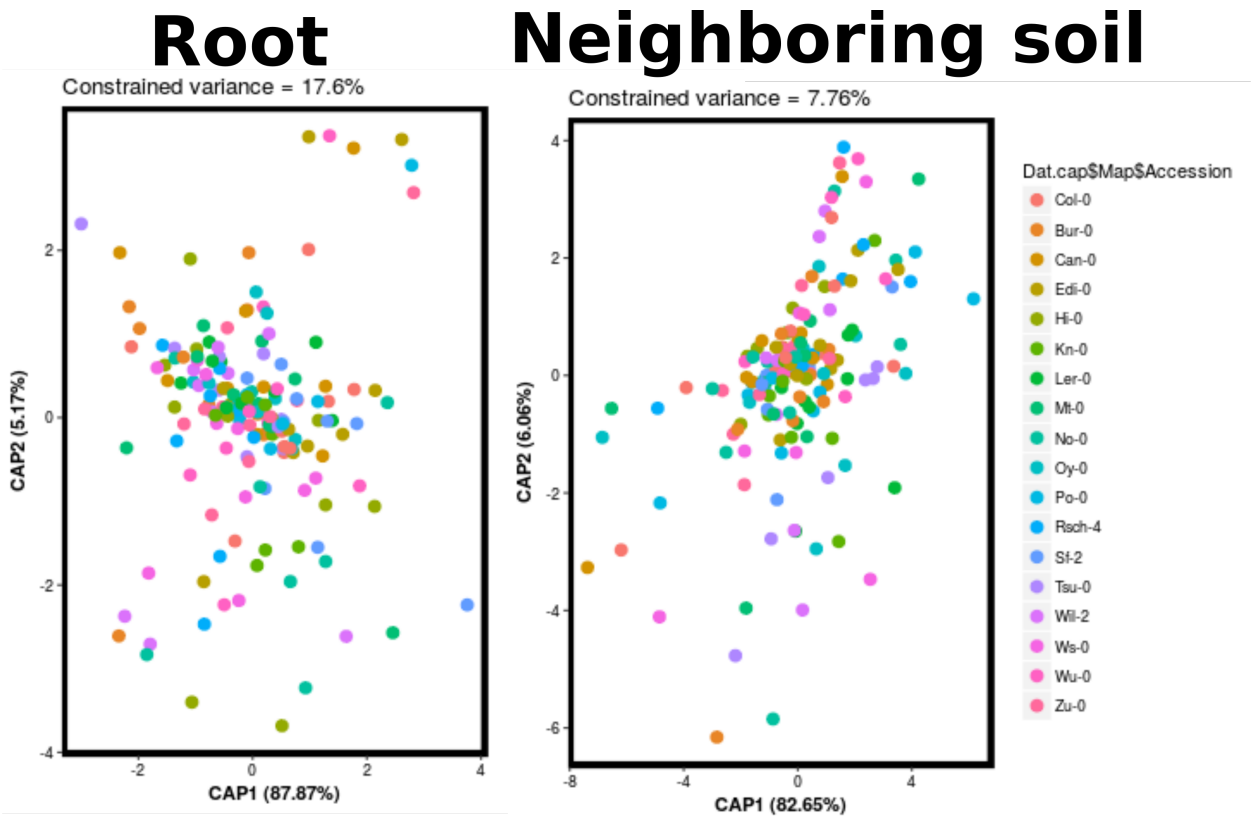


Figure 3.10: **CAP analysis of bacterial composition of root and neighboring soil samples from 18 Arabidopsis accessions**. Plant accession explainedv17.6% of the variance in community composition among root samples (p-value = 0.01499; permutation), while it explained only 7.76% of the variance among neighboring soil samples (p-value = 0.8541).

3.5 Estimating heritability of the root microbiome

We observed quantitative differences in presence/absence and relative abundance of specific strains across genetically diverse plant hosts. This observation suggests that the root microbiome can be viewed to some extent as an extended plant phenotype that could be genetically mapped. We decided to test a reduced complexity synthetic bacterial community in 18 diverse Arabidopsis accessions that are parents of the Arabidopsis MAGIC population (Kover et al., 2009) and parents of several recombinant inbred lines (RIL) population.

We used constrained ordination (section 3.7.6.1) independently on root and neighboring soils samples. The first ordination axis explained most of the constrained variance even after removing a few outliers (87% in roots, 82% in neighboring soils). Plant accession explained

17.6% of the variance in community composition among root samples (p-value = 0.01499; permutation, section 3.7.6.1), while it explained only 7.76% of the variance among neighboring soil samples (p-value = 0.8541; permutation, section 3.7.6.1). Thus, we observed measurable and heritable differences in the root microbiome composition between Arabidopsis accessions.

If root community composition represents heritable variation, then we should be able to measure a difference in the distribution of community dissimilarities of pairs of plants of the same accession in contrast with pairs of plants from different accessions, similar to what has been observed in twin studies of the human gut microbiome (Goodrich et al., 2014). We calculated those distributions for root and neighboring soil, and we found that dissimilarities are smaller between root samples than neighboring soil samples (F-value 443.774; $df = 1$; p-value $< 2x10^{-16}$; ANOVA), but no significant difference between intra- and inter-accession dissimilarities (F-value 0.446; $df = 1$; p-value = 0.504; ANOVA). This can be due to a number of factors like limited power and the fact that this approach does not control for several technical covariates like sequencing plate and depth.

We measured the largest ever reported heritability for plant microbiome, and thus our results provide evidence for the presence of heritable variation in the root microbiome, though the contribution of genetics remains small, and it is still unclear if it is feasible to dissect the root microbiome via genetic mapping. Twin studies of heritability in the human microbiome did not find significant differences until the sample size was increased from hundreds of twin pairs (for comparison this work uses 18 inbredd accessions) to thousands (Turnbaugh et al., 2009a; Goodrich et al., 2014). It is also possible that the 13 strains measured in this study are too few and not representative enough to capture the genetic variability associated with microbial composition.

3.6 Discusion

Previous work has shown that the root microbiome composition differs between soil types (Lundberg et al., 2012; Bulgarelli et al., 2012), and the microbiome of extremophile plants has been shown to have features that might explain their tolerance (Yuan et al., 2016; Dombrowski

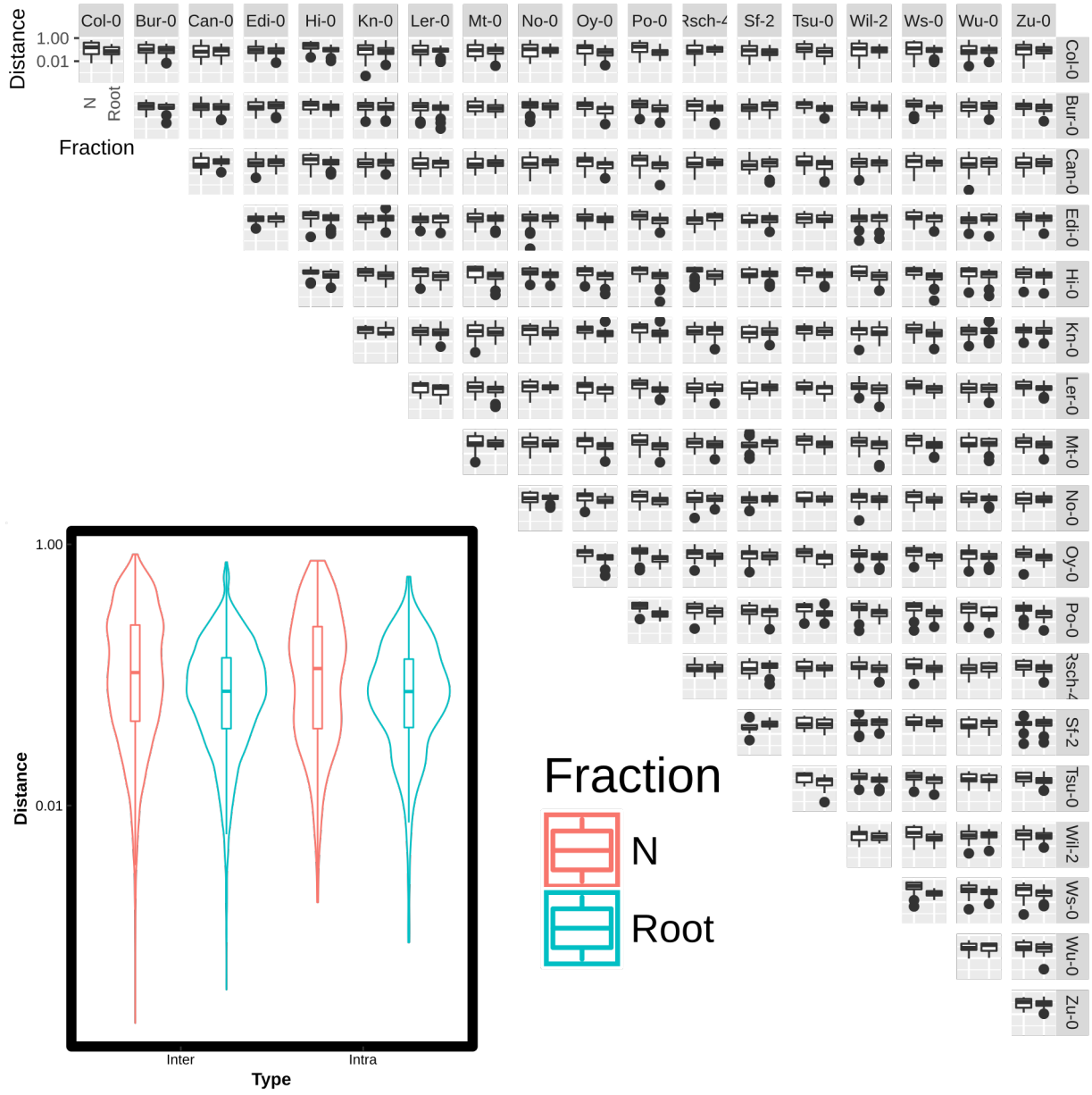


Figure 3.11: **Pairwise community dissimilarities**. Each panel in the grid (upper triangle) shows the distribution of pairwise Bray-Curtis dissimilarities (in log10 scale) between samples of the same fraction, same biological replicate and the genotypes of the corresponding column and row. Panels in the diagonal correspond to intra-accession differences while panels above the diagonal represent inter-accession differences. Violin plots in the lower triangle show the aggregated distributions of all panels in the upper triangle. Communities from roots are more similar between them than communities of neighboring soils F-value 443.774; $df = 1$; p-value $< 2x10^{-16}$; ANOVA), but there is no difference between intra- and inter-accession distributions (F-value 0.446; $df = 1$; p-value $= 0.504$; ANOVA).

et al., 2017). However, it is unclear whether those differences represent bacterial adaptations to the abiotic factors that influence plant-associated environment, or whether they represent plant adaptations that help them to cope with challenging environmental conditions.

By systematically varying both the abiotic environment and the host genotype, we were able to distinguish between changes that are due to a combination of adaptation to the abiotic environment and the plant response (when changing the abiotic composition), and changes that are exclusively due to genetically encoded differences between plant hosts, and thus are subject to selection on the plant. Overall, we observed more and stronger prevalence and relative abundance differences in response to changes in the nutrient composition than in response to different plant hosts, suggesting that the majority of the microbial abundance differences measured in different environments are the result of microbial competition and adaptation to their specific abiotic conditions. This is consistent with the different generational scales of plants and bacteria, and with the idea that plant colonization is a microevolutionary process for bacteria (Herrera Paredes and Lebeis, 2016). Despite this, we found a significant amount of variation in bacterial prevalence and relative abundance across plant hosts. Consistent with previous reports, we found that the variation in the *A. thaliana* species encompasses the variation of other Brassicaceae (Schlaeppi et al., 2014). Interestingly, we found that most of the variation attributable to genotype showed a clear pattern of bacterial exclusion in the majority of genotypes, while some genotypes (notably Oy-0) partially lost the ability to effect that exclusion. This suggests that a relatively simple host genetic architecture underlies these differences.

We profiled the root colonization patterns in 18 inbred *A. thaliana* accessions using a simplified synthetic community, and we detected the largest broad sense heritability in plant microbiome composition reported to date (17.6%). However, our study remains underpowered as we found no significant differences in intra- and inter-accession pairwise dissimilarity distributions. Exploiting natural variation to identify the genetic determinants that modulate root microbiome will probably require the optimization of the synthetic community and larger

plant diversity panels.

We have presented a reduced-complexity experimental and analytical platform that can be used to disentangle the microbial adaptation and the selection exerted by the host on specific microbes. Our approach is flexible and modular enough that can be easily adapted to test specific hypothesis or as a hypothesis generating tool in novel contexts.

3.7 Methods

3.7.1 Synthetic community experimental procedures

Synthetic community experiments were performed by filling 4 in^2 pots with calcined clay (Diamond Pro drying agent). Autoclaving the filled pots and then inoculating them with 40 mL a mix of media and a mix of 63 or 15 diverse bacterial strains (see section 3.7.4). Each bacterial strain was grown independently in liquid 2xYT media at 28°C. Equal volumes of bacteria cultures were mixed and the total mix OD was adjusted so that each pot receives 10^5 C.F.U/mL of media.

Seeds were surface sterilized following protocols described before (Lundberg et al., 2012), and stratified in the dark at 4°C for 3 days. Seeds were placed in a suspension of dH_2O, so that 100 μL of the suspension had an average of 6 seeds.

Pots inoculated with media bacteria were sowed with 100 μL of the seed suspension. The number of seeds per pot follows a Poisson distribution and given the sample numbers guarantees that all pots get at least a couple of seeds. We also kept a few pots without plants for the whole experiment.

Sowed pots were placed randomly in trays according to atmospheric-noise derived random numbers (`https://www.random.org/`), and trays were covered with transparent lids. Lids were removed 2 week after sowing, and pots were thinned to 1 plant per pot. Trays were periodically reshuffled in the growth chamber to minimize location effects. Plants were grown in a short-day photoperiod (8 h light) through the whole experiment and watered as needed with dH_2O from the top to simulate rainfall (Lundberg et al., 2012).

Plants were harvested at seven weeks post sowing with a protocol adapted from our

previous work in natural soils (Lundberg et al., 2012). Briefly, roots were cleaned with at least two rounds of washing in autoclaved dH_2O to remove all the calcined clay. The cleaned roots correspond to a combination of epi- and endo-phytic bacteria and were denominated the Root fraction, and were snap frozen with liquid nitrogen and freeze-dried for storage before DNA extraction. The soil from pots containing a plant was also collected, we refer to those as neighboring soils (N), and were processed by suspending 10 mL of the soil in 25mL of autoclaved dH_2O; after vortexing, these soils were filtered in with a 100 μm mesh and the filtrate was concentrated down to a 1.5 mL tube via centrifugation. The concentrate was snap frozen with liquid nitrogen and stored at -80°C before DNA extraction. Bulk soils (i.e. soils samples from pots with no plant) were processed in the same manner as the neighboring soils.

3.7.2 DNA extraction

DNA extraction was performed with the MoBio Power Soil htp kit. We followed the manufacturer's instructions except for the following two changes: 1) we pre-homogenized the freeze-dried root samples by bead beating them with three sterile 4 mm glass beads in sterile 2 mL tubes using the FastPrep tissue homogenizer, we re-suspended the homogenate in the bead solution from the Power Soil kit and proceeded with the protocol. 2) We eluted DNA in 100 μL of dH_2O instead of the 50 μL of solution C6 from the kit.

3.7.3 Library preparation and sequencing

Library preparation for 16S gene profiling was done following the method from Lundberg et al. (2013), with the adaptation described in Castrillo et al. (2017). Briefly we amplified the 16S gene with primers 338F and 806R, and we multiplexed all four 96-well plates into a single MiSeq library that was sequenced with a 600-cycle V3 kit, in a 300x2 run at UNC.

3.7.4 Synthetic community composition

For the main experiments, we decided to maximize diversity, in order to increase our chances of observing relative abundance changes in response to nutrient or plant genotype. Based on Sanger sequencing of the 16S gene from our culture collection, we selected 61 strains which was the maximum number of strains that we could chose while maintaining

the ability to differentiate by sequencing (every pair of strains had at least one mismatch in the V4 region of their 16S gene). Our strain collection is derived mainly from roots of Brassicaceae plants growing in one of two natural soils. We also included *E. coli* DH5α as a reference strains that is not adapted to the soil or root environment, but that grows well in all the media used. The full list of strains, their taxonomy, and their genome sequence (when available) are presented in table 3.3, and indicated in the column *Big*.

ID	Name	OID	Big	Small	Phylum	Class	Order	Family
3	*Pseudomonas sp.* BZ64	2513237142		x	P	γ-proteobacteria	Pseudomonadales	Pseudomonadaceae
8	*Chryseobacterium sp.* UNC8MFCol	2529292577	x		B	Flavobacteriia	Flavobacteriales	Flavobacteriaceae
10	*Agrobacterium sp.* 10MFCol1.1	2521172663		x	P	α-proteobacteria	Rhizobiales	Rhizobiaceae
27	*Bacillus flexus* 27Col1.1E	2522125133	x		F	Bacilli	Bacillales	Bacillaceae 1
29	*Rhodococcus sp.* 29MFTsu3.1	2519899643	x		A	Actinobacteria	Actinomycetales	Nocardiaceae
33	*Agrobacterium sp.* 33MFTa1.1	2561511224	x	x	P	α-proteobacteria	Rhizobiales	Rhizobiaceae
36	*Pseudomonas mandelii* 36MFCvi1.1	2521172653		x	P	γ-proteobacteria	Pseudomonadales	Pseudomonadaceae
40	*Flavobacterium sp.* 40S8	2563366720	x		B	Flavobacteriia	Flavobacteriales	Flavobacteriaceae
41	*Bacillus sp.* UNC41MFS5	2563366514	x		F	Bacilli	Bacillales	Bacillaceae 1
47	*Polaromonas sp.* JS666 UNC47MFTsu3.1	2636416056	x		P	β-proteobacteria	Burkholderiales	Comamonadaceae
50	*Pseudomonas sp.* KD5	2228664007	x		P	γ-proteobacteria	Pseudomonadales	Pseudomonadaceae
57	*Rhizobium sp.* 57MFTsu3.2	2228664006	x		P	α-proteobacteria	Rhizobiales	Rhizobiaceae
69			x		A	Actinobacteria	Actinomycetales	Microbacteriaceae
79	*Dyella japonica* UNC79MFTsu3.2	2556921674	x		P	γ-proteobacteria	Xanthomonadales	Xanthomonadaceae
80			x		F	Bacilli	Bacillales	Paenibacillaceae 1
105	Bacillus sp. 105MF	2517572206	x		F	Bacilli	Bacillales	Bacillaceae 1
109	Leifsonia sp. 109	2522572063	x		A	Actinobacteria	Actinomycetales	Microbacteriaceae
125	Bacillus sp. UNC125MFCrub1.1	2561511073	x		F	Bacilli	Bacillales	Bacillaceae 1
135	Arthrobacter sp. 135MFCol5.1	2517572123		x	A	Actinobacteria	Actinomycetales	Micrococcaceae
136	Streptomyces sp. 136MFCol5.1	2636416059	x		A	Actinobacteria	Actinomycetales	Streptomycetaceae
138	Luteibacter sp. UNC138MFCol5.1	2593339266	x		P	γ-proteobacteria	Xanthomonadales	Xanthomonadaceae
140	Streptomyces sp. 140Col2.1E	2563366508		x	A	Actinobacteria	Actinomycetales	Streptomycetaceae

161	Arthrobacter sp. 161MFSha2.1	2517572124	x		A	Actinobacteria	Actinomycetales	Micrococcaceae
174	Methylobacterium sp. 174MFSha1.1	2590828856	x		P	α-proteobacteria	Rhizobiales	Methylobacteriaceae
181	Paenibacillus sp. 181MFCol5.1	2639762524	x		F	Bacilli	Bacillales	Paenibacillaceae 1
199			x		B	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
217	Paenibacillus sp. UNC217MF	2563366516	x		F	Bacilli	Bacillales	Paenibacillaceae 1
267	Mycobacterium sp. UNC267MFSha1.1M11	2593339259	x		A	Actinobacteria	Actinomycetales	Mycobacteriaceae
273	Terracoccus sp. 273MFTsu3.1	2522125155	x		A	Actinobacteria	Actinomycetales	Intrasporangiaceae
279	Caulobacter sp. UNC279MFTsu5.1	2590828858	x		P	α-proteobacteria	Caulobacterales	Caulobacteraceae
299	Streptomyces canus 299MFChir4.1	2521172643	x	x	A	Actinobacteria	Actinomycetales	Streptomycetaceae
302	Phyllobacterium sp. UNC302MFCol5.2	2563366739	x		P	α-proteobacteria	Rhizobiales	Phyllobacteriaceae
303	Streptomyces sp. 303MFCol5.2	2521172626	x		A	Actinobacteria	Actinomycetales	Streptomycetaceae
313			x		P	β-proteobacteria	Burkholderiales	Oxalobacteraceae
314	Curtobacterium sp. 314Chir4.1	2521172612	x		A	Actinobacteria	Actinomycetales	Microbacteriaceae
322	Bacillus sp. UNC322MFChir4.1	2574179748	x		F	Bacilli	Bacillales	Bacillaceae 1
327	Promicromonospora sukumoe 327MF-Sha3.1	2522572130	x		A	Actinobacteria	Actinomycetales	Promicromonosporaceae
339	Rhodococcus erythropolis 339MF-Sha3.1	2643221496	x	x	A	Actinobacteria	Actinomycetales	Nocardiaceae
348	Nocardia sp. 348MFTsu5.1	2521172629	x		A	Actinobacteria	Actinomycetales	Nocardiaceae
358	Caulobacter sp. UNC358MFTsu5.1	2565956508	x		P	α-proteobacteria	Caulobacterales	Caulobacteraceae
360	Mycobacterium sp. 360MFTsu5.1	2521172630	x		A	Actinobacteria	Actinomycetales	Mycobacteriaceae
362	Arthrobacter sp. UNC362MFTsu5.1	2563366511	x		A	Actinobacteria	Actinomycetales	Micrococcaceae
363	Rhodococcus sp. UNC363MFTsu5.1	2563366512	x		A	Actinobacteria	Actinomycetales	Nocardiaceae
370	Ochrobactrum sp. 370MFChir3.1	2643221500	x		P	α-proteobacteria	Rhizobiales	Brucellaceae
371			x	x	B	Sphingobacteriia	Sphingobacteriales	Sphingobacteriaceae
374			x		A	Actinobacteria	Actinomycetales	Nocardiaceae
376	Burkholderia bryophila 376MF-Sha3.1	2521172625	x	x	P	β-proteobacteria	Burkholderiales	Burkholderiaceae
468			x		F	Bacilli	Bacillales	Planococcaceae
496	Paenibacillus sp. UNC496MF	2593339199	x		F	Bacilli	Bacillales	Paenibacillaceae 1
499	Paenibacillus sp. UNC499MF	2593339197	x		F	Bacilli	Bacillales	Paenibacillaceae 1

501			x		P	γ-proteobacteria	Xanthomonadales	Xanthomonadaceae
220	Sphingomonas sp. 220AMFTsu3.1	2643221532	x		P	α-proteobacteria	Sphingomonadales	Sphingomonadaceae
305	Sphingomonas sp. UNC305MFCol5.2	2565956511	x		P	α-proteobacteria	Sphingomonadales	Sphingomonadaceae
345A	Nocardioides sp. UNC345MFTsu5.1	2582580751	x	x	A	Actinobacteria	Actinomycetales	Nocardioidaceae
CL13	Bacillus sp. UNCCL13	2639762621	x		F	Bacilli	Bacillales	Bacillaceae 1
CL14	Variovorax paradoxus CL14	2643221508	x	x	P	β-proteobacteria	Burkholderiales	Comamonadaceae
CL18	Streptomyces sp. UNC401CLCol	2563366515	x		A	Actinobacteria	Actinomycetales	Streptomycetaceae
CL19	Bosea sp. UNC402CLCol	2579779168	x		P	α-proteobacteria	Rhizobiales	Bradyrhizobiaceae
CL20	Curtobacterium sp. UNCCL20	2595698215	x		A	Actinobacteria	Actinomycetales	Microbacteriaceae
CL21	Ralstonia sp. UNC404CL21Col	2558309150	x	x	P	β-proteobacteria	Burkholderiales	Burkholderiaceae
CL41	Agrobacterium sp. UNC420CL41Cvi	2529292583	x		A	Actinobacteria	Actinomycetales	Micrococcaceae
CL52	Paenibacillus sp. UNCCL52	2563366513	x	x	F	Bacilli	Bacillales	Paenibacillaceae 1
CL69	Acinetobacter sp. UNC434CL69Tsu2S25	2593339129	x		P	γ-proteobacteria	Pseudomonadales	Moraxellaceae
CL81	Bacillus sp. UNCCL81	2593339131	x	x	F	Bacilli	Bacillales	Bacillaceae 1
CL82			x		P	γ-proteobacteria	Xanthomonadales	Xanthomonadaceae
CL9	Mycobacterium sp. UNCCL9	2576861824	x		A	Actinobacteria	Actinomycetales	Mycobacteriaceae
Ecoli	Escherichia coli DH5α		x		P	γ-proteobacteria	Enterobacteriales	Enterobacteriaceae

Table 3.3: **Isolates used in synthetic community experiments**. Phylum abbreviations: P = Proteobacteria, F = Firmicutes, B = Bacteroidetes, A = Actinobacteria.

For the heritability experiment we decided to simplify the community. We chose 15 strains that had differences in relative abundance between genotypes (371 & CL52) and between media types (345A, CL14 & CL21). We also included strains (including some that were not in the original community) so that there were two representatives of the genera Agrobacterium (10 & 33), Pseudomonas (3 & 36) and Streptomyces (135 & 299), which are commonly found in the root and rhizosphere. A further four strains were included to ensure phylogenetic diversity (135, 339, 376 & CL81). The full list of strains used, their taxonomy, and genome

(when available) are presented in table 3.3, and are indicated in column *Small*.

3.7.5 Sequence processing

Sequences were processed as defined before (Lebeis et al., 2015; Castrillo et al., 2017). Briefly, sequences were processed via MT-Toolbox (Yourstone et al., 2014) which merged paired reads and identified the outer index. Then sequences where demultiplexed based on frameshift lenghts and inner barcode sequences (Lundberg et al., 2013; Castrillo et al., 2017). Reads were trimmed to from the 5' end up to a length of 330bp. Reads were then mapped with USEARCH6 (Edgar, 2010), at 98.5% identity to a reference database containing Sanger sequences of the 16S gene from all isolates, as well as plant nuclear and organellar rRNA genes. A count table was built from the mapping results using a QIIME script (Caporaso et al., 2010), and plant-derived counts were discarded. The resulting count tables are the basis for analysis.

For the heritability experiment, the same pipeline approach was used, but reads were trimmed from the 3' end to a final length of 335 bp due to low quality in the second read for most pairs, and the trimmed reads were quality-filtered with sickle (Joshi and Fass, 2011) using default parameters.

3.7.6 Ordination

For Principal Component Analysis (PCA) the count tables were converted to percent abundant tables by dividing the counts of each bacteria in each sample by the total number of reads in that sample, using the `normalize` function from the AMOR package (Herrera Paredes, 2016), and PCA was performed directly on the normalized table.

For Canonical Analysis of Principal Coordinates (CAP) distances between samples were calculated directly on the unnormalized count table with the Bray-Curtis dissimilarity , and the `capscale` function from the vegan package was used Oksanen et al. (2014). CAP was performed separately on the Neighboring soil and root samples, constraining by media or genotype and conditioning on experiment and sequencing depth with one of the following

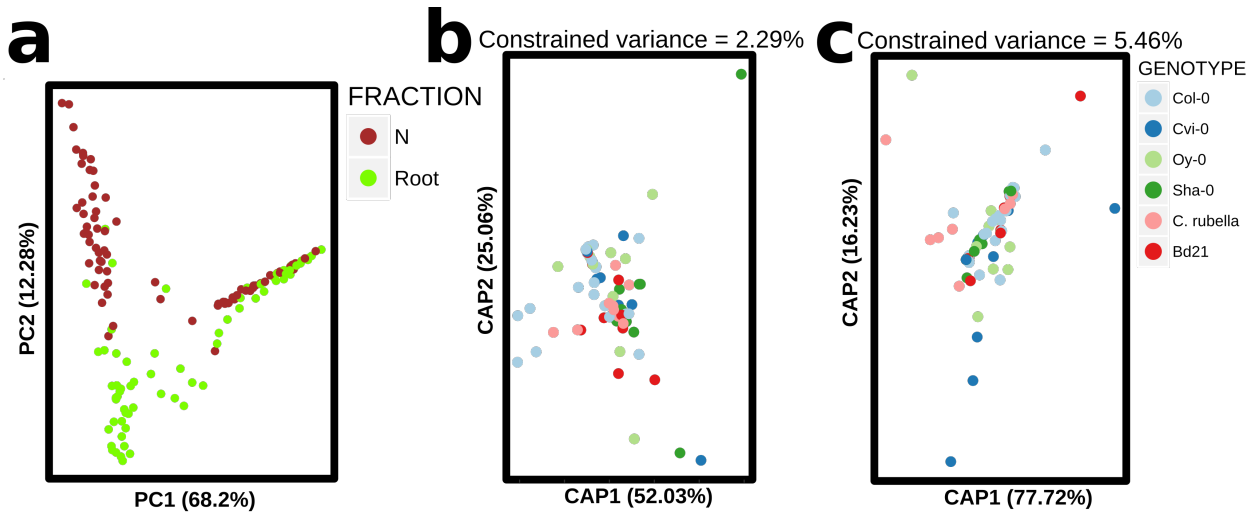


Figure 3.12: Ordination results from samples form different plant genotypes.

formula:

$$Counts = MEDIA + Condition(EXPERIMENT) + Condition(DEPTH) \quad (3.1)$$

$$Counts = GENOTYPE + Condition(EXPERIMENT) + Condition(DEPTH) \quad (3.2)$$

PCA showed a clear separation between sample fractions (Figs. 3.2a and 3.12a). The effect of media and genotype was small but significant, and it was stronger in the root samples (Figs. 3.2c and 3.12c) than in the neighboring soils (Figs. 3.2b and 3.12b)

3.7.6.1 Heritability experiment For the heritability experiment, we performed canonical analysis of principal coordinates (CAP) on the relative abundance table with Bray-Curtis dissimilarity. We performed CAP separately on neighboring soils (N) and root samples, conditioning by experiment, sequencing plate and depth using the following formula:

$$RA = Accession + Condition(Exp) + Condition(Plate) + Condition(Depth) \quad (3.3)$$

To test the significance of the Accession term, we used the `anova.cca` function from the vegan package (Oksanen et al., 2014), with 1000 permutations.

3.7.7 Identifying robust colonizers

Presence of a strain in a sample was defined as at least 5 reads in a given sample, though similar results were obtained by lowering this threshold to 1 read, or raising it to 10 (not shown). Then we defined *robust colonizers* as those strains that had a probability of being present in Col-0 roots that was statistically significantly higher than 50% (q-value < 0.05, one-sided binomial test) (Lebeis et al., 2015; Castrillo et al., 2017). For this test, we considered only roots of plants grown either on 1/4 MS media or 1/25 MS media. Isolates that didn't pass our stringent statistical threshold are labelled *sporadic colonizers*.

3.7.8 Testing from presence/absence differences

To identify quantitative differences in bacterial prevalence (presence/absence) across different fractions, media and host genotypes. We utilized a regularized logistic regression with the ridge penalty. We used the implementation in the glmnet R package (Friedman et al., 2010). The objective function that must be minimized to fit the regularized logistic regression is defined by the following equation:

$$-\ell/n + \lambda * \frac{1}{2}||\beta||_2^2 \tag{3.4}$$

where ℓ is the log-likelihood from the logistic regression, n is the number of samples, $||\beta||_2^2$ is the ridge penalty and λ is the weight we give the the regularization penalty (Friedman et al., 2010).

To determine the value of λ, we use k-fold cross-validation, with $k = 10$, on each isolate and select the value of λ that minimizes the model's deviance. Once the value of λ has been determined, we use permutation ($N = 1000$) to estimate a p-value for each model coefficient, using the same λ value in all permutations. This p-values were corrected for multiple testing using the method by Storey and Tibshirani (2003), and a q-value < 0.05 was used as a threshold for significance.

We used an overparameterized design matrix to fit the model, where each level of each

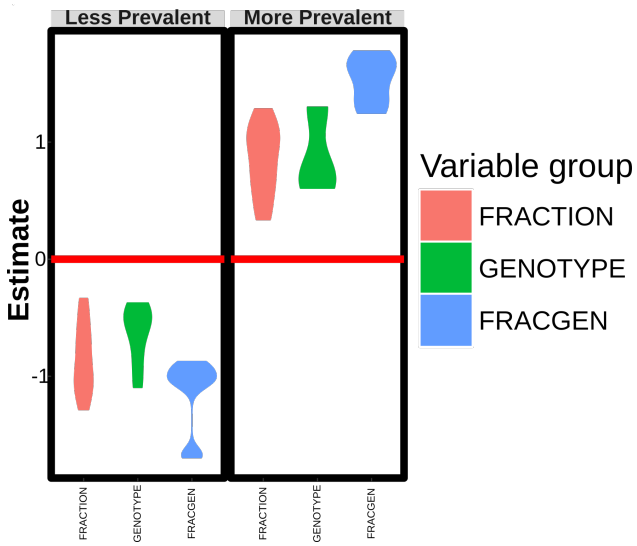


Figure 3.13: Genotype has a larger effect on presence/absence inside the root. Violin plots showing the effect sizes of different variables on the presence/absence of individual strains. Y-axis indicates the coefficient values from the ZINB model; it can be interpreted as the log fold-change in relative abundance between samples of different fraction, genotype or combination of fraction and genotype (FRACGEN). Only significant (q-value < 0.05) effects are plotted.

variable is represented. This design matrix facilitates interpretation and is able to handle linearly dependent variables thanks to the parameter regularization. The coefficients from this model represent the change in the log odds of a given isolate being found in samples of a particular type, with positive values meaning that samples of that group are more likely to contain that isolate. The regularization term shrinks the coefficients towards zero so that only isolates that have strong evidence for an effect end up being significant, and the values of the significant coefficients for each model are plotted in Figs. 3.4 and 3.13

The observed counts are binarized using 5 reads as a threshold for presence. Once the count matrix is converted into a prevalence matrix, it is fit to a model defined by one of the following formulas:

$$\begin{aligned} Prevalence = FRACTION + MEDIA + FRACMEDIA + \\ EXPERIMENT + LOGDEPTHK \end{aligned} \tag{3.5}$$

$$Prevalence = FRACTION + GENOTYPE + FRACGEN + EXPERIMENT + LOGDEPTHK \quad (3.6)$$

3.7.9 Identifying relative abundance differences

We used a linear modelling approach to identify changes in relative abundances due to multiple variables. We tested both a Zero-Inflated Negative Binomial (ZINB) model that we have previously described in Lebeis et al. (2015) (section 4.9.7.2), and the empirical Bayes moderated dispersion parameter estimation in a negative binomial GLM as implemented in the R package edgeR (Robinson et al., 2010; McCarthy et al., 2012). Overall, we found high consistency between both methods (above 75% in all variables; not shown), but edgeR tended to provide fewer significant results.

We chose the ZINB approach, which takes into account the sparsity of the data. And we used similar formulas as in section 3.7.8, but we noticed that sequencing plate was significant when considering relative abundances (while it wasn't when considering presence absence) so we included it in the formula for media and plant genotype according to the following formulas:

$$Abundance = FRACTION + MEDIA + FRACMEDIA + EXPERIMENT + PLATE + LOGDEPTHK \quad (3.7)$$

$$Abundance = FRACTION + GENOTYPE + FRACGEN + EXPERIMENT + PLATE + LOGDEPTHK \quad (3.8)$$

CHAPTER 4

Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa[1]

Immune systems distinguish 'self' from 'nonself' to maintain homeostasis and must differentially gate access to allow colonization by potentially beneficial, nonpathogenic microbes. Plant roots grow within extremely diverse soil microbial communities but assemble a taxonomically limited root-associated microbiome. We grew isogenic *Arabidopsis thaliana* mutants with altered immune systems in a wild soil and also in recolonization experiments with a synthetic bacterial community. We established that biosynthesis of, and signaling dependent on, the foliar defense phytohormone salicylic acid is required to assemble a normal root microbiome. Salicylic acid modulates colonization of the root by specific bacterial families. Thus, plant immune signaling drives selection from the available microbial communities to sculpt the root microbiome.

Recognition of plant pathogens in leaves leads to dramatic changes in transcription, synthesis of defense phytohormones and antimicrobial compounds, and elaboration of physical barriers (Dodds and Rathjen, 2010; Jones and Dangl, 2006). Defense phytohormones are structurally diverse plant secondary metabolites that integrate plant immune system output responses while repressing cell growth and proliferation. Salicylic acid (SA), jasmonic acid

[1]Most of the content of this chaper has been published before as a peer-reviewed article (Lebeis et al., 2015). The text has been been ligthly edited and re-arranged to facilitate reading. The figure order has been changed to match the updated text order. Section and subsection headers have been added for easier navigation. Several minor mistakes have been amended. Numerous supplementary files were made available online at the time of publication, and are not included here; they will be referred to as Supplementary Table or Supplementary Dataset and can be obtained at `http://science.sciencemag.org/content/early/2015/07/15/science.aaa8764.figures-only`.

(JA), and gaseous ethylene mediate localized and systemic plant immune responses (Belkhadir et al., 2014; Pieterse et al., 2012). Nonspecific systemic acquired resistance is mediated by SA in leaves (Fu and Dong, 2013). In contrast, induced systemic resistance in leaves can be triggered by specific rhizobacteria colonizing roots and is mediated by JA and ethylene (Pieterse et al., 2012). SA and JA act antagonistically in responses to infection by biotrophs, at least in leaves (Huot et al., 2014). The defense phytohormones control a set of overlapping signaling sectors, each contributing to the regulation of plant defense via transcriptional and biosynthetic output in leaves (Kim et al., 2014).

Accessions of *Arabidopsis thaliana* show variation in defense phytohormone profiles after infection, even though they share similar root-associated bacterial microbiota (Bulgarelli et al., 2012; Kliebenstein et al., 2002; Lundberg et al., 2012). Previous studies examined the roles of defense phytohormones in shaping the wild-type root microbiome by using single mutant lines defective in their biosynthesis or perception, or exogenous defense hormone application in combination with bacterial culturing and/or lower-resolution profiling methods. No generalizable clarity has emerged to date (Bakker et al., 2013; Mendes et al., 2013). We therefore compared the bacterial root microbiome of wild-type *A. thaliana* accession Col-0 with a set of isogenic mutants lacking biosynthesis of, and/or signaling dependent on, at least one of the following: SA, JA, and ethylene. We focused on those with multiple mutations that eliminated overlapping defense-signaling sectors Fig. 4.1A and table S1) (Katagiri and Tsuda, 2010). We anticipated that this experimental design would reveal the contributions of plant defense phytohormones to wild-type root microbiome composition.

4.1 Defense phytohormone mutant genotypes

Plant-associated microbial communities promote plant productivity by improving accessibility to nutrients, producing plant growth stimulating factors, and inducing protection against pathogen infection and various abiotic stresses (Bulgarelli et al., 2013; Vorholt, 2012). The plant immune system detects microbes using highly polymorphic external and internal receptors, which recognize both general microbe-associated molecular patterns and specific

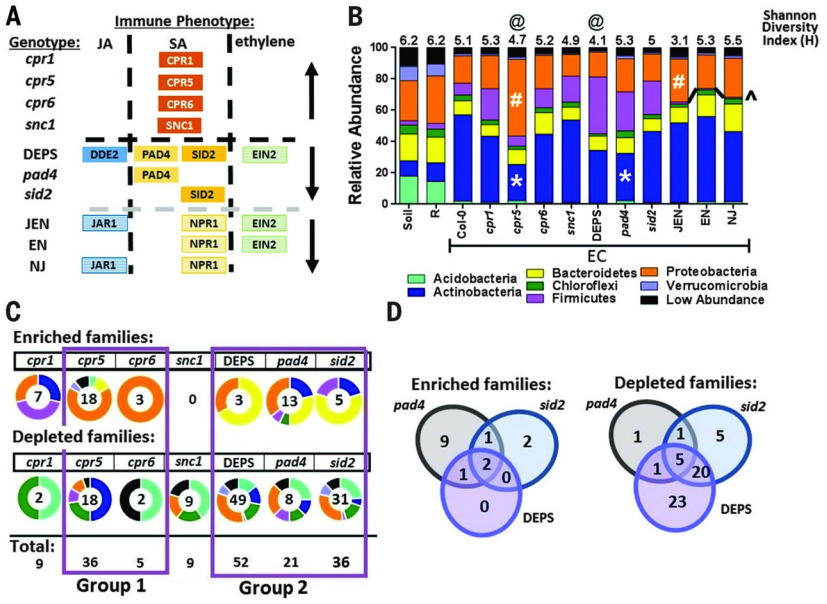


Figure 4.1: **Defense phytohormone mutants have altered root bacterial communities compared with those of wild-type plants**. **A** JA, SA, and ethylene mutants (names at left) derived from wild-type Col-0. Upward and downward black arrows at right indicate hyper- and hypo-immune mutants, respectively. **B** Phyla distributions were separated into sample fractions [soil, Col-0 rhizosphere (R), or EC] and plant genotypes. Shannon diversity indices are listed above each bar. Asterisk indicates a phylum significantly lower than Col-0 EC at $p - value < 0.001$; pound sign indicates a phylum significantly higher than Col-0 EC at $p - value < 0.05$; and caret indicates that JEN, EN, and NJ Firmicutes relative abundances were significantly lower than Col-0 EC at $p - value < 0.04$; @ indicates that the Shannon diversity index is significantly lower than Col-0 EC at $p - value < 0.001$ (all ANOVA with post hoc Tukey test). **C** The phyla distribution [circles color-coded as in (B)] of bacterial families identified as either enriched or depleted in ECs of each mutant compared with Col-0. The number of families in each category is noted inside each donut. Groups defined by means of Monte Carlo testing of Manhattan distances. **D** Venn diagrams showing the overlap of (left) enriched or (right) depleted group 2 families from (B).

pathogen virulence molecules. Salicylic acid (SA) biosynthesis is induced by immune receptor-mediated recognition of microbial pathogens that require living host tissue (biotrophs) (Spoel and Dong, 2008). By contrast, ethylene and jasmonic acid (JA) biosynthesis are induced by pathogens that cause and exploit host cell death (necrotrophs); the consequence of their action contributes to limiting necrotrophic infections (Spoel and Dong, 2008). In order to determine the role for the phytohormones salicylic acid, jasmonic acid, and ethylene production and signaling in controlling microbiome community composition, we examined the microbial

communities of roots in a variety of *Arabidopsis thaliana* defense phytohormone mutants (Fig. 4.1A and table S1). We used four hyper-immune mutants (*cpr1*, *cpr5*, *cpr6*, and *snc1*) previously characterized to constitutively produce enhanced levels of salicylic acid and constitutive defense signaling through salicylic acid in leaves (Bowling, 1994; Clarke et al., 1998; Kirik et al., 2001; Zhang et al., 2003). We investigated two classes of immunocompromised mutants, which either lack pathogen-induced biosynthesis of salicylic acid (*sid2*; (Dewdney et al., 2000)), or produce salicylic acid, but lack sensitivity to it (*pad4*; (Glazebrook et al., 1996)). We examined the role of salicylic acid biosynthesis or signaling in combination with a loss in jasmonic acid (JA) biosynthesis and ethylene sensitivity (with *dde2 ein2 pad4 sid2* (DEPS); (Tsuda et al., 2009)). In the second class of immunocompromised mutants, we examined the role of salicylic acid sensitivity in combination with jasmonic acid sensitivity (with *npr1 jar1* (NJ); (Clarke, 2000)), ethylene sensitivity (with *ein2 npr1* (EN); (Clarke, 2000)), or sensitivity to salicylic acid, jasmonic acid, and ethylene (with *jar1 ein2 npr1* (JEN); (Clarke, 2000)).

Root expression of each of these genes in wild type Col-0 roots was confirmed via Genevestigator's plant biology database (`https://genevestigator.com/gv/doc/intro_plant.jsp`) with the exception of CPR6, which was not in the database. We confirmed the salicylic acid hyper-accumulation phenotypes in leaves of *cpr1*, *cpr5*, *cpr6*, and *snc1*, and the absence of salicylic acid in biosynthetic *sid2* mutant leaves. However, we noted low salicylic acid levels in roots of all genotypes grown in wild Mason Farm soil (Fig. 4.14A). Further, *cpr1*, *cpr5*, *cpr6*, *snc1*, *pad4*, and *sid2* seedlings were grown axenically for 18 days in vertical plates as described in section 4.9.1.2 for tissue to measure salicylic acid accumulation in axenic conditions (Fig. 4.14B). Finally, root morphological differences between genotypes grown on agar could not explain the observed overlap in microbiome differences from wildtype (Fig. 4.14C-D; section 4.3).

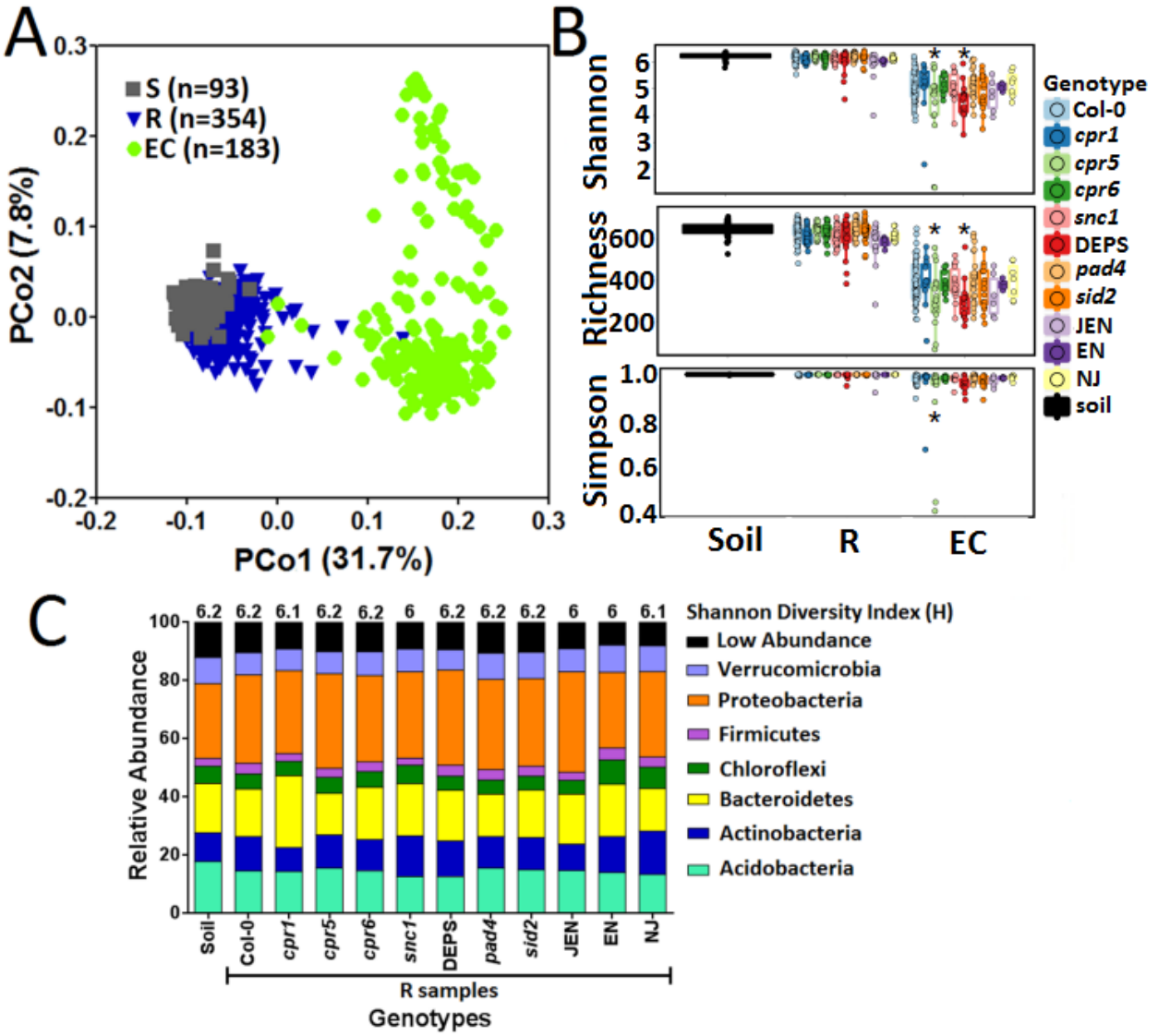


Figure 4.2: **Sample fraction drives differences in alpha and beta diversity of root microbiome communities**. **A** Principal Coordinate Analysis (PCoA) of pairwise normalized, weighted UniFrac distances between the samples considering rarified to 1,000 abundance of all OTUs. **B** Shannon diversity index, richness, and Simpson index for bulk soil, rhizosphere (R), and endophytic compartment (EC) samples for each genotype with the median represented by the bar and the 25^{th} and 75^{th} percentiles represented by the box. Asterisk indicates significantly lower than Col-0 EC samples at $p - value < 0.001$ by ANOVA test with post hoc Tukey test. **C** Phyla distributions were separated into sample fractions (Soil or Rhizosphere) and plant genotypes. Shannon Diversity indices are listed above each bar. There were no significant differences in the Shannon Diversity or phyla abundances.

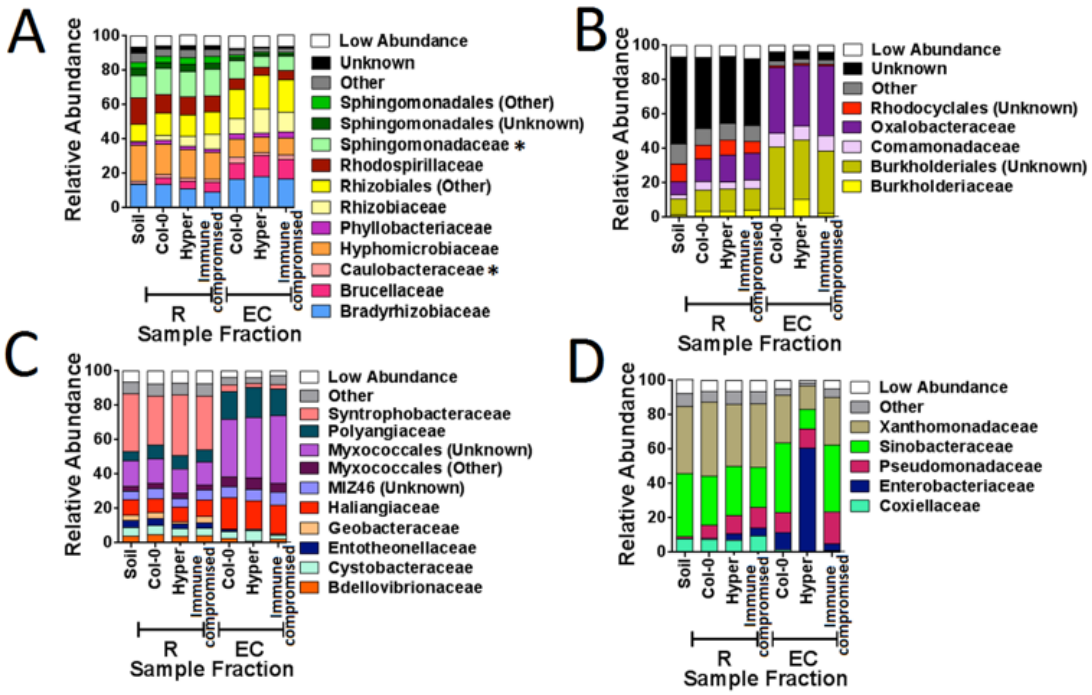


Figure 4.3: **Differential abundance of Proteobacteria families in different sample fractions**. Relative abundance of Proteobacteria families in the α (A), β (B), δ (C) and γ (D) orders in bulk soil, rhizosphere (R), and endophytic compartment (EC) sample fractions. Asterisk in (A) indicates that these families are significantly less abundant in EC-hyper samples compared to EC-Col-0 samples by ANOVA and post hoc Tukeys test, $p-value < 0.05$.

4.2 Overall diversity patterns

Through sequencing the 16S rRNA gene, we profiled bacterial communities of rhizosphere (soil directly adjacent to the root) and endophytic compartment (EC) from roots grown in a previously characterized wild soil from the University of North Carolina Mason Farm biological preserve, as well as unplanted bulk soil (Figs. 4.17, 4.4, 4.15 and 4.2, and tables S2 to S4, and methods 4.9.2) (Lundberg et al., 2012). Sample fraction (soil, rhizosphere, or endophytic compartment) and the differentiation of endophytic samples from bulk soil and rhizosphere explained the largest proportions of variance across the bacterial communities examined (table S5) (Bulgarelli et al., 2012; Lundberg et al., 2012). Endophytic bacterial communities were less diverse than bulk soil and rhizosphere communities (Figs. 4.1B and 4.2), with reduced representation of Acidobacteria, Bacteroidetes, and Verrucomicrobia and enrichment of Actinobacteria and Firmicutes [analysis of variance (ANOVA), $q-value < 0.05$]. Individual Proteobacteria families were either enriched or depleted in endophytic communities as compared with those of bulk soil and rhizosphere samples (Fig. 4.3, and methods 4.9.7.2). These results are consistent with distributions of bacterial phyla from *A. thaliana* roots grown in four wild soils (Bulgarelli et al., 2012; Lundberg et al., 2012). Interestingly, and in contrast with bacterial patther, we found α- and β-diversity differences between bulk soil, rhizosphere and endophytic compartment fungal communities as measured by sequencing the ITS intergenic region, and thus we focus on bacterial communities ahead (Fig. 4.17 and section 4.9.3.2).

Plant genotype affected phylum-level bacterial root endophytic community composition [4.3 to 5.0%, canonical analysis of principal coordinates (CAP)] (Figs. 4.1B and 4.16; and methods 4.9.7.1, 4.9.5.1 and 4.9.5.2) (Anderson and Willis, 2003), with both hyperimmune *cpr5* and immunocompromised quadruple *dde1 ein2 pad4 sid2* mutant communities displaying lower α-diversity indices than that of the wild type (Figs. 4.1B and 4.2B, and section 4.1). The relative abundance of Firmicutes was lower in immunocompromised *jar1 ein2 npr1*, *ein2 npr1*, and *npr1 jar1* mutants, which all lack response to SA (Fig. 4.1A-B, and table

S1). Actinobacteria were less abundant in *cpr5* and *pad4* endophytic samples, whereas Proteobacteria were more abundant in *cpr5* and *jar1 ein2 npr1* (Figs. 4.1A-B and 4.16, and section 4.9.5.1). Only mutants that lacked all three defense hormone signaling systems exhibited diminished survival that correlated with the presence of unidentified oomycete sequences in the root microbiota of survivors (Fig. 4.4 and section 4.9.4.7). The vast majority of these sequences corresponded to two Operational Taxonomic Units (OTUs) that were ∼20bp shorter than bacterial amplicons, and matched mitochondrial sequences from the oomycete genera *Phythophthora* and *Pythium* (Fig. 4.4). Our identification of oomycete sequences closely related to known plant pathogens is consistent with increased susceptibility of these mutant lines to infection (Tsuda et al., 2009; Clarke, 2000). Presumably as a consequence, both the JEN and DEPS mutants survived poorly on wild soil over the experimental time course, resulting in a lower number of replicates (table S3).

4.3 Salicylic acid genetic status explains differential abundances of specific taxa

To determine which taxonomic groups associate differentially with each variable of interest, we took a linear modeling approach. We first collapsed OTUs assigned to the same bacterial family (section 4.9.4.7), by aggregating their counts into a family-level count table. We decided to focus mainly on family-level abundances, because most of the data (i.e. the Roche 454 census experiments) is based on fragments of only 220bp, and it has been previously shown that only a small portion of sequences can be accurately given a genus level assignment (Guo et al., 2013), and it has been suggested that genus level assignments should only be performed with at least 250bp sequences (Liu et al., 2008). We also prefer family-level over OTU based analysis, because taxonomic families likely represent monophyletic groups while OTUs can be (and many are) paraphyletic (Koeppel and Wu, 2013). Despite all of these drawbacks, we analyzed the OTU-level count table as well using exactly the same model specification that we used in the family-level analysis, and we observed similar trends (Fig. 4.6).

We previously found that only OTUs with at least 25 reads in at least each of 5 different

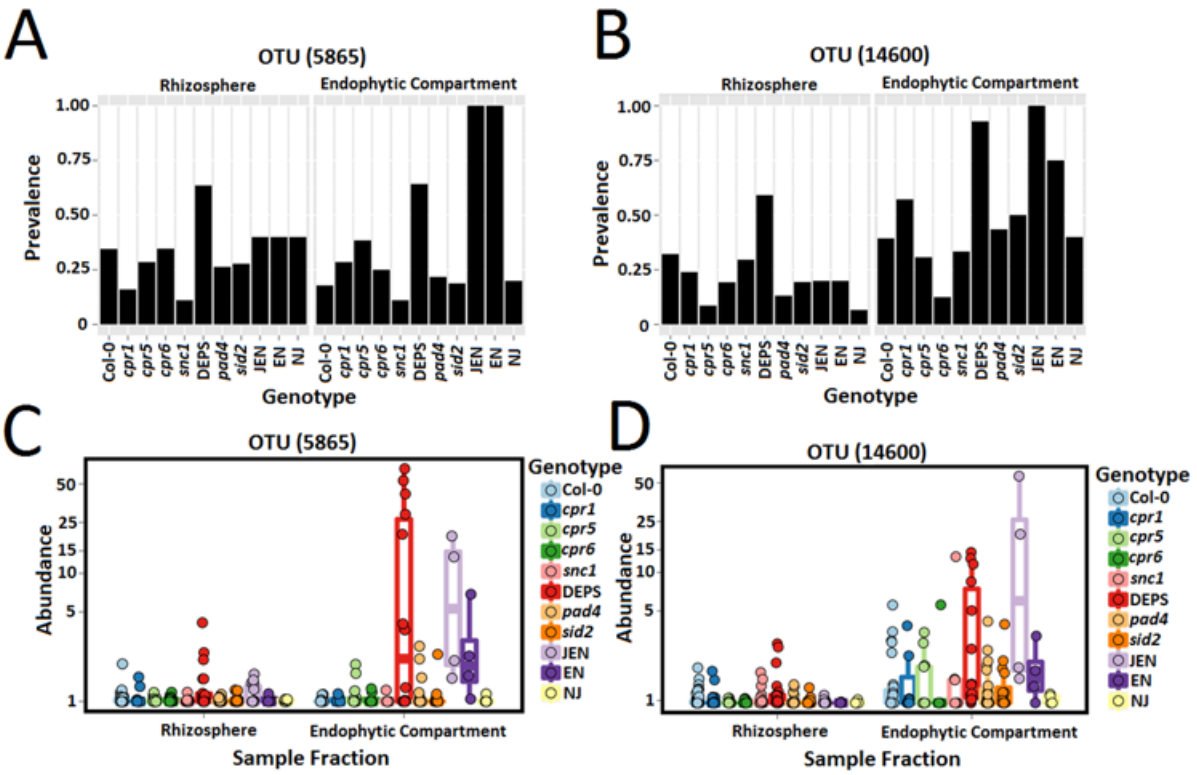


Figure 4.4: **DEPS and JEN root microbiome communities contain a disproportionate number of oomycete mitochondria reads**. The prevalence of the top two OTUs matching oomycete mitochondria, OTU 5865 **A** and OTU 14600 **B** in each genotype. The percent abundance (over total non-plant reads) of OTU 5865 **C** and OTU 14600 **D** in Rhizosphere or Endophytic Compartment samples of each genotype is shown.

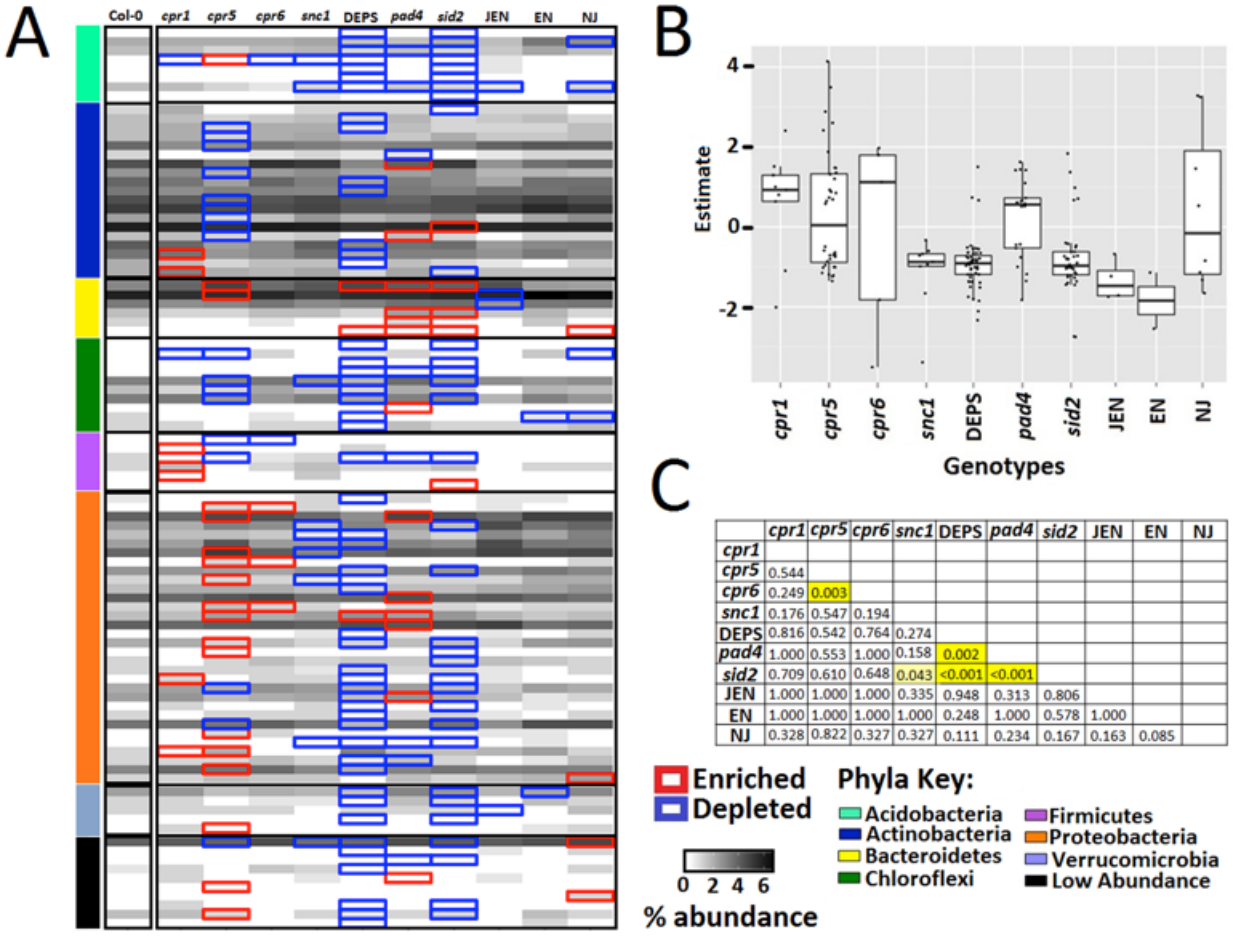
samples, produce reproducible abundances, and we defined these as *measurable* taxa (Lundberg et al., 2012). We restricted our analysis to these measurable taxa, and applied a Zero-Inflated Negative Binomial (ZINB) model (Fig. 4.18 and section 4.9.7.2).

Using the ZINB approach, we identified bacterial families and operational taxonomic units (OTUs) in the root endophyte microbiome of each mutant plant line that were differentially abundant as compared with wild-type plants (Figs. 4.1C, 4.5 and 4.6; tables S6 and S7). Both the number of differentially abundant bacterial taxa and their identity differed in endophytic samples from mutants. Among 52 differentially abundant families in surviving *dde1 ein2 pad4 sid2* mutant endophytic samples, nearly all were depletions (Figs. 4.1C and 4.5), which is consistent with this mutant's decreased α-diversity (Fig. 4.1B). Differentially

abundant bacterial families were consistent with the significant relative phyla differences observed in specific defense hormone mutants (Figs. 4.1B and 4.5A). In *cpr5*, for example, nine Actinobacteria families were identified with decreased relative abundance, and 12 Proteobacteria families were identified with increased relative abundance, in comparison with wild type (Figs. 4.1C and 4.3, and table S6). These differences demonstrate that defense phytohormones modulate root microbiome composition at multiple taxonomic levels from phylum to family.

We then compared the enrichment and depletion profiles across the mutant genotypes in order to identify shared patterns (Figs. 4.1C, 4.5C and 4.6B; section 4.9.7.3). We used a Monte Carlo test based on the Manhattan distance between enrichment/depletion profiles for each pair of mutants (section 4.9.7.3. Two striking genotype groups were observed at the family level (Fig. 4.1C). Group 1 mutants constitutively produce and accumulate salicylic acid, whereas group 2 mutants either accumulate less salicylic acid or cannot respond to it. These two genotype groups exhibited complementary patterns of differentially abundant Proteobacteria families: In group 1, these were α- and β-Proteobacteria, whereas in group 2, they were γ-Proteobacteria (table S6 and Fig. 4.5A). Within genotype group 2, nearly all of the differentially abundant bacterial families in *sid2* were shared with *pad4* and *dde1 ein2 pad4 sid2*, especially those families depleted as compared with the wild type. Half of the *dde1 ein2 pad4 sid2* depletions were apparently SA-independent (Figs. 4.1D and 4.5B).

We reanalyzed the data to ask whether the differential family abundances observed in specific mutant groups remained consistent at higher taxonomic (OTU) resolution (Fig. 4.6; table S4, tab B; table S7; section 4.9.7.3). We largely recapitulated mutant groups 1 and 2 at OTU resolution (Fig. 4.6A-B). If the plant selected bacteria at a low (genus or species) taxonomic level, we would expect that only one or a few abundant OTUs would drive, and thus correlate with, family-level analyses. However, we observed that a number of OTUs from across the abundance range matched family-level enrichment profiles (Fig. 4.6C-F). These results suggest that defense phytohormones, particularly salicylic acid, modulate taxonomic



	cpr1	cpr5	cpr6	snc1	DEPS	pad4	sid2	JEN	EN	NJ
cpr1										
cpr5	0.544									
cpr6	0.249	0.003								
snc1	0.176	0.547	0.194							
DEPS	0.816	0.542	0.764	0.274						
pad4	1.000	0.553	1.000	0.158	0.002					
sid2	0.709	0.610	0.648	0.043	<0.001	<0.001				
JEN	1.000	1.000	1.000	0.335	0.948	0.313	0.806			
EN	1.000	1.000	1.000	1.000	0.248	1.000	0.578	1.000		
NJ	0.328	0.822	0.327	0.327	0.111	0.234	0.167	0.163	0.085	

Figure 4.5: **Genotype differentially abundant (DA) family enrichments and depletions**. **A** Grid depicting the abundances for each family (grey scale), illustrating the overlap of differentiating families that are either enriched (red outline) or depleted (blue outline) in each mutant compared to the Col-0 abundances organized by phyla (indicated on the left side). **B** Dots represent families with a significantly different abundance between each mutant and Col-0 according to the ZINB model. The magnitude of these predictions is represented by the estimate on the Y-axis with enriched family represented by positive numbers and depletions represented by negative numbers. **C** Table of the p-values of the Monte Carlo testing of Manhattan distances between the enrichment and depletion profiles for each genotype pairing. The significance level for the pale yellow cell is p-value < 0.05, while the significance level for the dark yellow cells are p-value < 0.003.

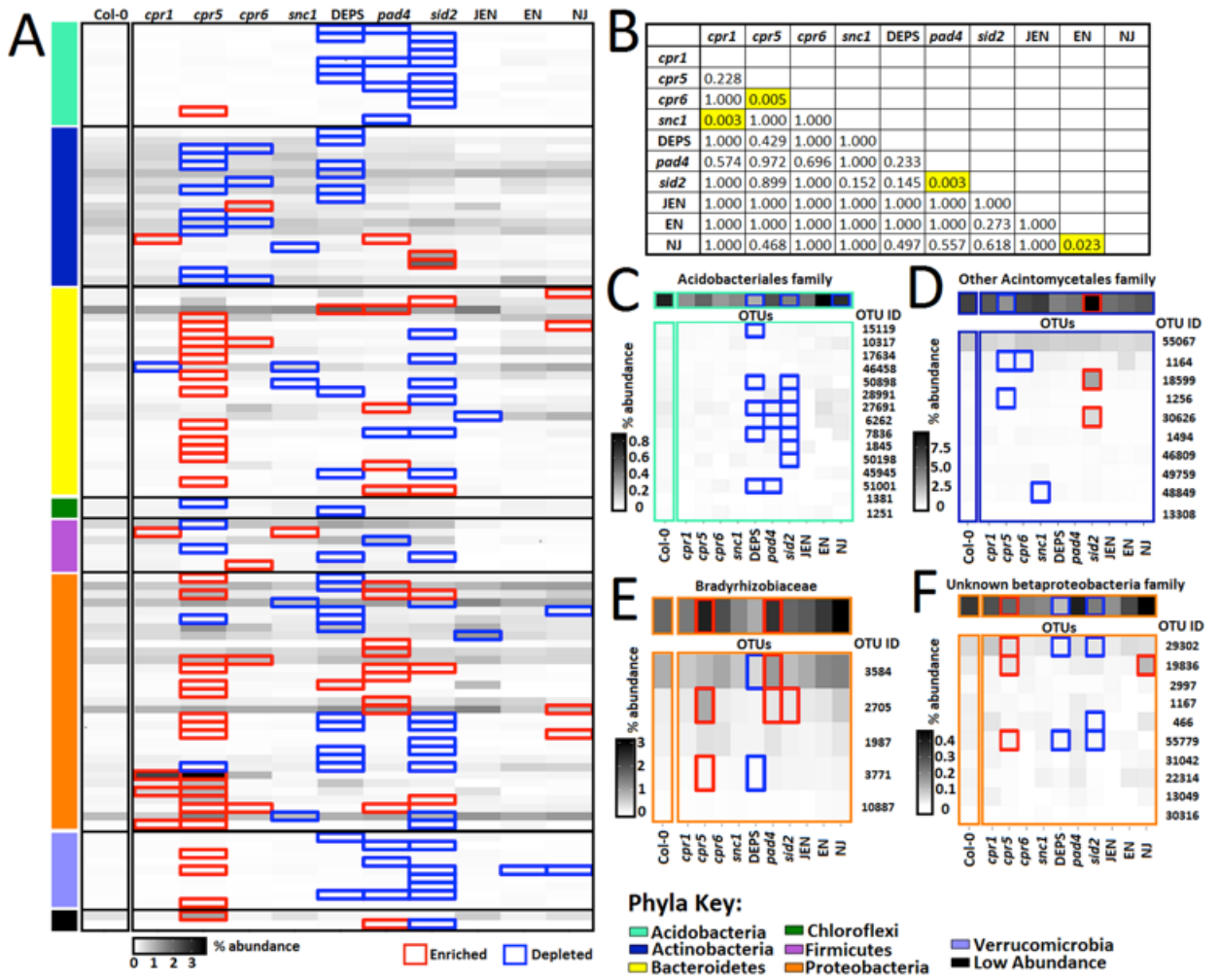


Figure 4.6: **Genotype differentially abundant (DA) OTU enrichments and depletions**. **A** Grid depicting the abundances for each OTU (grey scale) illustrating the overlap of differentiating OTUs that are either enriched (red outline) or depleted (blue outline) in each mutant compared to the Col- 0 abundances organized by phyla (indicated on the left side, color-coded to Figure 1). (B) Table of the p-values of the Monte Carlo testing of Manhattan distances between the enrichment and depletion profiles for each genotype pairing. The significance level for the yellow cell is p-value < 0.05. The majority of families defined above are represented by only one *measurable* OTU so we focused on families with at least five measurable OTUs to address consistency between OTU and family level analyses. **C-F** Grids depicting the abundances of individual OTUs (grey scale), illustrating the overlap and consistency of differentiating OTUs that are either enriched (red outline) or depleted (blue outline) in each mutant compared to the Col-0 abundances within four families (grey scale above each grid) from figure S6: Acidobacteriales family (C), other Acintomycetales family (D), Bradyrhizobiaceae (E), and unknown β-proteobacteria family (F).

groups of bacteria at the family level in the root, and not by altering the abundance of a small number of dominant strains within each differentially abundant family.

4.4 Phytohormone mutants have an abnormal core microbiome

We next asked whether the bacterial families affected by the plant defense phytohormone mutants corresponded to taxa that were normally either enriched or depleted in wild-type roots compared with bulk soil. To that end, we sought to define a set of bacterial enrichments and depletions that are robust to technical choices. We re-sequenced a subset of samples from a single experiment (3 soil samples, 3 Col-0 R samples, and 90 EC samples from nine genotypes, table S3) using the Illumina MiSeq platform and two different hypervariable regions of the 16S rRNA gene (Fig. 4.7A; section 4.9.3.4). We used the same ZINB model approach on each dataset to identify family-level enrichments with respect to soil while controlling for batch effects within each platforms. We noted that the Illumina MiSeq gave much more consistent results, regardless of the variable region, than the Roche 454 instrument. Nevertheless, both platforms and all variable regions recapitulated the differences between EC and R samples and the α- and β-diversity patterns (Fig. 4.17). Further, we found that even when different 16S rRNA gene regions are assessed across sequencing platforms (MiSeq V4 vs. 454 V8), the correlation between taxonomic profiles is $\sim$80% (Fig. 4.7B). Finally, bacterial families that were enriched or depleted consistently in all three bacterial datasets (Illumina V4, Illumina V8 and Roche 454 V8) according to the ZINB model (Fig. 4.7C-D; Table S4) were considered to be 'technically robust', and represent a core set of enrichments and depletions that are insensitive to technical variation and thus are likely to represent true biological differences.

We identified 19 enriched and 23 depleted families in endophytic samples of wild-type roots compared with soil (Fig. 4.7C-D; table S8). Consistent with phyla-level analyses (Fig. 4.1B), 79% of the bacterial families enriched in endophytic samples were Actinobacteria or Proteobacteria. Further, 55% of the endophytic-enriched families in SA mutants are Actinobacteria or Proteobacteria (tables S6 and S8). A similar pattern was observed in

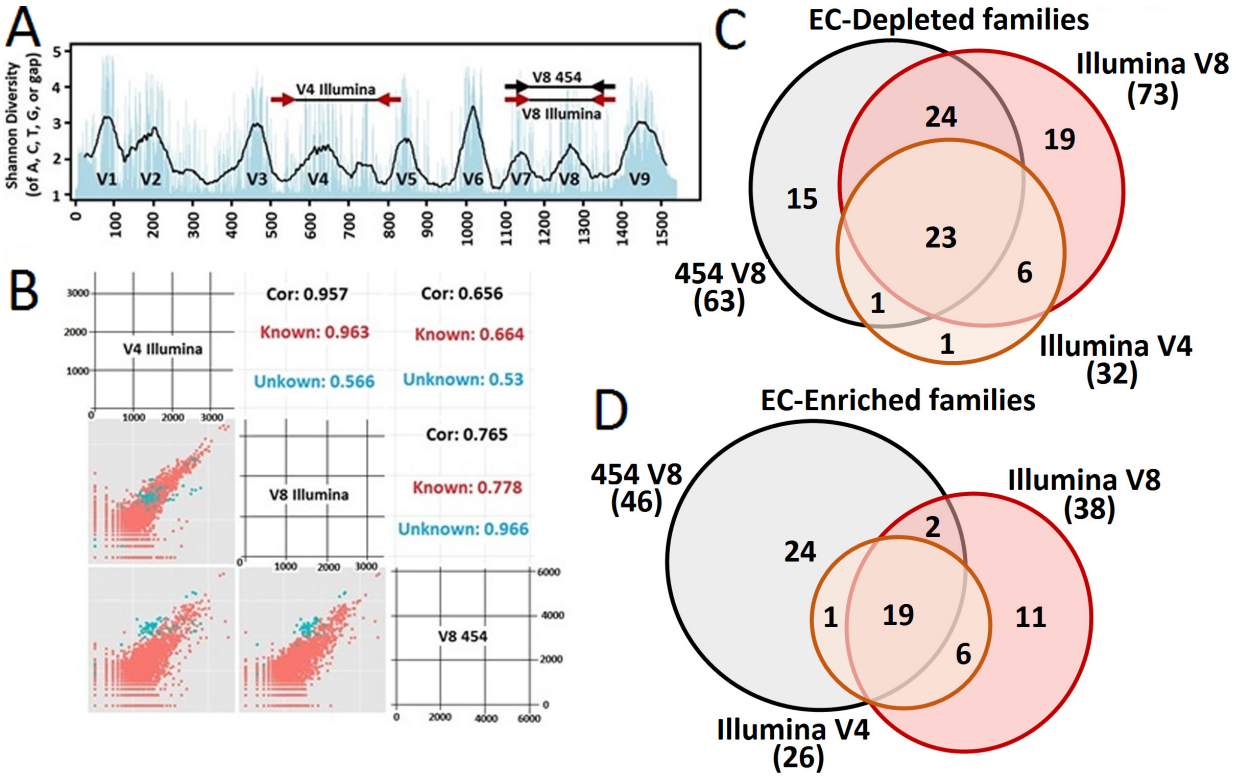


Figure 4.7: **Technical reproducibility between variable regions and sequencing platforms**. **A** A schematic of the three 16S rRNA gene sequencing strategies used. **B** The reproducibility of family-level abundances between each sequencing strategy pairwise comparison for both taxonomically known (red dots) and unknown (blue dots) families with the calculated correlation. Venn diagrams showing the overlap of EC-depleted (**C**) and EC-enriched (**D**) families. The 19 EC-enriched and 23 EC-depleted families in all sequencing strategies are listed in table S8.

the OTU-level analysis, in which 42 and 48% of the endophytic-enriched bacterial families contained at least one OTU that is further enriched in the phytohormone mutants (tables S7 and S8).

Six of the 19 endophytic-enriched families (table S8) were depleted in the *cpr5* mutant that constitutively produces salicylic acid (table S6), suggesting that these six bacterial families are sensitive to SA or SA-dependent processes. Five different endophytic-enriched families (table S8) were further enriched in group 2 mutants that lack salicylic acid biosynthesis and signaling (table S6). Thus, these five bacterial families are candidates for taxa whose colonization is normally limited by wild-type levels of SA and/or SA-dependent processes. In contrast, 12 of the 23 endophytic-depleted families (table S8) were further depleted in group 2 mutants but not in group 1 mutants. Hence, these endophytic-depleted families may require SA-dependent processes to maintain even their very low abundance in the wild-type endophytic compartment (tables S6 and S8). Thus, salicylic acid is required to modulate the assembly of a normal root microbiome. In its absence, core root bacterial community composition is substantially altered. However, these changes to the bacterial microbiome are not sufficient to alter survival of these mutants in this particular wild soil.

4.5 Microcosm recapitulation of the root microbiome

We then asked whether bacteria isolated from roots can colonize sterile roots in the context of a defined but complex synthetic bacterial community. In order to validate our sequencing results and associations found by the ZINB analysis of our census data, we performed three independent microcosm reconstitution experiments (table S3). Each experiment consisted of sterile *A. thaliana* seedlings (wild type and defense phytohormone mutants), planted in a sterilized calcined clay substrate, and inoculated with a simplified synthetic community (SynCom) of bacteria (section 4.9.6).

The synthetic community was composed of a mix of 38 bacterial strains (table S9) that could each be readily differentiated by 16S rRNA gene amplicon sequencing; these isolates were isolated from surface sterilized *A. thaliana* roots grown in either MF soil, or another

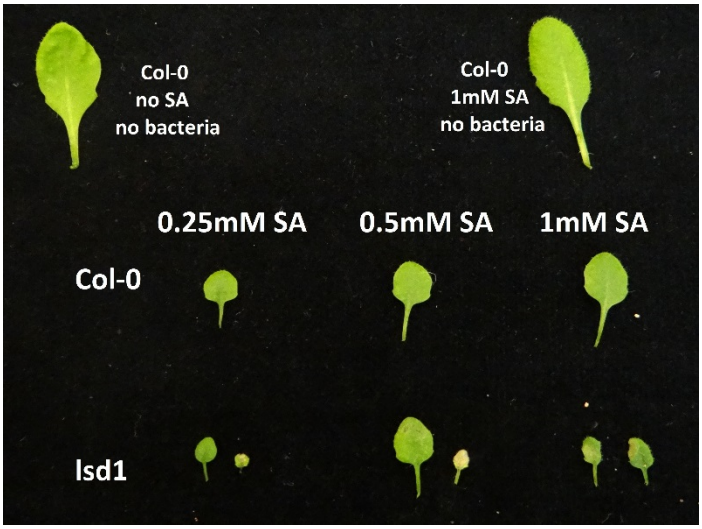


Figure 4.8: **Induction of Runaway Cell Death (RCD) in lsd1 mutants grown in the SynCom with salicylic acid treatment of leaves**. Col-0 and *lsd1* were grown in SynCom. 0, 0.25 mM, 0.5 mM, or 1 mM salicylic acid (SA) was applied to their leaves. 96 hours later RCD was assessed.

previously characterized wild soil from Clayton, North Carolina (Lundberg et al., 2012), plus laboratory *E. coli* DH5α (table S9). Strains were selected from a set of isolates in order to maximize the number of strains with differentiable 16S rRNA genes so that they could be accurately quantified via amplicon sequencing. Sixteen SynCom strains (table S9) were members of 10 families enriched in endophytic compartments of wild-type plants as compared with soil (table S8), and 18 strains matched family OTUs altered in plant defense hormone mutants (tables S6 and S9). Further, 21 of the 38 strains belonged to families that matched endophytic-enriched OTUs from a published census of plants grown in wild Mason Farm soil (Lundberg et al., 2012).

We applied exogenous salicilyc acid (0.5 mM) every 3 days to leaves and soil of additional plants as part of our 8-week synthetic community experiment. Roughly half of the samples for each experiment were sprayed with 0.5 mM salicylic acid every 3 days, which is above physiological levels (Bi et al., 1995), but can induce systemic acquired resistance (Spoel and Dong, 2008). This treatment can also induce runaway cell death in a mutant that is hyper-responsive to salicylic acid via activation of an immune receptor, *lsd1* (Bonardi et al.,

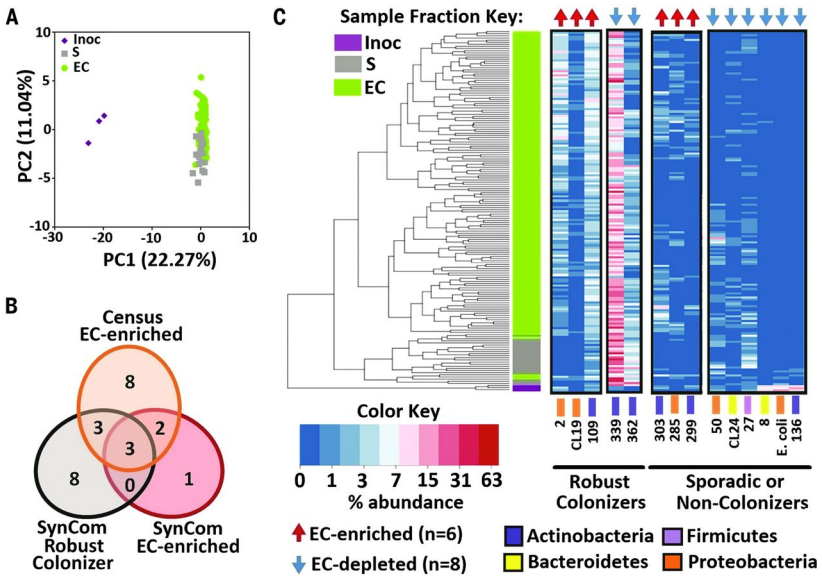


Figure 4.9: **A 38-member synthetic community recapitulates differentiated microbiome colonization**. **A** Principal coordinates analysis showing the inoculum (purple diamonds), soil (gray squares), and EC (green circles) samples. **B** The overlap of SynCom members that were robust colonizers of Col-0 EC (black), EC-enriched (red), or matched EC-enriched families from the census of roots grown in wild Mason Farm soil (orange) (Fig. 4.1). **C** Hierarchical clustering and heat map showing percent abundance ($\log_2$-scale) of selected isolates. Sample clustering splits by fraction (left) and EC samples grouping by biological replicate. Isolates are grouped by their presence in the majority of Col-0 EC samples (Robust colonizers) or absence in the majority of Col-0 EC samples (sporadic or non-colonizers). Isolates are color-coded to phyla as in Fig. 4.1. Isolates that were significantly more abundant (red arrows) or less abundant (blue arrows) in EC with respect to bulk soil are denoted along the top.

2011). Under our synthetic community experiment control, this treatment elicited runaway cell death in control *lsd1* plants (Aviv et al., 2002), but not Col-0 leaves 96 hours after spraying (Fig. 4.8). Plant roots and bulk soil controls were harvested when an inflorescence meristem formed. Unlike the Mason Farm soil experiments, only EC and bulk soil fractions were collected due to the granular texture of the calcined clay that made rhizosphere harvest difficult.

Both bulk soil and endophytic compartment microbiomes changed over 8 weeks after SynCom inoculation (Figs. 4.9A and 4.10). Fourteen of the 38 SynCom strains were 'robust colonizers' (Fig. 4.10C, table S9, section 4.9.7.5). Six of these 14 are from families predicted to be endophytic-enriched in roots from our Mason Farm soil census (Fig. 4.9B, overlapping

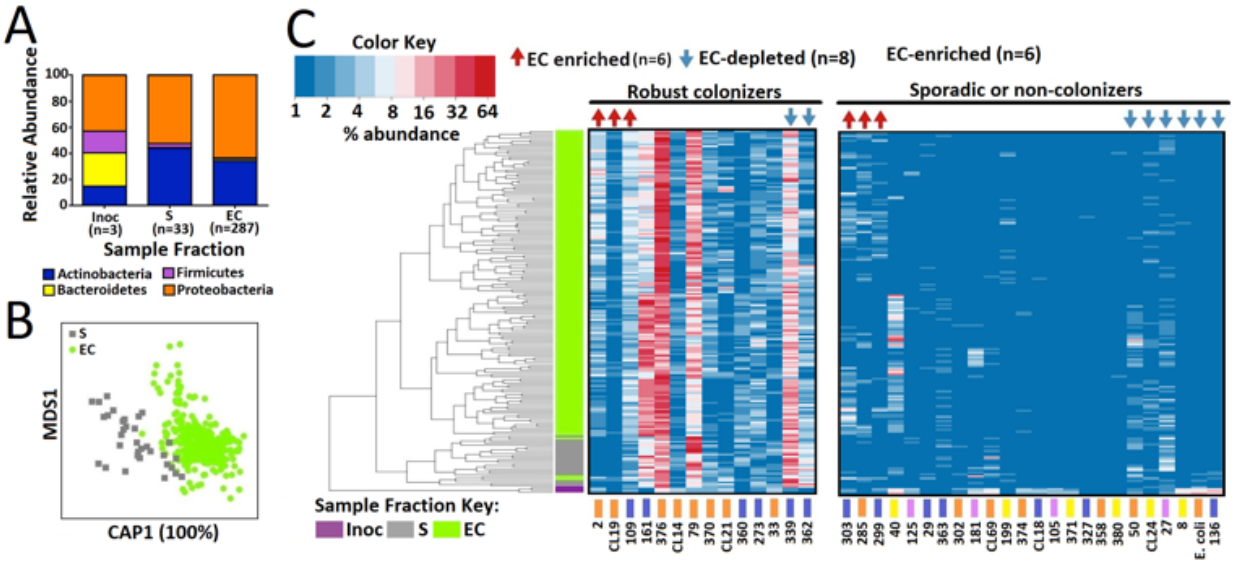


Figure 4.10: **Synthetic community differentiates sample fractions**. **A** Phyla distributions in the synthetic community (SynCom) inoculum, soil, or EC fraction samples from all genotypes. **B** CAP analysis to showing the contribution of sample fraction to overall community composition. **C** Hierarchical clustering and heat map showing percent abundance ($\log_2$ scale) of selected isolates. Sample clustering split by fraction (left), with EC samples grouping by biological replicate. Isolates are grouped by their presence in the majority of Col-0 EC samples (Robust colonizers) or absence in the majority of Col-0 EC samples (Sporadic or non- colonizers). Isolates color-coded to phyla as in Fig. 4.1. Isolates that were significantly more abundant (red arrows) or less abundant (blue arrows) in EC with respect to bulk soil are denoted along the top.

black and orange circles; table S9), corroborating their ability to colonize roots. We identified six 'SynCom EC-enriched' isolates and eight 'SynCom EC-depleted' isolates (Fig. 4.9C; table S4e; section 4.9.7.6). Five of the six SynCom EC-enriched strains belong to families also predicted to be endophytic-enriched in roots from the Mason Farm soil census (Fig. 4.9B, overlapping orange and red circles, and table S9), supporting their categorization as endophytic compartment-enriched families (table S8). Thus:

1. Some but not all SynCom isolates robustly colonized the endophytic compartment of host plants in these mesocosms.

2. The soil and endophytic microbiomes still differed in this context.

3. There was considerable overlap in enrichments and depletions between the SynCom and wild soil colonization experimental platforms at the family level.

4.6 Salicylic acid modulates the abundance of specific isolates

Seven bacterial isolates were differentially abundant between wild type and the defense phytohormone mutants in the synthetic community experiments (Fig. 4.11; section 4.9.7.6), including at least one representative from each of the four phyla present in the inoculum (table S9). Six of the seven isolates were either depleted (*Streptomyces sp.* 136, *Chryseobacterium sp.* 8, *Pseudomonas sp.* 50, and *Escherichia coli*) or were sporadic or noncolonizers (*Bacillus sp.* 125 and *Brevundimonas sp.* 374). Four of these six overlapped with families predicted to be differentially abundant across genotypes in our Mason Farm soil census (Fig. 4.11 and table S6), and six of seven (all except *Bacillus sp.* 125) were enriched in the defense phytohormone mutants (Fig. 4.11C). The profiles of differentially abundant isolates in *pad4* and *sid2* mutants overlapped (Fig. 4.11C).

Exogenous SA application to our SynCom experiments also affected bacterial community composition in both bulk soil and endophytic compartment samples (CAP 0.3 to 1.5%) (Fig. 4.12A; table S5; sections 4.9.7.4 and 4.5), which is consistent with rhizosphere changes in plants treated with salicylic acid or jasmonic acid (Carvalhais et al., 2014; Doornbos et al.,

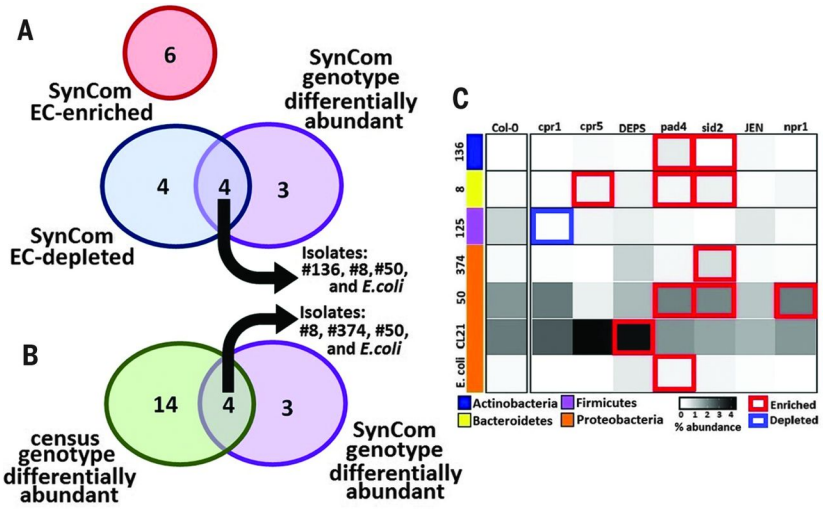


Figure 4.11: **Defense phytohormone mutants exhibit increased abundance of EC-depleted microbes**. **A** Overlap of SynCom EC-depleted (Fig. 4.9C) and SynCom isolates differentially abundant in defense phytohormone mutants (SynCom genotype differentially abundant). No SynCom EC-enriched isolates (Fig. 4.9B-C) were affected by plant genotype. **B** Overlap of the same SynCom genotype differentially abundant isolates from (A) compared with isolates present in the SynCom from families that were genotype differentially abundant in the wild soil census (green circle) (table S8). **C** Heat map of isolates (color-coded by phylum as in Fig. 4.1) differentially abundant between defense phytohormone mutants and Col-0. Grayscale shows the mean abundance of the corresponding isolate (rows) in the EC of a given genotype (columns). Genotype differentially abundant families predicted as enriched or depleted by the ZINB model are boxed in red or blue, respectively (supplementary materials, materials and methods 6f).

2011). Two isolates were enriched [*Flavobacterium sp.* 40 (Bacteroidetes) and *Terracoccus sp.* 273 (Actinobacteria)] and one depleted [*Mitsuaria sp.* 370 (β-Proteobacteria)] in the presence of exogenous salicylic acid (Fig. 4.12B; table S9; section 4.9.7.4).

These data integrate our synthetic community experiments with our wild soil census and demonstrate increased abundance in the SA-deficient mutants of isolates that were 'sporadic or non-colonizers' across all wild soil endophytic samples. Thus, altering SA production and signaling in the host plant prevents it from fully excluding bacterial taxa that a wild-type plant shuns.

4.7 Reconstituting the effect of salicylic acid *in vitro*

We then asked whether the effect of salicylic acid on our syntheti community experiments is direct, and thus can be recapitulated in a further simplified system. To that end, we performed *in vitro* growth curves with varying concentrations of salicylic isolates with selected isolates. *Terracoccus sp.* 273 abundance was higher in both SA-treated bulk soil and root endophytic samples (Fig. 4.13A), and its growth was enhanced by SA in liquid media (Fig. 4.13B; section 4.9.6.3), although its genome contains no obvious SA catabolism genes (taxon IDs in table S9; section 4.9.7.7). In contrast, *Mitsuaria sp.* 370 was depleted in endophytic samples treated with SA and grew less well in its presence (Fig. 4.13C-D). *Streptomyces sp.* 303 was weakly enriched in SA-treated samples (q-value < 0.07) (Fig. 4.13E), grew on minimal media with 0.5 mM SA as a sole carbon source (Fig. 4.13F), and contains orthologs to a previously characterized *Streptomyces* SA-degradation operon (Fig. 4.12D; table S9; section 4.9.7.7). Among two other Streptomyces strains in the SynCom inoculum (#136, #299; table S9) and two additional Actinobacteria that were significantly associated with salicylic acid treatment prior to multiple testing correction (#29 and #362), the only obvious salicylic acid metabolism gene was an salicylic acid dioxygenase found in *Arthrobacter sp.* #362 (Ferraroni et al., 2013; Hintner et al., 2001) (section 4.9.7.7). Thus, the broader effects of SA on microbiome composition consist of both direct and indirect effects on the physiologies of individual community members from limited, specific taxa.

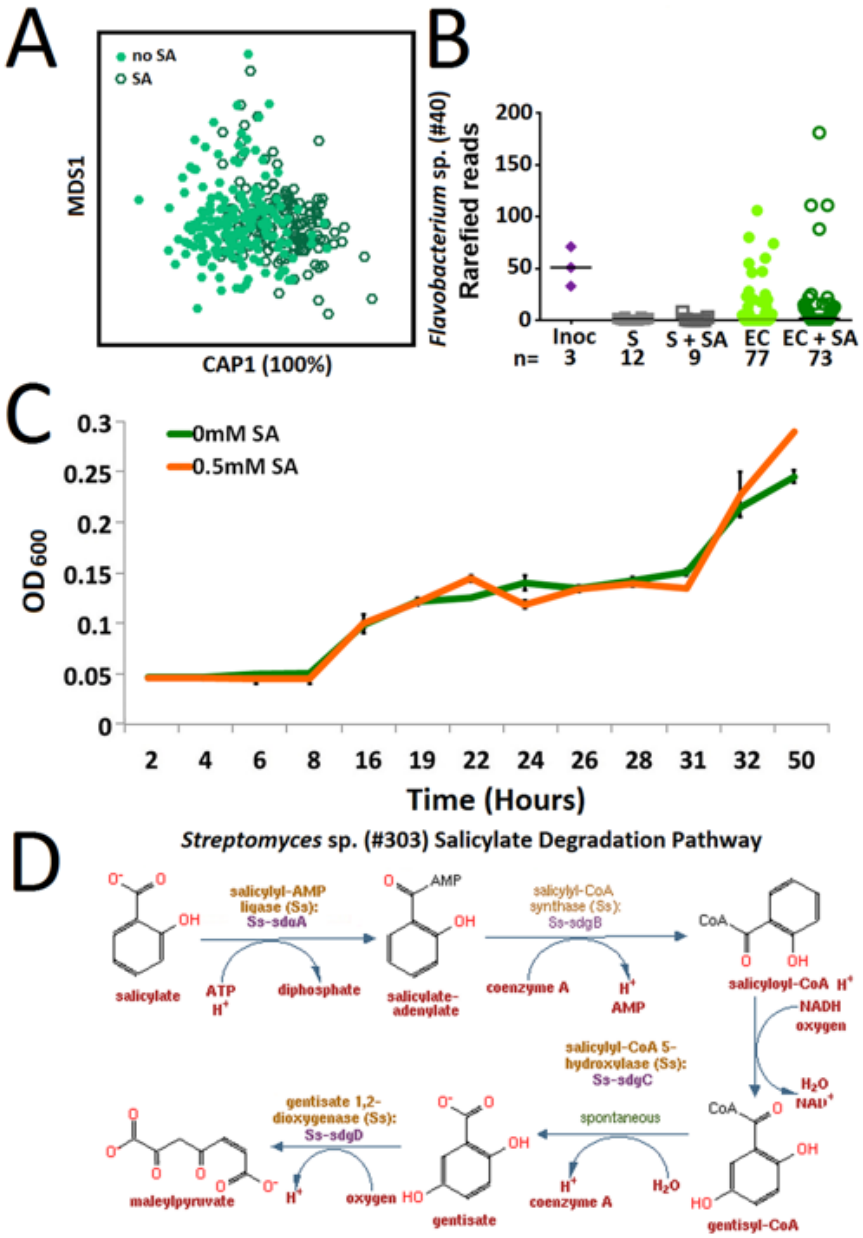


Figure 4.12: **Salicylic acid treatment affects SynCom composition, but did not affect growth of *Flavobacterium sp.* #40 in SynCom or in liquid growth curves**. **A** CAP analysis of the full count matrix to identify the contribution of salicylic acid (SA) treatment to community composition. **B** Dot plot of 400 rarefied consensus sequences from isolate #40 from synthetic community inoculum (purple diamonds), soil (grey squares), and EC samples (light/dark green circles) for both salicylic acid (SA) treated (open symbols) and untreated (closed symbols). No group of samples were significantly different from any others. **C** Optical density of isolate #40 grown in phosphate buffered 1/10 LB with either 0 (green line) or 0.5 mM (orange line) salicylic acid (SA) added. **D** Salicylate degradation pathway (MetaCyc) present in *Streptomyces sp.* (#303) genome contains all 4 genes in this pathway (% identities to each: sdgA-98%, sdgB-98%, sdgC-96%, and sdgD-94%).

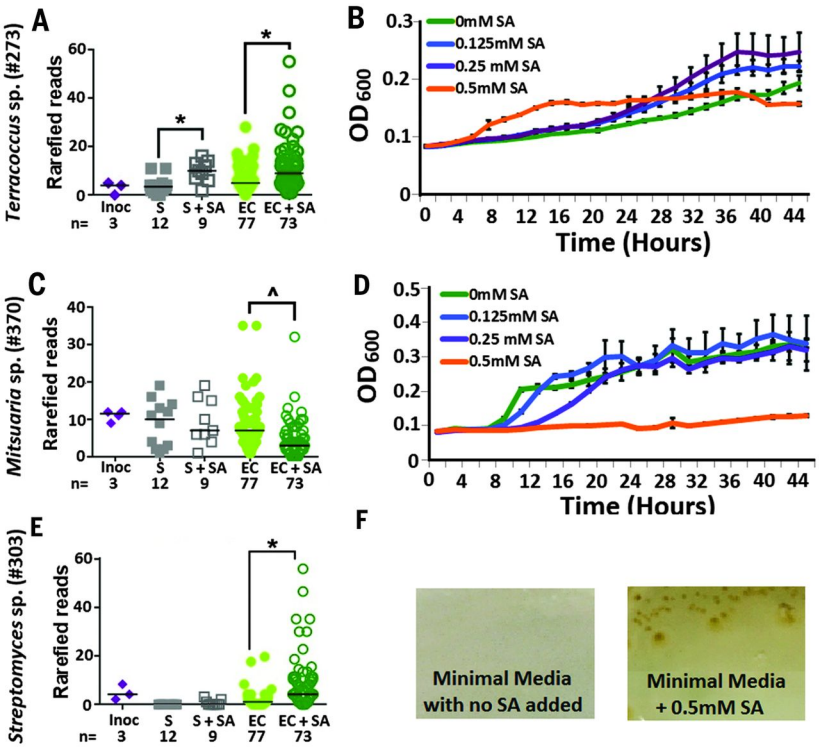


Figure 4.13: **Salicylic acid directly affects synthetic community isolates**. **A** *Terracoccus sp.* (273) reads from 400 rarefied consensus sequences for the SynCom inoculum (purple diamonds), soil (gray squares), and EC samples (green circles) from SA-treated (open symbols) and -untreated (closed symbols) plants. Asterisk indicates significantly different between sample treatments at p-value < 0.006 by Mann-Whitney test. **B** Optical density of *Terracoccus sp.* (273) grown in buffered 1/10 LB with 0 (green), 0.125 (blue), 0.25 (purple), or 0.5 mM (orange) SA added. **C** *Mitsuaria sp.* (370) reads as in (A). Caret indicates significantly different between EC sample treatments at p-value < 0.0001 by means of Mann-Whitney test. **D** Optical density of *Mitsuaria sp.* (370) grown as in (B). **E** *Streptomyces sp.* (303) reads. Asterisk indicates significantly different between EC sample treatments at p-value < 0.001 by means of Mann-Whitney test. **F** Streptomyces sp. (303) aggregates in liquid cultures but grows on minimal media agar with 0.5 mM SA as the sole carbon source.

4.8 Conclusion

We demonstrate that plant defense phytohormones sculpt the root microbiome in characteristic ways. Elimination of all three defense phytohormone signaling sectors results in abnormal microbial profiles in the root, which may be linked to lowered survival in a wild soil. Salicylic acid, a key immune regulator in leaves, also modulates the composition of the root microbiome. Plants with altered salicilic acid signaling have root microbiomes that differ in the relative abundance of specific bacterial families as compared with those of wild type. It will be of interest to address whether and how the extra- and intracellular plant immune system receptor systems further condition root bacterial community composition. We demonstrated that different bacterial strains could make use of salicylic acid in different ways, whether as a growth signal or as a carbon source. Thus, salicylic acid influences the microbial community structure of the root. This may occur by gating bacterial taxa as a consequence of salicylic acid function in homeostatic control of immune system outputs, or via as-yet-undefined effects on microbe-microbe interactions and root physiology. Together, our results show that a central regulator of the plant immune system, largely uncharacterized in the root, directly influences root microbiome composition. Our results could open new avenues for modulating the root microbiome to enhance crop production and sustainability.

4.9 Supplemental information

4.9.1 Plant measurments

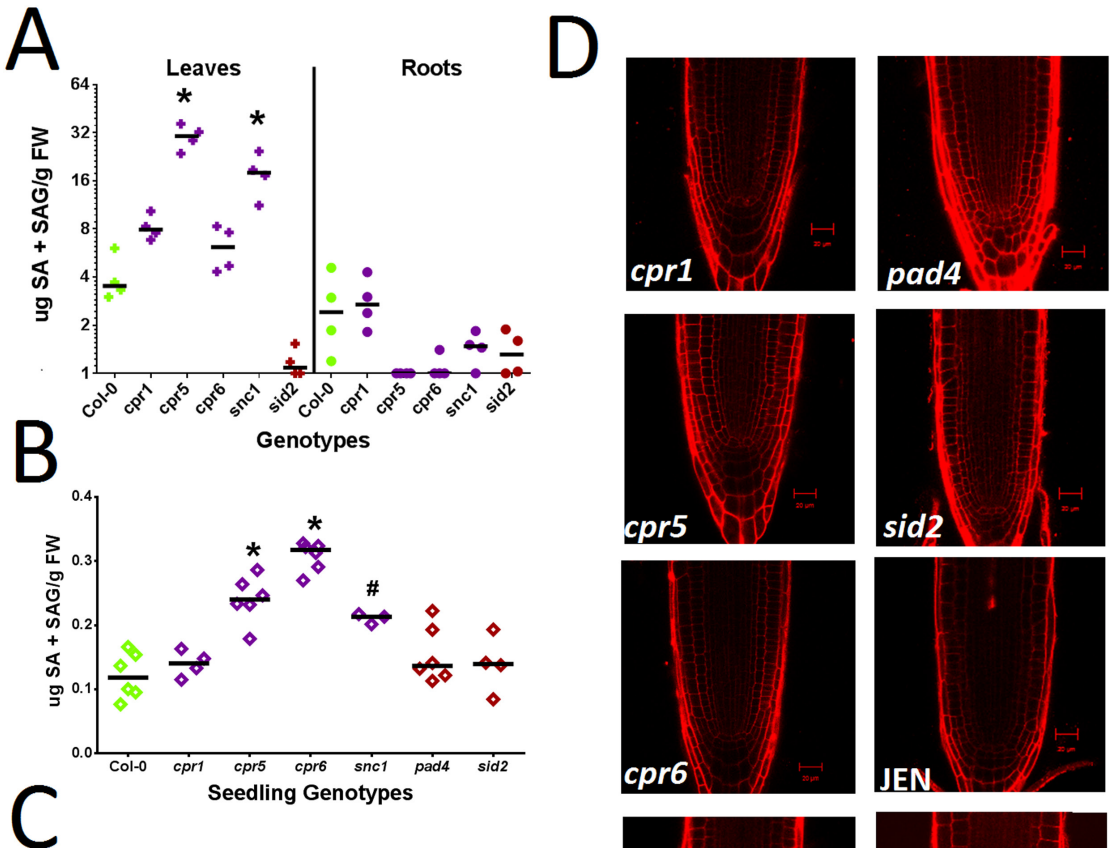
4.9.1.1 Measuring salicylic acid production in leaves and roots Previously, production of salicylic acid has been measured in the leaves of many of the mutants used in this study (Bowling, 1994; Clarke et al., 1998; Kirik et al., 2001; Zhang et al., 2003). For measuring salicylic acid production in the leaves and roots in MF soil, hyper-responsive mutants (cpr1, cpr5, cpr6, and snc1) as well as negative control (sid2) and isogenic wild type (Col-0) were grown in Mason Farm soil as described in section 4.9.2.2 with the exception that 4-5 seedlings of each genotype were grown in a 4.5 pot together to increase the amount of plant material harvested for each sample. When the inflorescence meristem formed, plants

were harvested and four 10 0mg samples of leaves and roots were taken. We also grew seedlings axenically for 18 days in vertical plates as described in section 4.9.1.2 for tissue to measure salicylic acid accumulation in axenic conditions. In all cases samples were snap frozen and stored at -80°C until SAG levels were assessed biochemically. Briefly, the levels of total salicylic acid and salicylic acid glucoside (SAG) were determined for each genotype using the *Acinetobacter sp.* ADPWH_lux biosensor (DeFraia et al., 2008).

4.9.1.2 Measuring root length and morphology For both root length and root morphology measurements, surface sterilized mutant seeds and control seeds were grown vertically on plates containing 1/2 strength Murashige and Skoog (MS) salt mixture, 1% sucrose, 2.5 mM 2-(N-morpholino) ethanesulfonic acid (pH 5.7), and 0.5% phytoagar for 7 days. Root lengths were measured using ImageJ (Abramoff et al., 2004), and the Student's t-test was used to determine statistical significance (Fig. 4.14C). Seedlings were stained in 10 mg/ml propidium iodide for 0.5 to 2 minutes and mounted in water. Imaging was on a Zeiss LSM710 confocal laser-scanning microscope using the 488-nm laser line for excitation and a 40x water objective (Fig. 4.14D).

4.9.2 Census study experimental procedures

4.9.2.1 Soil collection and preparation For each experiment, we collected the top 20 cm of earth from Mason Farm (MF), which is managed by the North Carolina Botanical Garden. This site is free from pesticide and fertilizer use and has low human disturbance, providing a fairly stable soil source. Soil micronutrient analysis was used to define this as a loam soil with a variety of nutrients and a pH of 6 (Lundberg et al., 2012). Soil was dried and crushed using an aluminum mallet. After crushing, debris was removed by sifting, resulting in a very fine soil. To improve drainage, soil is mixed 2:1 volume with steamed and autoclaved sand. The resulting soil mixture is used to grow plants in 2 x 2 inch square pots for each experiment.



C

Genotype	Normalized root length	Root Morphology
Col-0	1 ± 0.1	Normal
cpr1	1 ± 0.1	Normal
cpr5	1 ± 0.1	Normal
cpr6	0.9 ± 0.2	QC & columella disorganization
snc1	1 ± 0.2	QC divisions
DEPS	1 ± 0.2	QC defects, possible early endo divisions
pad4	Not measured	QC & columella defects
sid2	1 ± 0.1	Normal
JEN	0.8 ± 0.2	QC divisions, shorter meristems
EN	Not measured	Possible shorter meristems
NJ	Not measured	QC defect

Figure 4.14: **Salicylic acid production in MF soil and root morphology of defense phytohormone mutants**. **A** Representative of salicylic acid (SA) measurements performed three times in leaves and roots grown in MF soil (n= 4 for each type of sample; section 4.9.1.1). Asterisk indicates statistically higher than Col-0 (p-value < 0.0001) by ANOVA with Bonferroni multiple test correction. **B** Representative of salicylic acid (SA) measurements performed on axenically grown seedlings (n=3-6 for each type of sample; section 4.9.1.1). Asterisk indicates statistically higher than Col-0 (p-value < 0.0001) and pound sign indicates statistically higher than Col-0 (p-value < 0.005) by ANOVA with Bonferroni multiple test correction. **C** Overview of root morphology at the root tip of each defense phytohormone mutant with representative images (**D**) (section 4.9.1.2).

4.9.2.2 Seed sterilization, germination and plant growth All seeds were surfaced-sterilized by treatment with 70% ethanol with 0.1% Triton-X100 for 1 minute, 12 minutes of treatment with freshly made bleach solution (10% household bleach and 0.1% Triton-X100), and 3 rinses with sterile distilled water. Seedlings grown from such seeds have previously been shown to not contain endophytic microbes, and this treatment eliminates any seed-borne microbes on the seed surface (Lundberg et al., 2012). Seeds were stratified at 4°C in the dark for 3 days and germinated on 0.5% agar containing 1/2 strength Murashige and Skoog (MS) vitamins and 1% sucrose for 1 week at 24°C under 18 hours of light. Healthy 1 week old seedlings were aseptically transplanted from the MS germinating plates into sterile 2.5 inch square pots filled with Mason Farm soil prepared as described in section 4.9.2.1. We also included 'bulk soil' controls, which were pots without plants added to them and were randomly interspersed among the planted pots. All pots, including bulk soil controls, were watered from the top with non-sterile distilled water to avoid chlorine and other tap water additives 2-3 times a week. In order to promote large rosette and root growth, plants were grown in growth chambers with short day, 8 hours of light at 21°C and 16 hours of dark at 18°C until the formation of an inflorescence meristem (Lundberg et al., 2012).

4.9.2.3 Harvesting and DNA extraction Plants and bulk soil controls were harvested and their rhizosphere and endophytic compartment microbial communities isolated as previously described (Lundberg et al., 2012). At the formation of an inflorescence meristem, the above ground plant organs were aseptically removed and loose soil was physically removed until only soil within 1 mm from the root surface remained. The roots were placed in a clean and sterile 50 mL conical tube containing 25 mL of phosphate buffer (6.33g of $NaH_2PO_4{*}H_20$, 16.5 g of $Na_2HPO_4{*}H_20$, and 200 μL Silwet L-77 in 1 L of water). Rhizospheres (R) were separated from the roots by vortexing the root system in buffer at maximal speed for approximately 15 seconds. The resulting turbid solution was filtered through a sterile 100 μm nylon mesh cell strainer (BD Biosciences) into another sterile 50 mL conical tube to filter out

plant material, sand, and other large debris. The filtrate was centrifuged in 2 steps to form tight pellets (averaging 250 mg), defined as our rhizosphere (R) sample. Bulk soil samples were taken by discarding the top 1 cm of soil from the pot, homogenizing the remaining soil on a sterile work surface, and scooping approximately 250 mg of the mixed soil into a buffer tube and following the same protocol as rhizosphere samples. To isolate the endophytic compartment (EC) microbial community, roots were rinsed in sterile distilled water and debris was aseptically removed with tweezers. Roots were subsequently placed in new sterile phosphate buffer for sonication to remove soil or microbial aggregates remaining on the root surface using a Diagenode Bioruptor set on the low frequency for five minutes (five 30 s bursts followed by five 30 s rests). The clean sonicated roots constitute the EC samples. All Bulk soil, R, and EC samples were flash frozen and stored at -80°C until DNA was extracted with the 96-well format MoBio PowerSoil kit. For the EC samples, we performed a pre-homogenization step by lyophilizing the root samples, placing them in a 2 mL tube with 3 glass beads (4 mm), snap freezing again and running through a cycle on the MPBio FastPrep 24 for 20s at 4.0 m/s. This pre-homogenization allowed us to grind the tissue before adding lysis buffer and ensure that the kit was able to work efficiently.

4.9.3 Massive parallel sequencing library preparations

4.9.3.1 454 16S library preparation and pyrotag sequencing 454 pyrosequencing libraries were created in triplicate using the same protocol as in Lundberg et al. (2012) and sequencing was performed at the Joint Genome Institute and Roche. The raw data from the 454 survey experiment is available in the Short Read Archive (ERP010780), and the processed OTU representative sequences are in Supplementary Dataset 3.

4.9.3.2 Illumina library preparation and sequencing at JGI Three sets of primers were used to amplify the V4 region of the 16S rRNA gene (515F-806R), V8 region of the 16S rRNA gene (1114F-1392R), and ITS intergenic transcribed spacer (ITS4-ITS9) (table S2). In each case, the reverse primer had a unique molecular barcode for each sample. This allowed multiplexing of 92 samples for V4, 48 samples for V8, and 92 for ITS. PCR reactions with

~20ng template were performed with 5 Prime Hot Master Mix in triplicate along with a positive and negative control to reveal contamination. The PCR program used was 94°C for 3 min followed by (94°C for 45 sec, 50°C for 1 min, 72°C for 1.5 min) x 35 cycles, followed by 72°C for 10 min and then cool down to 4°C. Reactions were purified using 1.2X volume of AMPureXP magnetic bead and quantified with Qubit HS assay. Amplicons were pooled in equal amounts following qualitative analysis with a Bioanalyzer. Pooled amplicons were then diluted to 10 μM and submitted for qPCR for quality control. For family-level microbiome comparisons, samples were sequenced on an Illumina MiSeq machine at the Joint Genome Institute with a target cluster density of 500K/mm^2 . Each sample was spiked with approximately 25% PhiX control to increase sequence diversity. The data from the Illumina re-sequencing on the JGI portal at `http://genome.jgi.doe.gov/Immunesamples/Immunesamples.info.html`, and the processed OTU representative sequences are in Supplementary Dataset 3.

4.9.3.3 Illumina library preparation for SynCom experiment Illumina libraries for the SynCom experiments were created using the same protocol as in Lundberg et al. (2013), which allows counting of original template molecules, and sequencing was performed at UNC. The raw data for the SynCom experiments is available in the Short Read Archive (ERP010863).

4.9.3.4 Libraries for technically robust enrichments and depletions Four libraries were prepared: V8, V4 with peptide nucleic acid (PNA) (Lundberg et al., 2013), V4 without PNA and ITS2 (section 4.9.3.2. Each MiSeq lane was multiplexed to 48 samples. Sequences from each lane were run through the DOE JGI iTags pipeline at the DOE JGI for basic quality control. We noted that the two V4 libraries (with and without PNA) gave identical results so we combined them into one abundance table.

4.9.4 Processing of sequencing data

4.9.4.1 Sequence processing pipeline Sequences from each platform, library preparation method and experimental design were first pre-processed as described below (Fig. 4.15;

sections 4.9.4.2, 4.9.4.3 and 4.9.4.4) into a fasta file containing high quality sequences matched to a given sample on the fasta headers. The resulting sequences were then converted into a count table either by clustering into Operational Taxonomic Units (OTUs) (section 4.9.4.5) or mapping to known isolates' 16S rRNA gene (section 4.9.4.6). Representative sequences from OTUs were further taxonomically annotated (section 4.9.4.7). Samples that had less than 1,000 usable reads in the census were pooled in silico with samples of the same fraction, developmental stage, genotype and experiment that also had less than 1,000 usable reads to provide enough depth for statistical analyses (section 4.9.4.8). A number of off-the-shelf tools were used, and in-house Perl scripts filled the gaps (section 4.9.4).

4.9.4.2 Pre-processing Roche 454 census experiments As each 454 plate was sequenced, raw reads from individual plates were immediately run through Pyrotagger (Kunin and Hugenholtz, 2010) to diagnose plate quality (based on the number of reads passing quality checks) and determine if a plate needed to be re-sequenced. Plates with a reasonable number of long, high quality raw reads with matching barcodes were processed and quality controlled following the pipeline defined in Lundberg et al. (2012) (Fig. 4.15). Briefly, reads were trimmed to 220 bp and short reads removed, low quality reads were removed using default quality control settings in QIIME-1.3.0 (Caporaso et al., 2010) with the split_libraries.py script, and individual reads were matched to sequence barcodes.

4.9.4.3 Pre-processing Illumina MiSeq census experiment MiSeq lanes with a high number of sequence pairs matching barcodes and successful merging of paired-ends were used for downstream analysis. An in-house pipeline was implemented in Perl to process these sequences with the following steps:

1. Sequence pairs were identified and unpaired sequences were discarded.

2. Reads were trimmed to 165bp and merged using FLASH (Magoč and Salzberg, 2011) (options: -m 30 -M 165 -x 0.25 -r 165 -f 282 -s 20), any read pair that did not merge

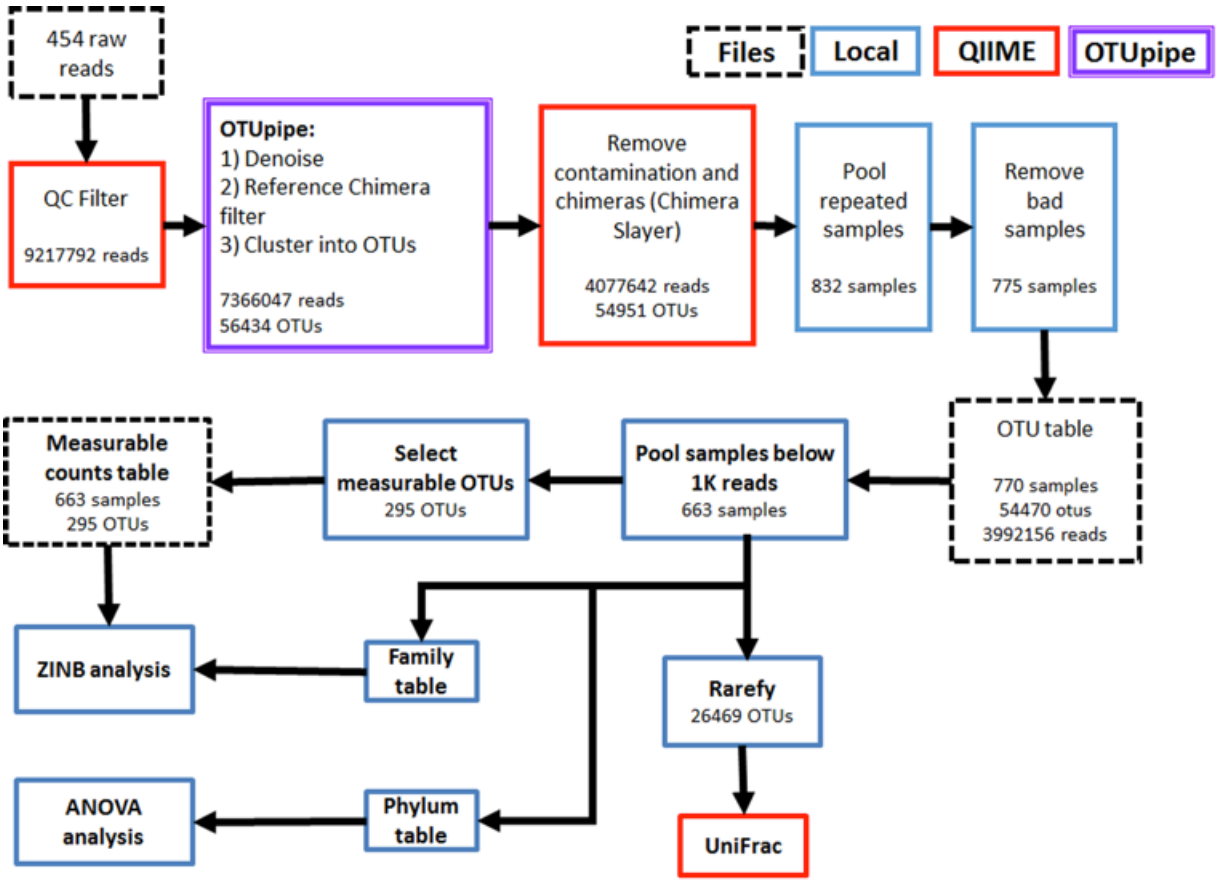


Figure 4.15: **Processing pipeline for Roche 454 census experiments**. This flowchart is order of events that occur in processing the sequencing data. Boxes with dashed black lines represent files. Boxes with blue lines describe events that occur locally using custom scripts. Boxes with red lines describe steps that occur through QIIME. Boxes with double purple lines describe events that occur using OTUpipe. For full details see supplementary information (Method 3).

was discarded.

3. Expected primer sequences were matched to the merged sequences using standard regular expression techniques, primer sequences were removed and the resulting 'in silico amplicons' were kept; any sequences without primer matching were discarded.

4. For the V4 region only, sequences shorter than 240 bp were removed because the primers used for this region also amplify oomycete mitochondrial genes.

5. Sequences were de-multiplexed.

4.9.4.4 Pre-processing Illumina MiSeq synthetic community experiments Libraries were prepared following the protocol from Lundberg et al. (2013). MiSeq reads were processed with MT-Toolbox (Lundberg et al., 2013; Yourstone et al., 2014). Briefly, sequence pairs were merged with FLASH (Magoč and Salzberg, 2011) and merged sequences were binned by molecule tag (MT). The resulting bins were used to correct for PCR and sequencing errors and biases. Only MTs with at least 3 merged sequences were kept for downstream analysis.

4.9.4.5 Clustering sequences into OTUs For the census experiments (both from 454 and MiSeq), the high quality sequences were clustered into Operational Taxonomic Units (OTUs) using custom made implementation of OTUpipe (`http://www.drive5.com/usearch/manual/otupipe.html`) with USEARCH6 (Edgar, 2010). Our implementation performs the following steps:

1. De-replicate sequences.

2. De-noising by clustering at 99% identity.

3. Cluster de-noised sequences at 97% to define OTUs.

4. Identify chimeric sequences using both a reference-based and a de-novo chimera detection step.

Sequences from Roche 454 were further scanned for chimeric OTUs using ChimeraSlayer (Haas et al., 2011) as implemented in QIIME (Caporaso et al., 2010). The number of reads matching a given OTU were counted for each sample and a count table was generated for each set of libraries (454, Illumina V4 with PNA, Illumina V4 without PNA, Illumina V8 and Illumina ITS2). Comparison of the Illumina V4 with PNA and V4 without PNA showed a very high degree of reproducibility (not shown) and thus the resulting count tables were combined to generate a single Illumina V4 count table.

4.9.4.6 Mapping MT consensus to isolate 16S genes For the synthetic community experiments, every high quality consensus sequence produced by MT-Toolbox (Yourstone et al., 2014) was mapped with BWA version 0.7.10-r78 (Li and Durbin, 2009) to a reference set of sequences made up of the Sanger 16S sequence from the 38 isolates in the synthetic community, as well as to known plant nuclear and organellar rRNA genes. Up to 3 mismatches were allowed during mapping and the number of consensus sequences matching to each isolate or host sequences were used to create a count table for downstream analysis.

4.9.4.7 OTU and isolate annotation We profiled the bacterial and the fungal communities by high-throughput sequencing of segments of the 16S rRNA gene and intergenic transcribed spacer (Figs. 4.7 and 4.17; sections 4.9.3.1 and 4.9.3.2). For each prokaryotic dataset (454 V8, Illumina MiSeq V8 an Illumina MiSeq V4), representative sequences from each OTU were given a taxonomic annotation using the RDP classifier (Wang et al., 2007) as implemented in QIIME 1.3.0. The 2011/02/04 Greengenes database was used as a training set. OTU representative sequences were also BLASTed (Camacho et al., 2009) against: i) a modified Greengenes database that includes plant and oomycete-derived sequences, and ii) the GOLD database (`http://drive5.com/uchime/gold.fa`). Any OTU annotated as plant, archaea or oomycete-derived (nuclear or organellar) by any of the three methods was removed

from downstream analysis. For the fungal ITS dataset, OTUs were classified by BLAST against the UNITE database (`https://unite.ut.ee/`) which was modified to contain the *A. thaliana* nuclear and organellar ITS region.

Profiles of the strongly immunocompromised *jar1 ein2 npr1* (JEN) triple mutant and the *dde2 ein2 pad4 sid2* (DEPS; table S1) quadruple mutant contained a disproportionate abundance of sequences not classified as bacteria, despite our use of bacteria-specific 16S rRNA gene primers (table S2). Because oomycete prevalence and abundance were otherwise rare across samples (Fig. 4.4), we removed these sequences during sequence processing (section 4.9.4.3) in order to focus on the alterations in the respective bacterial communities.

For the synthetic community experiments, sequences were classified 'isolate' (matching one of the isolates added), 'contamination' (matching a plant derived sequence), or 'unmapped' (not mapping anything in the reference set); both contamination and unmapped reads were removed for downstream analysis. The resulting counts, after removing host contamination, are referred to as the *usable* reads/counts/portion of the data, and are the basis for statistical analysis, where the total number of *usable* reads per sample is defined as the sampling depth for that sample.

4.9.4.8 *In silico* pooling of samples in census experiments In the 454 dataset, some DNA samples were barcoded and sequenced on multiple plates in an effort to achieve adequate depth. The resulting OTU counts from barcodes corresponding to the same original DNA sample were pooled (added) *in silico* after processing, but prior to any statistical analysis. Any barcode with 50 or less total reads was discarded, but samples that had between 50 and 1,000 usable reads were matched with samples from the same experiment, fraction and genotype and, when possible, pooled to obtain samples with at least 1,000 reads that were amenable to rarefaction. To allow for direct comparison between the Illumina and 454 datasets, samples that were pooled in the 454 dataset were also pooled in all the Illumina datasets regardless of their depth.

4.9.5 Microbial quantification procedures

4.9.5.1 CARD-FISH We used CARD-FISH (Eickhorst and Tippkötter, 2008) to show that the relative abundance decrease in Actinobacteria in the salicylic acid signaling mutant *pad4* EC samples compared to wildtype Col-0 EC controls was due to a decrease in the absolute number of metabolically active Actinobacteria in *pad4* EC tissue (Fig. 4.16A). On the other hand, the relative abundance increase of Proteobacteria in *cpr5* roots was due to a lower total number of other types of metabolically active Eubacteria (Fig. 4.16B).

We applied a previously described protocol (Lundberg et al., 2012; Eickhorst and Tippkötter, 2008). Briefly, several root systems from bolting plants grown in Mason Farm soil were fixed using 4% formaldehyde in PBS at 4°C for 3 h, washed twice in PBS and stored in 1:1 PBS:molecular-grade ethanol at -20 C. Bulk Mason Farm soil, rhizosphere, and ground EC samples from 3 sets of Col-0, *cpr5*, or *pad4* samples were pooled and harvested as described above. Samples were made equal by mass and probe sonicated for 5 minutes in 30 sec bursts. The sample suspension was diluted 1:500 in water and applied to a 25 mm polycarbonate filter with a pore size of 0.2 mm (Millipore) using a vacuum microfiltration assembly. Filters were embedded in 0.2% low-melting point agarose and dried. Prepared filters were treated with lysozyme solution (1 h at 37°C, 10 mg ml^{-1} ; Fluka) and achromopeptidase (30 min at 37°C, 60 U ml^{-1} ; Sigma) and subsequently washed. Endogenous peroxidases were inactivated with methanol treatment amended by 0.15% H_2O_2 at room temperature for 30 min and washed again. Probes targeting either the 16S or the 23S rRNA of eubacteria (EUB338 (5'-GCTGCCTCCCGTAGGAGT-3',35% formamide), actinobacteria (HGC69a (5'-TATAGTTACCACCGCCGT-3', 25% formamide), proteobacteria (1:1:1, ALF968 (5'-GGTAAGGTTCTGCGCGTT- 3', 20% formamide), (5'-Bet42a (5'-GCCTTCCCACTTCGTTT-3', 35% formamide), and Gam42a (5'-GCCTTCCCACATCGTTT-3', 35% formamide)) and the negative control (NON338 (5'-ACTCCTACGGGAGGCAGC-3', 30% formamide) were defined using probeBase (Loy et al., 2007), labeled with enzyme horseradish peroxidase on the 5' end (Invitrogen), diluted in

hybridization buffer (final concentration of 0.19 ng ml^1) with each probe's optimum formamide concentration, and hybridized at 35°C for 2 h. Unbound probes were washed away from samples in wash buffer (NaCl content adjusted according to the formamide concentration in the hybridization buffer) at 37°C for 30 min. Fluorescently labeled tyramide was used for signal amplification, and samples were washed before mounting on glass slides. For double CARD-FISH, samples went through a second round of the protocol, starting at the peroxidase inhibition with a second variety of fluorescently labeled tyramide used to be able to distinguish the signals from each probe. Filter sections were mounted on glass slides using Vectashield with DAPI (Vector Laboratories, catalogue no. H-1200) for mounting solution, and sealed with nail polish for storage. All microscopy images were made on a Nikon Eclipse E800 epifluorescence microscope.

For quantification of bacteria, positive EUB338 probe signals that co-localized with a DAPI signal were counted as Eubacteria. Positive Actinobacteria or Proteobacteria signals were counted as positive when the HGC69a probe or a combination of the ALF968, Bet42a, and Gam42a probes co-localized with both EUB338 and the DAPI signal. For each filter set, 20 fields were counted.

4.9.5.2 Differential eukaryotic 18S and prokaryotic 16S determination To measure the bacterial density in plant roots we used a protocol that simultaneously amplifies bacterial 16S and plant nuclear 18S, and calculated the ratio between these two groups of sequences across different genotypes. We refer to this method as density PCR (dPCR). Early attempts showed that the 16S:18S ratio was too low (data not shown) so we implemented a linear amplification step prior to exponential PCR. In the first step we performed 50 linear amplification steps with the 338F primer (5'-ACTCCTACGGGAGGCAGCA-3'). This primer amplifies bacterial 16S preferentially over organellar 16S. The linear amplification step was performed with the following reaction:

- 5μL of Kapa Enhancer

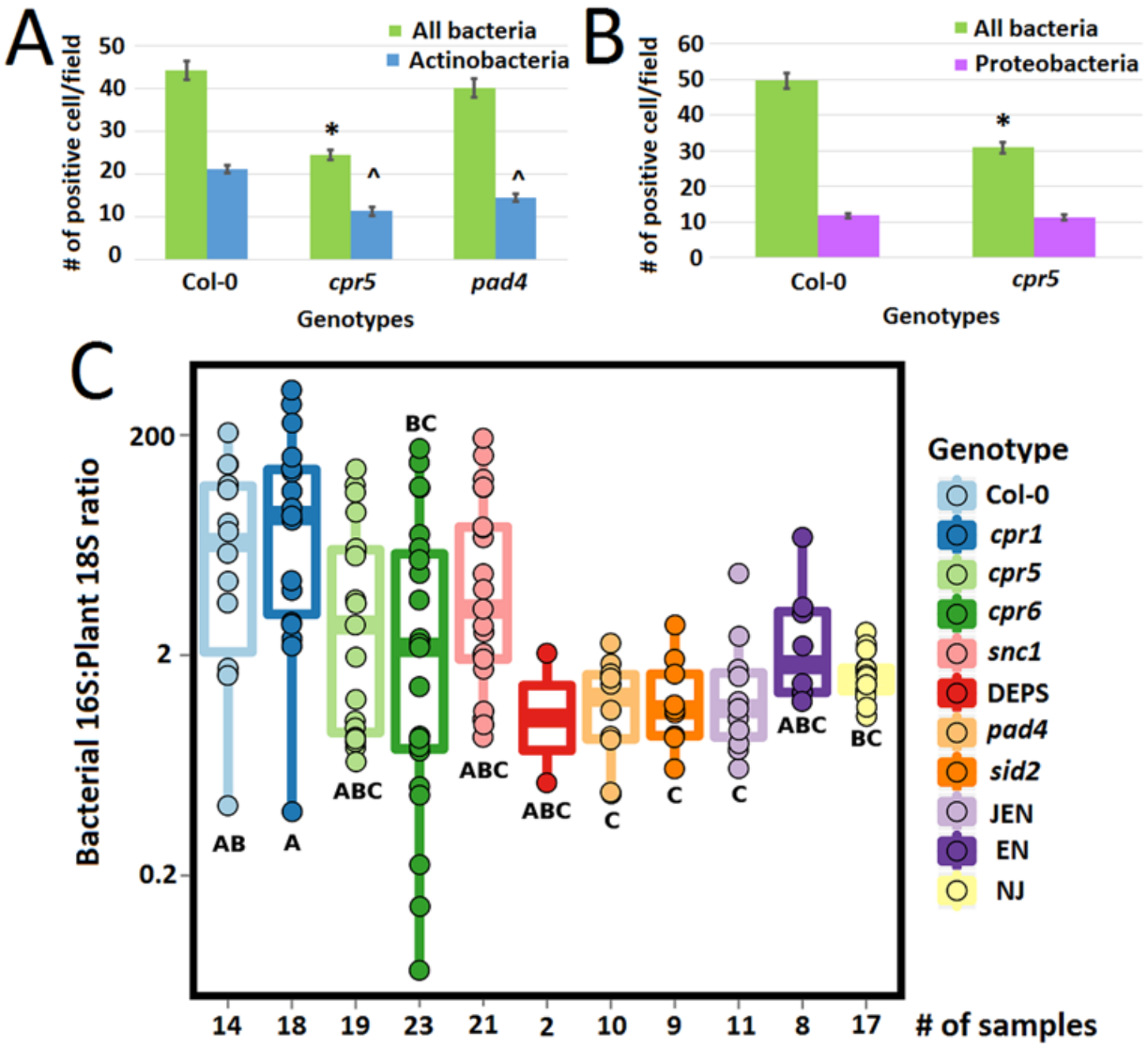


Figure 4.16: **The absolute quantification of bacteria in samples grown in MF soil**. CARD-FISH results from EC samples applied to filters for counts (section 4.9.5.1), and were probed for metabolically active Eubacteria (green) bacteria and Actinobacteria (blue) (**A**) or Proteobacteria (purple) (**B**). 20 fields were counted for each genotype with mean and standard error of the mean (s.e.m.) shown. Asterisk indicates significantly lower than Col-0 (p-value < 0.001). Caret indicates significantly lower than Col-0 Actinobacteria counts (p-value < 0.001). **C** The ratio of bacteria 16S to plant 18S sequences in EC samples (section 4.9.5.2). A, B, and C labels denote results from a Tukey's HSD test. Genotypes that do not share any letters are statistically different.

- 5μL of Kapa Buffer A
- 0.4μL of 5μM 338F
- 0.375μL of mixed PNAs (1:1 mix of 100μM pPNA and 100μM mPNA)
- 0.5μL of Kapa dNTPs
- 0.25μL of Kapa Robust Taq
- 8μL of dH_2O
- 5μL DNA

With the following thermocycling program:

1. 95°C for 45 seconds
2. 50 cycles of:
   (a) 95°C for 15 seconds
   (b) 78°C for 5 seconds (PNA annealing)
   (c) 60°C for 30 seconds (338F annealing)
   (d) 72°for 30 seconds
3. 12°C for 5 minutes
4. 4°C for ever

This linead amplification product is bead cleaned, and followed up by the molecular tagging protocol as described previously (Lundberg et al., 2013) (section 4.9.3.3, but substituting the tagged primer 806R (806R_f1-806R_f6, ST2) for tagged 926R (926R_f1-926R_f4, ST2). Primer 926R is universal (while 806R is bacteria specific) thus allowing to amplify nuclear 18S templates. For the forward primer we used the bc1 modification suggested by Lundberg et al. (2013) (515F_bc1_f1-515_bc1_f6, ST2). The following reaction was used:

- 5μL of Kapa Enhancer
- 5μL of Kapa Buffer A
- 0.4μL of 5μM 515F TAGGED
- 0.375μL of mixed PNAs (1:1 mix of 100μM pPNA and 100μM mPNA)
- 0.5μL Kapa dNTPs
- 0.25μ Kapa Robust Taq
- 13μL DNA (all the elution volume from linear amplification step)

For one cycle with the following thermocycling program:

1. 95°C for 60 seconds
2. 78°C for 5 seconds (PNA annealing)
3. 60°C for 30 seconds (515F annealing)
4. 72°for 60 seconds
5. 4°C for ever

After this one cycle is done, the reaction is removed from the the thermocycler and placed on ice. While on ice, the following was added:

- 0.4μL of 5μM 926R TAGGED
- 1.6μL of dH_2O

And follow it with one cycle with the following program:

1. 95°C for 60 seconds
2. 78°C for 5 seconds (PNA annealing)

3. 50°C for 60 seconds (926R annealing)

4. 72°for 60 seconds

5. 12°C for 5 minutes

6. 4°C for ever

The molecule-tagged product from this reaction is bead cleaned, and the cleaned product is used as input for the exponential PCR witht the following reaction:

- 12.5μL of Kapa HiFi HotStart ReadyMix
- 0.375μL of mixed PNAs (1:1 mix of 100μM pPNA and 100μM mPNA)
- 2.5μL of 5μM index primer (ST2)
- 10μL of DNA (all the elution volume from the molecule tagging step)

With the following program:

1. 95°C for 45 seconds
2. 35 cycles of:
   (a) 95°C for 15 seconds
   (b) 78°C for 5 seconds (PNA annealing)
   (c) 60°C for 30 seconds (index primer annealing)
   (d) 72°for 30 seconds
3. 12°C for 5 minutes
4. 4°C for ever

We chose 192 samples covering all mutants in different experiments on MF soil, and we applied this density PCR (dPCR) protocol. There are only 96 index primers but we used combinations in the frameshift length of the molecule tagging to multiplex all 192 samples in one sequencing run, while keeping the average size the same. This is achieved by using the following combinations of primers for each plate during the molecule-tagging steps:

- Plate1
  - 515F_bc1_f1
  - 515F_bc1_f3
  - 515F_bc1_f5
  - 926R_f2
  - 926R_f4
- Plate2
  - 515F_bc1_f2
  - 515F_bc1_f4
  - 515F_bc1_f6
  - 926R_f1
  - 926R_f3

All primer sequences are available in table S2.

After applying the dPCR protocol to these samples, we ran each reaction on an agarose gel to confirm the presence of two bands of the right sizes (one for the 16S and a larger one for the 18S). Then we pooled 3μL of each reaction into a master mix and bead cleaned twice eluting in 200 uL. This library mix was run on an agarose gel to confirm the presence of two bands of the right size and the absence of primer dimer. This library master mix was quantified with pico green (Quant-IT) and loaded into an Illumina MiSeq instrument

(following the manufacturers protocol) using a 50-cycle V2 chemistry kit. Resulting sequences were demultiplexed and quality controlled with Sickle (Joshi and Fass, 2011) by removing any sequence that had at least one base with a Q-score < 30. The remaining sequences were matched to a reference set that included the *Arabidopsis thaliana* nuclear 18S rRNA gene, *Arabidopsis thaliana* organellar 16S rRNA gene and the 17 most abundant bacterial sequences in the Greengenes database. No mismatches were allowed during this phase. After mapping the sequences, a ratio of bacterial 16S to plant 18S was calculated (Bactratio) and the results were analyzed with ANOVA and a post-hoc Tukey test using the aov and tukeyHSD functions in R. Results are presented in Fig. 4.16C.

4.9.6 Synthetic community (SynCom) experimental procedures

4.9.6.1 Microbe isolation To isolate putative endophytic bacteria from root systems, samples were harvested as described in section 4.9.2.3, rinsed in several water washes and debris was removed with sterile tweezers. Cleaned roots were then surface sterilized with freshly made 10% household bleach with 0.1% Triton-X100 for 12 minutes. Following the bleaching, roots were rinsed once in sterile distilled water, then placed in 2.5% sodium thiosulfate to neutralize the bleach for 2 minutes, and rinsed once more with sterile distilled water. Small pea-sized chunks of resulting surface-sterilized roots were then pulverized fresh in an autoclaved 2 mL tube with 3 glass beads with 300 μL of PBS, using the MPBio FastPrep 24 for 20s at 4.0 m/s. 300 μL of 80% glycerol was then added to the crushed material for a final glycerol concentration of 40%. Tubes were then flash frozen and stored at -80°C. To isolate microbes, root material was diluted 1:100-1:1000 in sterile water and plated on a diverse set of low nutrient solid media plates including: 1/10 LB, 1/50 TSA, KB, 1/10 869, LB with 1% Humic acid, R2A, *Pseudomonas* Media, TSA with polymyxcin B. We also utilized media with sterile filtered MF soil as the nutrient source, and homogenized sterile roots as the carbon source of another media.

4.9.6.2 Synthetic community experimental setup *A. thaliana* seeds were surface-sterilized and germinated the same way as we did for the wild soil experiment (section 4.9.2.2).

Seedlings on MS plates were transferred to 2.5 inch square plastic pots (Kord Products Ltd.) containing (∼100 mL) sterilized (autoclaved) calcined clay (Diamond Pro Calcined Clay Drying Agent, (`http://www.diamondpro.com/Products/CalcinedClayDryingAgent`) pots supplemented with 40% volume (∼40 mL) of 1/4 MS (no carbon source), and inoculated with a synthetic community was composed of a mix of 38 bacterial strains (table S9).

Plants were harvested in the same way as the Mason Farm soil plants, except that no rhizosphere samples were produced (section 4.9.2.3. DNA from both bulk soil and EC samples was extracted using the MoBio PowerSoil kit. We utilized a recently published improvement of Illumina library preparation, which takes advantage of molecular tags (MT) to allow direct counting of original DNA templates in the sample, thus reducing PCR and sequencing errors and biases, as well as peptide nucleic acid to block amplification of host DNA (Lundberg et al., 2013). We sequenced the V4 region of the bacterial input (inoculum) as well as EC and bulk soil samples, with primers 515F and 806R (table S2) from three independent biological replicates.

4.9.6.3 Growth curves Growth curves were performed in 1/10 LB with 0.1 M phosphate buffer containing 0.01% yeast extract (Silva et al., 2007) and either 0, 0.125, 0.25, or 0.5 mM salicylic acid added. Grown at 28°C shaking at 150 rpm. Optical density at 600 nm was measured every 2 hours for 50 hours of growth using a Synergy 2 multi-detection microplate reader (BioTek). Supernatants were harvested from liquid cultures of #273 or #303 grown in 1/2 and 1/10 LB with either 0, 0.125, 0.25, or 0.5 mM salicylic acid added after 0, 24, or 48 hours of growth and total salicylic acid was measured as described in section 4.9.1.1. No loss of total salicylic acid signal was detected for either culture in any media conditions (data not shown).

For #303 growth on agar plate, minimal salts media ($(NH_4)_2SO_4$ 2g, K_2HPO_4 14g, KH_2PO_4 6g, sodium citrate 1g, $MgSO_4$ 0.2g per L) was supplemented with 0.5 mM salicylic acid in phosphate buffer, or phosphate buffer alone. #303 colonies were evident after 4 days

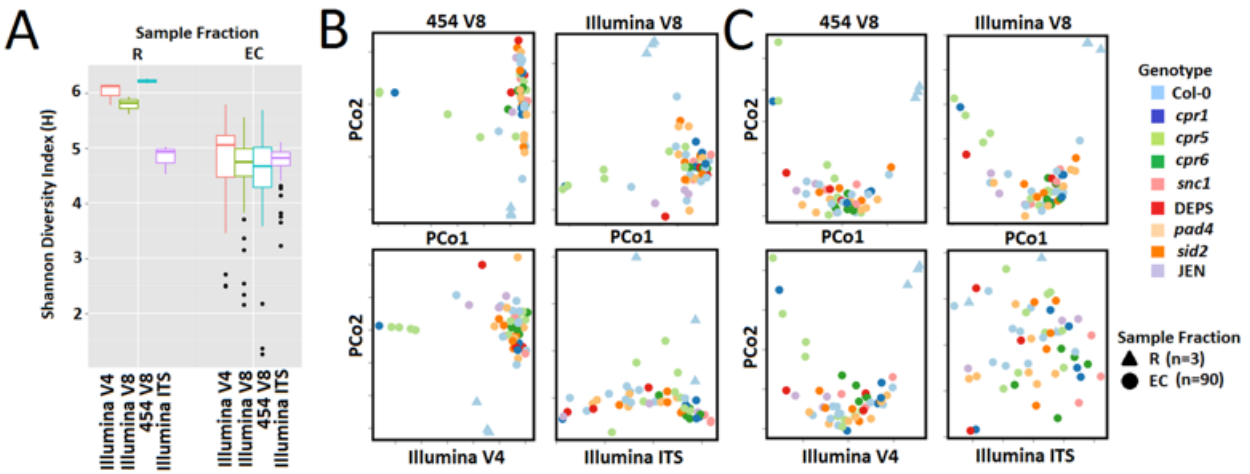


Figure 4.17: **Alpha and beta diversity for different 16S rRNA and ITS regions**. **A** Shannon diversity index (H) for Illumina V4, Illumina V8, 454 V8, and Illumina ITS in both R and EC samples. Principal Coordinate Analysis of weighted UniFrac (**B**) and unweighted UniFrac (**C**) R (triangles) and EC (circles) samples sequenced by Illumina V4, llumina V8, 454 V4, and Illumina ITS demonstrates that bacterial profiles differ between R and EC samples regardless of sequencing platform and variable region. In contrast ITS profiles are remarkably similar (both in alpha and beta diversity) between R and EC samples.

of growth on this media and after 2 days of growth on LB (Fig. 4.13F).

4.9.7 Statistical analysis

4.9.7.1 Diversity analysis for census experiments Alpha and beta diversity were calculated on count tables that were rarefied to 1,000 reads. Samples with less than this number of usable reads after pooling (section 4.9.4.8) were discarded. Alpha diversity (Shannon index, richness, Simpson index) metrics were calculated using vegan (Oksanen et al., 2014), and differences between groups were tested with ANOVA. Beta diversity metrics were calculated with QIIME (UniFrac) or vegan (Bray-Curtis), and Principal Coordinate Analysis (PCoA) was performed with labdsv (Roberts, 2016).

4.9.7.2 ZINB family and OTU-level analysis for census experiments A ZINB model (Zuur et al., 2009) acknowledges that some proportionof the observed zeroes in the count tables might not be biologically meaningful, but rather experimental error (Fig. 4.18, upper branch) and was therefore appropriate to use on our sparse family tables. At the same time, a ZINB model can focus on the variability associated with the variables of interest (Fig.

4.18, lower branch). A ZINB model achieves its purpose by combining a classic count GLM with a 'bad zero' generating process, and it links the two processes via a single parameter (p) that indicates the proportion (i.e. the probability) that a given zero is a 'bad zero' (Fig. 4.18; eq. 4.1):

$$f(y) = \begin{cases} \pi + (1-\pi) f_{nb}(y) & y = 0 \\ (1-\pi) f_{nb}(y) & y > 0 \end{cases} \tag{4.1}$$

Where π is the probability of a 'bad zero', f_{nb} is the negative binomial probability density function, and y is the observed counts of a given taxon in a given sample.

Like other linear modeling approaches, the ZINB model allows one to model a set of observations with a combination of variables. Besides the biological variables that interest usthe most (fraction, genotype and the interaction between the two), we included batch variables to control for technical error. We used two batch variables: experiment, which includes plant/harvest date, growth chamber, DNA extraction and soil dig; and plate which corresponds to library preparation and sequencing plate batches. The full set of variable is in the legend of table S4, and the sample metadata and design matrices for the model are in supplementary dataset SD2.

The implemented ZINB model depends on three parameters: i) π is the probability of a 'bad zero', ii) α is an over-dispersion parameter that quantifies the deviation of the count process from the standard Poisson assumption of equality between mean and variance, and iii) a vector of coefficients β that quantifies the association of counts with each variable of interest. Each of these parameters has to be estimated in a full ZINB model, but π and α can be fixed to a set value by making extra assumptions and simplifying the model (Fig. 4.18C). It is impossible to say *a priori* whether the extra assumptions made by simple models are justified, so we fit each of the four models from Fig. 4.18C on each family or OTU for each dataset, and then compared the model fits by means of the Akaike Information Criterion (AIC), which is a measure that combines the quality of the model fit while penalizing more

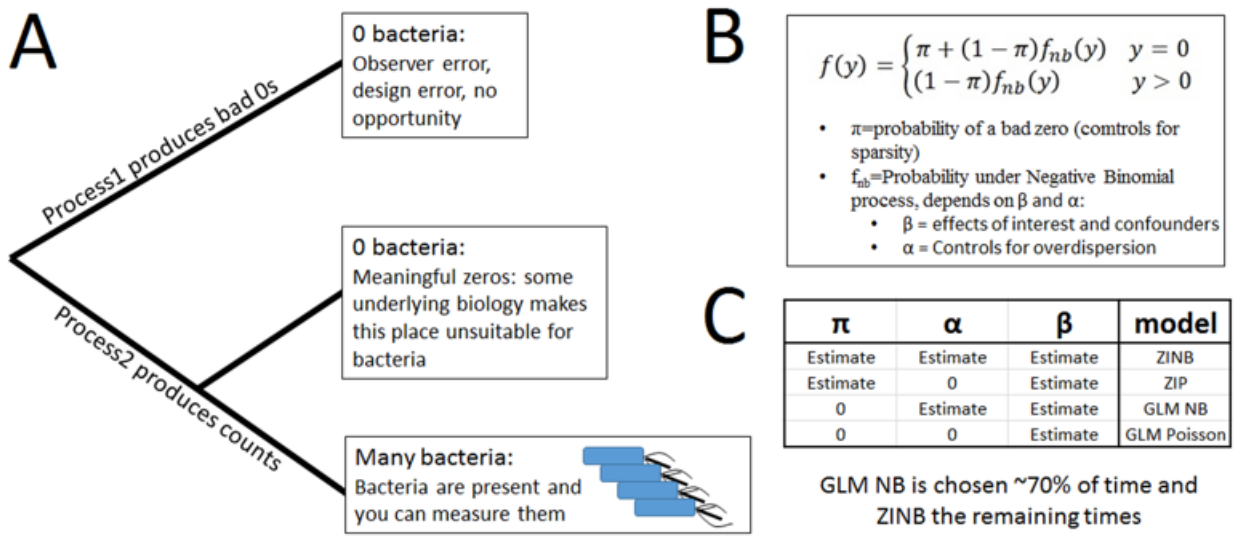


Figure 4.18: **Zero-Inflated Negative Binomial model**. The rationale (**A**) and formula (**B**) for the ZINB model is shown. (**C**) The models which were tested with this data set.

complex models, so that extra parameters are only includedwhen justified (Akaike, 1974).

Each of the four models was fit with the same design matrix plus the natural logarithm of the number of usable reads per sample (i.e. depth) (see legend sheet on table S4 for details on the variables, and supplementary dataset SD2 for the design matrices), and the best model was chosen for each family on each dataset based on the AIC. The design matrix was constructed in a way that the genotype coefficients represent the difference with respect to Col-0 wildtype, and the fraction coefficients represent the difference with respect to bulk soil samples. The resulting coefficents (β) were tested for significance with z-tests and corrected for multiple testing with the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Model fits were performed with the stats (R Core Team, 2014), MASS (Venables and Ripley, 2002) and pscl (Zeileis et al., 2008; Jackman, 2015) packages in R.

4.9.7.3 Comparison of enrichment profiles between genotypes

The ZINB model allowed us to identify the bacterial families and OTUs that are enriched or depleted in the EC of specific plant genotypes with respect to Col-0. In order to compare the enrichment/depletion profiles between genotypes, we developped a Monte Carlo test based on the Manhattan

distance between enrichment/depletion profiles or pairs of genotypes. First, each genotype is given a profile, which is a vector of numbers defined as following: each enriched family gets a value of 1, depleted families get a value of -1 and families that are not significantly different from Col-0 are given a value of 0. In this manner, each genotype gets an ordered vector of numbers, and such a vector can be compared directly to vectors of other genotypes. We chose the Manhattan distance, because given our definition of enrichment/depletion profiles, only families that are different contribute to the distance metric, and families that have opposite effects between two genotypes (i.e. families enriched in one genotype and depleted in another) contribute more than families where the difference is between effect and no effect. Notably, DEPS EC samples had 52 DA families, nearly all of which were depletions (Figs. 4.1C and 4.5). The decrease in alpha-Diversity observed in DEPS EC samples (Fig. 4.1B) likely reflects these depletions. This large number of depletions and low diversity in DEPS roots cannot be explained by their oomycete burden, since the equally oomycete-laden JEN EC samples exhibited only four DA families (Fig. 4.5A), and only one of these was shared with DEPS. Finally, to test whether the observed distances between genotypes are significant, we used a Monte Carlo procedure, by randomly permuting the order of the enrichment/depletion profiles 1,000 times and re-calculating the Manhattan distance in each instance. This approach provides an empirical null hypothesis that can be compared to the value observed on the original data, and an empirical p-value can be calculated as the proportion of cases in the simulation that have distance values at least as extreme as the distance from the real data. The table of p-values is provided in Fig. 4.5 for the family level analysis, and Fig. 4.6B for the OTU level analysis.

4.9.7.4 PCA and CAP analysis of synthetic community experiments For synthetic community data, the count table was rarefied to 400 consensus, and Principal Component Analysis was performed with the 'princomp' function of R. Canonical Analysis of Principal Coordinates (CAP) (Anderson and Willis, 2003) was performed using the 'capscale'

function of the vegan package (Oksanen et al., 2014) in R. CAP was performed on the full table of both the survey and the SynCom data and the constrained variation of fraction (Fig. 4.10B) and salicylic acid (Fig. 4.12A) was obtained after conditioning forevery other technical and covariate. The proportion of variance explained by each variable (table S5), was estimated as the proportion of the total variation explained by the constrained axis of CAP, and confidence intervals were obtained by bootstrapping the taxa of the count tables for 1,000 pseudoreplicates. For all of the CAP analysis, the CY Index, sometimes referred as Cao Index (Cao et al., 1997) was used as implemented in the 'vegdist' function of the vegan package.

4.9.7.5 Defining robust colonizers in synthetic community experiments We observed that some isolates were normally present in the vast majority of the SynCom EC samples, while others were rarely present. The presence/absence pattern in the root was not fully explained by the abundance in the soil or inoculum We defined robust colonizers as those isolates that have probability of being present in a given EC sample, that is significantly higher than 50% (q-value < 0.05, one-tailed binomial test, Benjamini-Hochberg correction). Presence was defined as the existence of one consensus sequence matching the given isolate, but almost identical results were obtained to raising this threshold to 5 consensus (data not shown). Only wild-type Col-0 root samples were used for this analysis, and so the list of robust colonizers represent bacteria that have a high chance of colonizing a wildtype plant. Isolates that fail to reject the null hypothesis in this test are dubbed sporadic or non-colonizers.

4.9.7.6 ZINB analysis of synthetic community experiments For the synthetic community experiments, we repeated the ZINB analysis performed on the census datasets (section 4.9.7.2, but at the isolate level since we chose the isolates to have easily differentiable 16S rRNA gene sequences on the basis of Sanger sequencing of their 16S rRNA gene. The same four model structures were used, and AIC was used to decide on the best model.

Hypothesis testing and multiple testing correction were done in the same manner as described in section 4.9.7.2. The same software was utilized. A different design matrix (corresponding to the experimental design differences) was used, and the variables included are described in the legend sheet of table S4.

4.9.7.7 Genomic analysis of isolates in synthetic community experiments Experimentally verified pathways that involve salicylic acid (SA, salicylate) were first obtained from MetaCyc (`http://www.metacyc.org/`). Five pathways were identified for SA degradation: salicylate degradation I, salicylate degradation II, salicylate degradation III, salicylate degradation IV, and enzyme salicylate 1,2-dioxygenase (accession number G-12243 MetaCyc). Of the two salicylic acid biosynthesis pathways, only one has evidence in bacteria (salicylate biosynthesis I) and so it was the only one used in our analyses. The amino acid sequence of all the characterized genes in this reaction were retrieved from the databases linked by MetaCyc (table S10c) and were used to perform a BLAST searches against the predicted ORFs of the isolates' genomes. BLAST searches were performed on the IMG/ER webserver with default parameters. The results of the best hit (identity percent, and query coverage) are given in table S10a-b. Yellow color in table S10 indicates a good homolog hit while green indicates matching annotations between query and subject regardless of the hit quality.

CHAPTER 5

Direct integration of phosphate starvation and immunity in response to a root microbiome[1]

Plants live in biogeochemically diverse soils that harbor extraordinarily diverse microbiota. Plant organs associate intimately with a subset of these microbes; this community's structure can be altered by soil nutrient content. Plant-associated microbes can compete with the plant and with each other for nutrients; they can also provide traits that increase plant productivity. It is unknown how the plant immune system coordinates microbial recognition with nutritional cues during microbiome assembly. We establish that a genetic network controlling phosphate stress response influences root microbiome community structure, even under non-stress phosphate conditions. We define a molecular mechanism regulating coordination between nutrition and defense in the presence of a synthetic bacterial community. We demonstrate that the master transcriptional regulators of phosphate stress response in Arabidopsis also directly repress defense, consistent with plant prioritization of nutritional stress over defense. Our work will impact efforts to define and deploy useful microbes to enhance plant performance.

Plant organs create distinct physical and chemical environments that are colonized by specific microbial taxa1. These can be modulated by the plant immune system (Lebeis et al., 2015) and by soil nutrient composition (Hacquard et al., 2015). Phosphorus is present in the biosphere at high concentrations, but plants can only absorb orthophosphate (Pi), a

[1]Most of the content of this chaper has been published as a peer-reviewed article (Castrillo et al., 2017). The text has been been ligthly edited and re-arranged to facilitate reading. The figure order has been changed to match the updated text order. Section and subsection headers have been added for easier navigation. Numerous supplementary files were made available online at the time of publication, and are not included here; they will be referred to as Supplementary Table or Supplementary Dataset and can be obtained at (`http://www.nature.com/nature/journal/vaop/ncurrent/full/nature21417.html`).

form also essential for microbial proliferation (Richardson and Simpson, 2011; Zhu et al., 2016) and scarce in soil (Raghothama, 1999). Thus, plants possess adaptive phosphate starvation responses (PSR) to manage low Pi availability that typically occurs in the presence of plant-associated microbes. Common strategies for increasing Pi uptake capacity include rapid extension of lateral roots foraging into topsoil where Pi accumulates7 and establishment of beneficial relationships with some soil microorganisms (Harrison, 2012; Hiruma et al., 2016). For example, the capacity of a specific mutualistic fungus to colonize Arabidopsis roots is modulated by plant phosphate status implying coordination between the PSR and the immune system (Hiruma et al., 2016; Hacquard et al., 2016). Descriptions of pathogen exploitation of PSR-immune system coordination are emerging (Zhao et al., 2013; Lu et al., 2014).

We demonstrate that Arabidopsis mutants with altered phosphate starvation responses (PSR) assemble atypical microbiomes, either in phosphate-replete wild soil, or during in vitro colonization with a synthetic bacterial community (SynCom). This SynCom competes for phosphate with the plant and induces PSR in limiting phosphate. PSR in these conditions requires the master transcriptional regulator PHR1 and its weakly redundant paralog, PHL1. The severely reduced PSR observed in *phr1 phl1* mutants is accompanied by transcriptional changes in plant defense leading to enhanced immune function. Negative regulation of immune system components by PHR1 is direct, as measured by target gene promoter occupancy, and functional, as validated by pathology phenotypes. Thus, PHR1 directly activates microbiome-enhanced response to phosphate limitation while repressing microbially-driven plant immune system outputs.

5.1 The root microbiome in plants with altered phosphate stress response

We linked PSR to the root microbiome by contrasting the root bacterial community of wild-type Arabidopsis Col-0 with three types of PSR mutants (Figs. 5.2a-b and 5.6; section 5.6.1; Supplementary Table 1). PSR, historically defined in axenic seedlings and measured by Pi concentration in the plant shoot, is variable across these mutants. In replete Pi and axenic

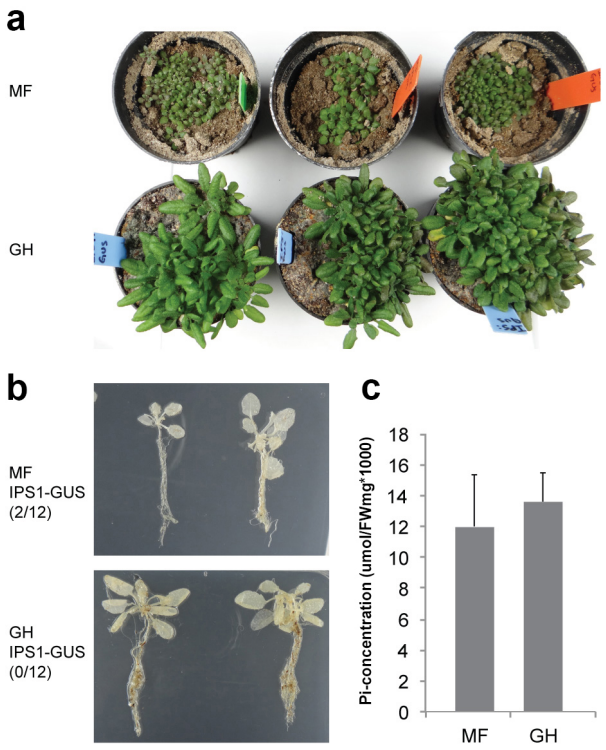


Figure 5.1: **Plants grown in Mason Farm wild soil or phosphate (Pi) replete potting soil do not induce PSR and accumulate the same amount of Pi**. **a**, Plants overexpressing the PSR reporter construct IPS1:GUS grown in Mason Farm wild soil (MF) or in phosphate (Pi) replete potting soil (GH) (250 ppm of 20-20-20 Peters Professional Fertilizer). **b**, Expression analysis of the reporter constructs IPS1:GUS (n= 12) shows lack of induction of PSR for both soils analyzed. In this construct, the promoter region of IPS1, highly induced by low Pi, drives the expression of GUS. Plants were grown in the conditions described in a. The number of GUS positive plants relative to the total number of plants analyzed in each condition is shown in parenthesis. **c**, Phosphate (Pi) concentration in shoots (n= 6) of plants grown in both soils analyzed shows no differences. Plants were grown in a growth chamber in a 15-h light/9-h dark regime (21°C day /18°C night). Images shown here are representative of the 12 plants analyzed in each case. Bars mean standard deviation.

conditions, *phr1* plants accumulate less free Pi than wild-type (Bustos et al., 2010); *pht1;1*, *pht1;1 pht1;4* and *phf1* accumulate very low Pi levels and express constitutive PSR (Shin et al., 2004; González et al., 2005); and *pho2*, *nla* and *spx1 spx2* express diverse magnitudes of Pi hyper-accumulation (Huang et al., 2013; Lin et al., 2013; Puga et al., 2014). We grew plants in a previously characterized wild soil (Lundberg et al., 2012) that is not overtly phosphate deficient (Fig. 5.1). Generally, the Pi concentration of PSR mutants grown in this wild soil recapitulated those defined in axenic conditions, except for *phf1* and *nla* which displayed the opposite phenotype to that observed in axenic agar, and *phr1* which accumulated the same Pi concentration as Col-0 (Fig. 5.2b). These results suggest that complex chemical conditions, soil microbes, or a combination of these can alter Pi metabolism in these mutants.

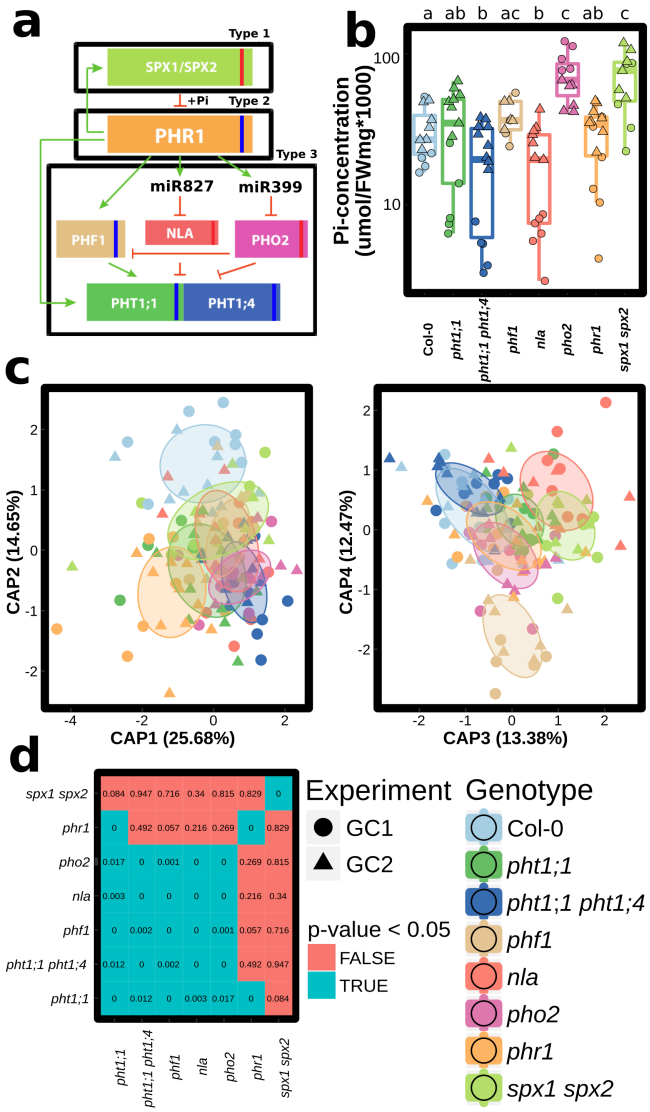


Figure 5.2: **Phosphate Stress Response (PSR) mutants assemble an altered root microbiota**. **a**, Diagram of PSR regulation in Arabidopsis. Red and blue stripes indicate whether these mutants hyper- or hypo-accumulate Pi, respectively, in axenic, Pi replete conditions. The master PSR regulator PHR1 is a Myb-CC family transcription factor (Bustos et al., 2010) bound under phosphate replete conditions by the negative regulators SPX1 and SPX2 in the nucleus (Puga et al., 2014). During PSR, PHR1 is released from SPX and regulates genes whose products include high-affinity phosphate transporters (PHT1;1 and PHT1;4)(Bustos et al., 2010). Transporter accumulation at the plasma membrane is

controlled by PHF1 (Lin et al., 2013), while PHO2 and NLA mediate PHT1 degradation (Huang et al., 2013; Lin et al., 2013). **b**, Phosphate (Pi) concentration in shoots of plant genotypes (grown in growth chambers, 16-h dark/8-h light regime, 21°C day 18°C night for 7 weeks) in a natural soil. Statistical significance was determined by ANOVA while controlling for experiment (indicated by point shape); genotype grouping is based on a post-hoc Tukey test, and is indicated by letters at the top; genotypes with the same letter are indistinguishable at 95% confidence. Biological replicate numbers are: Col-0 (n=12), *pht1:1* (n=13), *pht1;1 pht1;4* (n=14), *phf1* (n=9), *nla* (n=13), *pho2* (n=11), *phr1* (n=14) and *spx1 spx2* (n=11) distributed across two independent experiments. **c**, Constrained ordination of root microbiome composition showing the effect of plant genotype: *phr1* separates on the first two axes, *spx1 spx2* on the third axis and *phf1* on the fourth axis. Ellipses show the parametric smallest area around the mean that contains 50% of the probability mass for each genotype. Biological replicate numbers are: Col-0 (n=17), *pht1:1* (n=18), *pht1;1 pht1;4* (n=17), *phf1* (n=13), *nla* (n=16), *pho2* (n=16), *phr1* (n=18) and *spx1 spx2* (n=14) distributed across two independent experiments. **d**, Table of p-values from Monte Carlo pairwise comparisons between mutants at the OTU level. A significant p-value (cyan) indicates that two genotypes are more similar than expected by chance.

Bacterial root endophytic (EC) community profiles were consistent with previous studies (Lundberg et al., 2012; Lebeis et al., 2015). Constrained ordination revealed significant differences between bacterial communities across the Pi accumulation gradient represented by these PSR mutants [5.3 % constrained variance, canonical analysis of principal coordinates (CAP)] (Fig. 5.2c). Additionally, CAP confirmed that *phr1* and *spx1 spx2* carried different communities, as evidenced by their separation on the first three ordination axes, and that *phf1* was the most affected of Pi-transport mutant (Fig. 5.2c). Specific bacterial taxa had differential abundances inside the roots of mutant plants compared to wild-type. Mutants from the same PSR type had a similar effect on the root microbiome at a low taxonomic

level [97% identity Operational Taxonomic Unit (OTU)] (Fig. 5.2d), while they had no overlapping effect at a higher taxonomic level (Family, Fig 5.6g). This suggests that closely related groups of bacteria have differential colonization patterns on the same host genotypes. Importantly, we found that the enrichment and depletion profiles were better explained by PSR mutant signaling type rather than the mutants capacity for Pi accumulation: all of the Pi-transport-related mutants had a similar effect on the root microbiome, and the antagonistic PSR regulators *phr1* and *spx1 spx2* each exhibited unique patterns (Figs. 5.2a and d and 5.6f-g). Our results indicate that PSR components influence root microbiome composition in plants grown in a phosphate-replete wild soil, leading to alteration of the abundance of specific microbes across diverse levels of Pi accumulation representing diverse magnitudes of PSR.

5.2 Phosphate starvation response in a microcosm reconstitution

Our observations in a wild soil suggested complex interplay between PSR and the presence of a microbial community. Thus, we deployed a tractable but complex bacterial synthetic community (SynCom) of 35 taxonomically diverse, genome-sequenced bacteria isolated from the roots of Brassicaceae (nearly all from Arabidopsis) and two wild soils. This SynCom approximates the phylum level distribution observed in wild-type root endophytic compartments (Extended Data Fig. 3, Supplementary Table 1, Supplementary Table 2). We inoculated seedlings of Col-0, phf1 and the double mutant phr1 phl1 (a redundant paralogue of phr113) grown on agar plates in low or high Pi (Supplementary Text 2). Twelve days later, we noted that the SynCom had a negative effect on shoot Pi accumulation of Col-0 plants grown on low Pi, but not on plants grown on replete phosphate (Fig. 2a). As expected, both PSR mutants accumulated less Pi than Col-0; the SynCom did not rescue this defect. Thus, in this microcosm, plant-associated microbes drive a context-dependent competition with the plant for Pi.

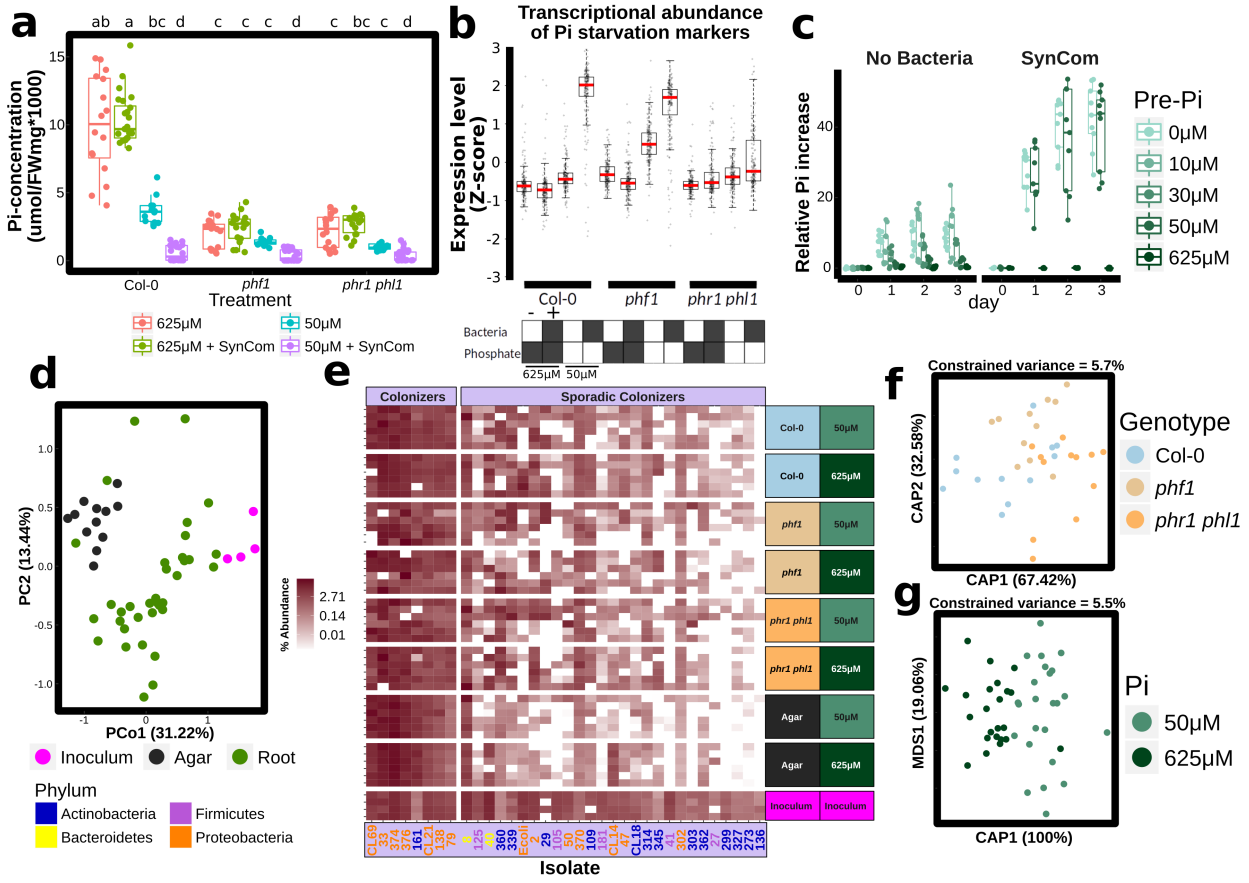


Figure 5.3: **A bacterial Synthetic Community (SynCom) differentially colonizes PSR mutants**. **a**, Pi concentration in shoots of plants grown on different Pi regimens with or without the SynCom. Biological replicates numbers are: Col-0 (n=16 (625 μM Pi), 24 (625 μM Pi + SynCom), 12 (50 μM Pi), 24 (50 μM Pi + SynCom)), *phf1* (n=16, 18, 12, 24) and *phr1 phl1* (n=16, 18, 12, 24) from three independent experiments. Statistical significance was determined via ANOVA while controlling for experiment; letters indicate the results of a post-hoc Tukey test. **b**, Expression levels of 193 core PSR genes. The RPKM expression values of these genes were z-score transformed and used to generate box and whiskers plots that show the distribution of the expression values of this gene set. Boxes at bottom indicate presence/absence of SynCom and Pi at the concentration indicated. This labeling is maintained throughout. Data is the average of 4 biological replicates. **c**, Functional activation of PSR by the SynCom. Plants were grown on five different Pi levels (0 μM Pi,

10 μM Pi, 30 μM Pi, 50 μM Pi and 625 μM Pi) without the SynCom (left) and on three different Pi levels (0 μM Pi, 50 μM Pi and 625 μM Pi) with the SynCom (right). Plants were then transferred to full (1 mM Pi) condition to evaluate the capacity of the plants for Pi accumulation over time (section 5.7.3). Shoots were harvested every 24 h for 3 days and Pi concentration was measured. Pi increase was calculated with equation 5.1. Shoots with SynCom-activated PSR accumulated approximately 20-40 times more Pi than non-inoculated shoots. Absolute Pi concentration values are available in Supplementary Table 4. For all Pi concentrations and SynCom treatments n=6 at day 0, and n=9 at all other time points, distributed across two independent experiments. **d**, PCoA of SynCom experiments showing that Agar and Root samples are different from starting inoculum. Biological replicate numbers are: Inoculum (n=4), Agar (n=12) and Root (n=35) across two independent experiments. **e**, Heatmap showing percent abundances of SynCom isolates (columns) in all samples (rows). Strain name colors correspond to Phylum (bottom left). Within each block, samples are sorted by experiment. For each combination of genotype and Pi level, there are n=6 biological replicates evenly distributed across two independent experiments, except for Inoculum for which there are n=4 technical replicates evenly distributed across two independent experiments. **f**, Constrained ordination showing the effect of plant genotype and **g**, media Pi concentration effect on the root communities. The proportion of total variance explained (constrained) by each variable is indicated on top of each plot; for g, remaining unconstrained ordination was subjected to multi-dimensional scaling (MDS); the first MDS axis (MDS1) is shown. For f and g, biological replicate numbers are: Col-0 (n=12), *phf1* (n=11), *phr1 phl1* (n=12), 50 μM Pi (n=24) and 625 μM Pi (n=23) distributed across two independent experiments.

We sought to establish whether PSR was activated by the SynCom. We generated a literature-based core set of 193 PSR transcriptional markers and explored their expression in transcriptomic experiments (Fig. 5.13a-b; Supplementary Table 3). In axenic low Pi

conditions, only the constitutive Pi-stressed mutant *phf1* exhibited induction of these PSR markers. By contrast, Col-0 plants expressed only a marginal induction of PSR markers compared to those plants grown at high Pi (Fig. 5.3b). This is explained by the purposeful absence of sucrose, a key component for the PSR induction in vitro (Karthikeyan et al., 2007) (Supplementary Text 2; Fig. 5.7) that cannot be used in combination with bacterial SynCom colonization protocols. Remarkably, the SynCom greatly enhanced the canonical transcriptional response to Pi starvation in Col-0 (Fig. 5.3b); this was dependent on PHR1 and PHL1 (Figs. 5.3b and 5.13b). Various controls validated these conclusions (Figs. 5.13, 5.7 and 5.8; section 5.6.2). Importantly, shoots of plants pre-colonized with SynCom on 0 or 50 μM Pi, but not on 650 μM Pi, accumulated 20-40 times more Pi than shoots from similarly treated non-colonized plants when subsequently transferred to full Pi conditions in the absence of additional bacteria (Fig. 5.3c and Supplementary Table 4). This demonstrates functional PSR activation by the SynCom. We thus propose that the transcriptional response to low Pi induced by our SynCom reflects an integral microbial element of normal PSR in complex biotic environments.

We evaluated agar- and root-associated microbiomes of plants grown with the SynCom (section 5.6.3; Figs. 5.3d-e and 5.11e-f; Supplementary Table 5). In line with results from plants grown in wild soil, we found that PSR mutants failed to assemble a wild-type SynCom microbiome (Fig. 5.3f). Some strains were differentially abundant across PSR mutants *phf1* and *phr1 phl1* (Figs. 5.3e-f and 5.11c), Pi concentration (Figs. 5.3g and 5.11d), or sample fraction (Fig. 5.11b, e and f). These results established a microcosm reconstitution system to study plant PSR under chronic competition with plant-associated microbes and allowed us to confirm that the tested PSR mutants influence root microbiome membership.

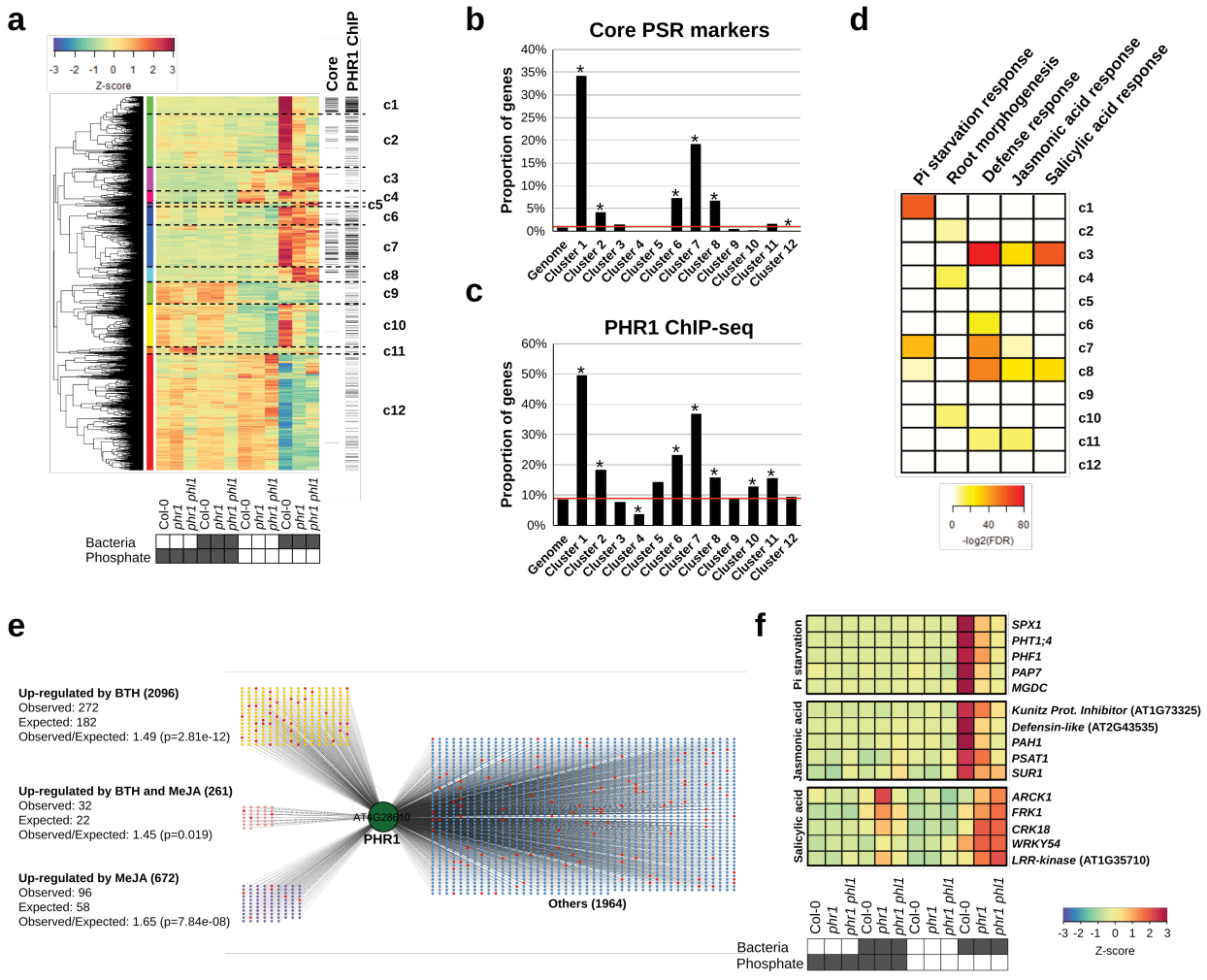


Figure 5.4: **PHR1 mediates interaction of the PSR and plant immune system outputs**. **a**, Hierarchical clustering of 3257 genes that were differentially expressed in the RNA-seq experiment. Plants were germinated on Johnson medium containing 0.5% sucrose supplemented with 1 mM Pi for 7 d, then transferred to 50 μM Pi or 625 μM Pi media (without sucrose) alone or with the Synthetic Community at 10^5 c.f.u/mL, for another 12 d (plates vertical). Columns on the right indicate genes that are core PSR markers (*core* lane) or had a PHR1 binding peak ('PHR1 ChIP' lane). **b**, Proportion of PSR marker genes per cluster. **c**, Proportion of PHR1 direct targets genes per cluster. The red line in b and c denotes the proportion of genes in the whole Arabidopsis genome that contain the analyzed feature. Asterisk denotes significant enrichment or depletion (p-value ≤ 0.05;

hypergeometric test). *d*, Summary of the Gene Ontology enrichment analysis for each of the twelve clusters. The enrichment significance is shown as -$\log_2$(FDR). White means no enrichment. The complete results are in Supplementary Table 9. **e**, The set of genes bound by PHR1 (At4g28610) in ChIP-seq experiments is enriched in genes that are up-regulated by BTH/SA and/or MeJA. Red nodes are core PSR marker genes. **f**, Example of genes bound by PHR1 and differentially expressed in our experiment. PSR marker genes (top) and JA response (middle) are more expressed in wild-type plants, whereas SA-responsive genes (bottom) exhibit higher transcript levels in *phr1* and *phr1 phl1*. The heatmaps show the average measurement of ten biological replicates for Col-0 and *phr1* and six for *phr1 phl1*. The color key (blue to red) related to a, and f, represents gene expression as Z-scores.

5.3 Coordination between phosphate stress response and immune system output

We noted that *phr1 phl1* and *phf1* differentially activated transcriptional PSR in the presence of our SynCom (Fig. 5.3b). Therefore, we investigated the transcriptomes of plants growing in the SynCom to understand how these microbes activate PHR1-dependent PSR. We identified differentially expressed genes (DEGs) that responded to either low Pi, presence of the SynCom, or the combination of both (hereafter PSR-SynCom DEGs) (section 5.6.4; Fig. 5.9a-b; Supplementary Table 6). Hierarchical clustering (Fig. 5.4a, Supplementary Table 7) revealed gene sets (c1, c2, c7 and c10) that were more strongly activated in Col-0 than in *phr1* or *phr1 phl1*. These clusters contained most of the core PSR markers regulated by PHR1 (Fig. 5.4b). They were also enriched in PHR1 direct targets identified in an independent ChIP-seq experiment (Fig. 5.4c, Supplementary Table 8), PHR1 promoter binding motifs (Fig. 5.13c), and genes involved in biological processes related to PSR (Fig. 5.4d and Supplementary Table 9). PHR1 surprisingly contributed to transcriptional regulation of plant immunity. Five of the twelve clusters (Fig. 5.4a; c3, c6, c7, c8 and c11) were enriched in genes related to plant immune system output; four of these were over-represented for jasmonic acid (JA)

and/or salicylic acid (SA) pathway markers (Fig. 5.4d, c3, c7, c8, and c11; Supplementary Table 9) and three of these four were enriched for PHR1 direct targets (Fig. 5.4c). SA and JA are plant hormone regulators of immunity and at least SA modulates Arabidopsis root microbiome composition (Lebeis et al., 2015).

To further explore PHR1 function in the regulation of plant immunity, we generated transcriptomic time course data for treatment-matched Col-0 seedlings following application of Methyl Jasmonate (MeJA) or the SA analogue Benzothiadiazole (BTH; Supplementary Table 10). We found a considerable over-representation of SA- and JA-activated genes among the PSR-SynCom DEGs (468 versus 251 predicted for SA and 165 vs. 80 predicted for JA; p-value < 0.0001, hypergeometric test) (Fig 5.9c-h; Supplementary Table 7). A large proportion of SA-responsive genes were more strongly expressed in *phr1* and *phr1 phl1* than in Col-0; these were strongly enriched for classical SA-dependent defense genes (Fig. 5.9d-e). A second group of SA-responsive genes that were less expressed in *phr1* and *phr1 phl1* than in Col-0 lacked classical SA-dependent defense genes and were weakly enriched for genes likely contributing to PSR (Fig 5.9d). By contrast, most JA-responsive genes exhibited lower expression in *phr1* and *phr1 phl1* (Fig. 5.9g-h), including a subset of 18 of 46 genes known or predicted to mediate biosynthesis of defense-related glucosinolates (Fig. 5.9i) (Schweizer et al., 2013). This agrees with the recent observation that *phr1* exhibited decreased glucosinolate levels during Pi starvation (Pant et al., 2015). Analyses of SA- and JA- up-regulated genes revealed enrichment of direct PHR1 targets (Fig. 5.4e), consistent with the converse observation that some PHR1-regulated clusters enriched in direct targets were also enriched in defense genes (Fig. 5.4c-d). Many of the SA- and JA- responsive genes were PSR-SynCom DEGs (Figs. 5.4f and 5.9c-h; Supplementary Table 7). Thus, PHR1 directly regulates an unexpected proportion of the plant immune system during PSR triggered by our SynCom.

5.4 PHR1 integrates plant immune system output and phosphate stress response

We tested whether PHR1 also controls the expression of plant defense genes under conditions typically used to study PSR (axenic growth, sucrose present, no microbiota involved). We performed RNA-seq in response to low Pi in these conditions and identified 1482 DEGs in Col-0 and 1161 DEGs in *phr1 phl1* (Figs. 5.5a-b and 5.10; Supplementary Table 11). A significant number of our BTH/SA-activated genes were also up-regulated in *phr1 phl1*, but not in Col-0 in response to low Pi (Fig. 5.4a-b; Supplementary Table 12). A large number of these overlapped with the defense genes induced in *phr1 phl1* by our SymCom (Fig. 5.5c; red ellipse, 113/337 = 33%; clusters c3 and c8 from Fig. 5.4a). At least 14/113 are direct PHR1 targets (Supplementary Table 12).

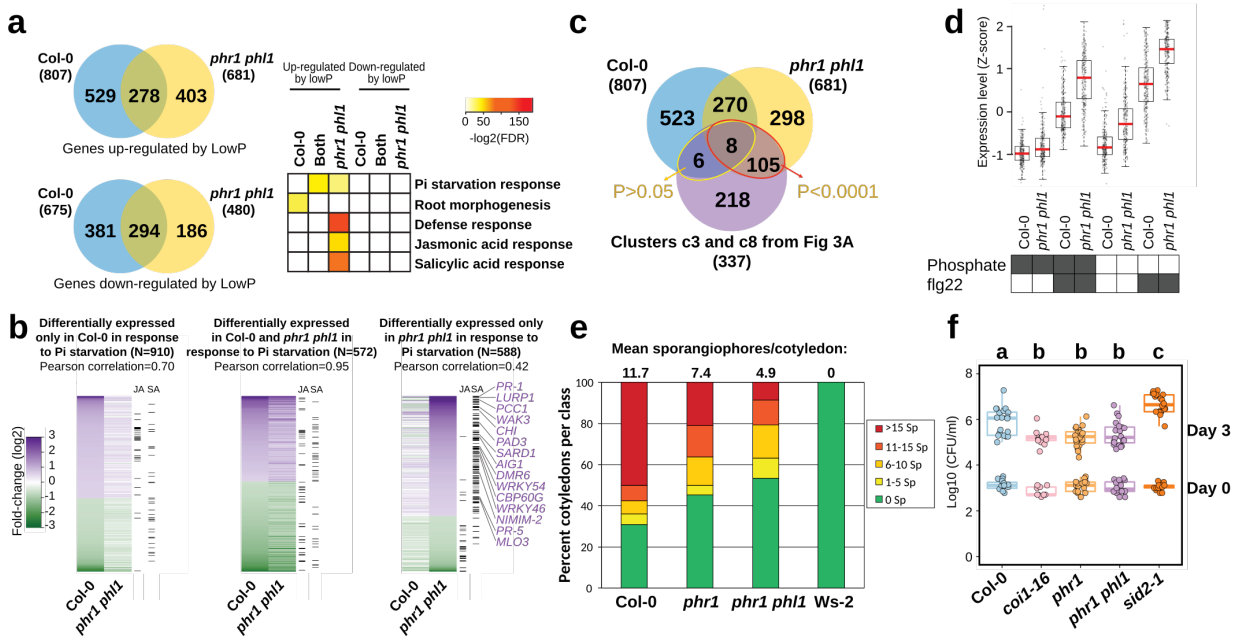


Figure 5.5: **Loss of PHR1 activity results in enhanced activation of plant immunity**. **a**, Venn diagram (left) showing the overlap between genes up-regulated and down-regulated in Col-0 and *phr1 phl1* in response to phosphate starvation. Gene ontology enrichment (right) analyses indicate that defense-related genes are up-regulated exclusively in *phr1 phl1*. The complete enrichment results are shown in Supplementary Table 14. Color key (white

to red) represents the gene ontology enrichment significance shown as -$\log_2$(FDR). White means no enrichment. **b**, Fold-change of genes differentially expressed in Col-0, *phr1 phl1* or in both genotypes in response to phosphate starvation. Columns on the right indicate whether each gene is also up-regulated by MeJA or BTH/SA. Arabidopsis plants were germinated on Johnson medium (1% sucrose) containing 1 mM Pi for 7 d in a vertical position and then transferred to the same medium containing 1% sucrose either alone or supplemented with 1 mM Pi for 12 d. **c**, Venn diagram showing the overlap among genes up-regulated in Col-0 and *phr1 phl1* during a typical PSR (from a) and the defense genes up-regulated in *phr1 phl1* in response to the SynCom (from Fig. 5.4a; clusters c3 and c8). The red ellipse indicates 113 defense genes that were up-regulated in *phr1 phl1* during classical PSR and during PSR triggered by the SynCom; yellow ellipse indicates the 14 genes up-regulated genes under the same conditions. *p*-values refer to enrichment results using hypergeometric tests. **d**, *phr1 phl1* exhibits enhanced transcriptional activation of 251 genes differentially expressed following chronic flg22 exposure. Averaged from six biological replicates. **e**, *phr1* exhibits enhanced disease resistance to the biotrophic oomycete pathogen *Hyaloperonospora arabidopsidis* isolate Noco2. Infection classes were defined by the number of asexual sporangiophores (Sp) per cotyledon and displayed as a color gradient from green (more resistant) to red (more susceptible); the mean number of sporangiophores per cotyledon is noted above each bar. Col-0 and Ws-2 represent susceptible and resistant controls, respectively. More than 100 cotyledons counted per genotype; the experiment was performed at least five times with similar results. *f*, *phr1* mutants exhibit enhanced disease resistance to the hemibiotrophic bacterial pathogen *Pseudomonas syringae* DC3000. The *coi1-16* (n= 9 (day zero), 13 (day three)) and *sid2-1* (n= 16, 20) mutants were controls for resistance and susceptibility, respectively. Col-0 (n=16, 20), *phr1* (n=17, 20), *phr1 phl1* (n=16, 20) and control plants were inoculated under typical experimental conditions: phosphate replete in non-axenic potting soil (Fig. 5.1). The experiment includes at least 9 biological replicates from three independent experiments. Statistical comparisons among genotypes were one-way

ANOVA tests followed by a post-hoc Tukey analysis; genotypes with the same letter above the graph are statistically indistinguishable at 95% confidence.

To underscore the role of PHR1 in the regulation of response to microbes, we analyzed the transcriptional profile of Col-0 and *phr1 phl1* plants exposed to the flagellin peptide flg22. We chose a chronic exposure to flg22 to mimic the condition of plants in contact with a microbiome. We found that *phr1 phl1* plants displayed higher expression of flg22-responsive genes (Rallapalli et al., 2014) than Col-0, independent of phosphate status (Supplementary Text 5; Figs. 5.5d and 5.10a-b; Supplementary Table 11; Supplementary Table 13). This indicates that PHR1 negatively regulates the immune response triggered by flg22.

We hypothesized that *phr1 phl1* would express an altered response to pathogen infection. The *phr1* and *phr1 phl1* mutants exhibited enhanced disease resistance against both *Hyaloperonospora arabidopsidis* isolate Noco2, and *Pseudomonas syringae* DC3000 (Fig. 5.5e-f). Collectively, these results confirm the role of PHR1 as a direct integrator of PSR and the plant immune system.

5.5 Conclusions

Plant responses to phosphate stress are inextricably linked to life in microbe-rich soil. We demonstrate that genes controlling PSR contribute to assembly of a normal root microbiome. Surprisingly, our SynCom enhanced the activity of PHR1, the master regulator of the PSR, in plants grown under limited phosphate. This led to our discovery that PHR1 is a direct regulator of a functionally relevant set of plant immune system genes. Despite being required for the activation of JA-responsive genes during PSR (Khan et al., 2016), we found that PHR1 is unlikely to be a general regulator of this response (Fig. 5.10c-e; Supplementary Table 12), but rather may fine-tune JA response in specific biological contexts.

We demonstrate that PSR and immune system outputs are directly integrated by PHR1 (and, likely, PHL1). We provide a mechanistic explanation for previous disparate observations that PSR and defense regulation are coordinated and implications that PHR1 is the key

regulator (Zhao et al., 2013; Lu et al., 2014; Hiruma et al., 2016; Khan et al., 2016). We provide new insight into the intersection of plant nutritional stress response, immune system function, and microbiome assembly and maintenance; systems that must act simultaneously and coordinately in natural and agricultural settings. Our findings will drive investigations aimed at enhancing phosphate use efficiency using microbes.

5.6 Supplementary text

5.6.1 General features of the root microbiota in wild soil

In line with many recent plant 16S root microbiome census experiments, we found that bacterial EC communities were less diverse than those in bulk soil (Figs. 5.6a-b). We recapitulated previously observed enrichments of Actinobacteria, Firmicutes and Proteobacteria inside the root, and depletions of Acidobacteria, Verrucomicrobia, and Chloroflexi (Fig. 5.6c and Supplementary Table 1). As expected, we found a large difference between root endophytic and bulk soil communities, followed by soil dig (Fig. 5.6d).

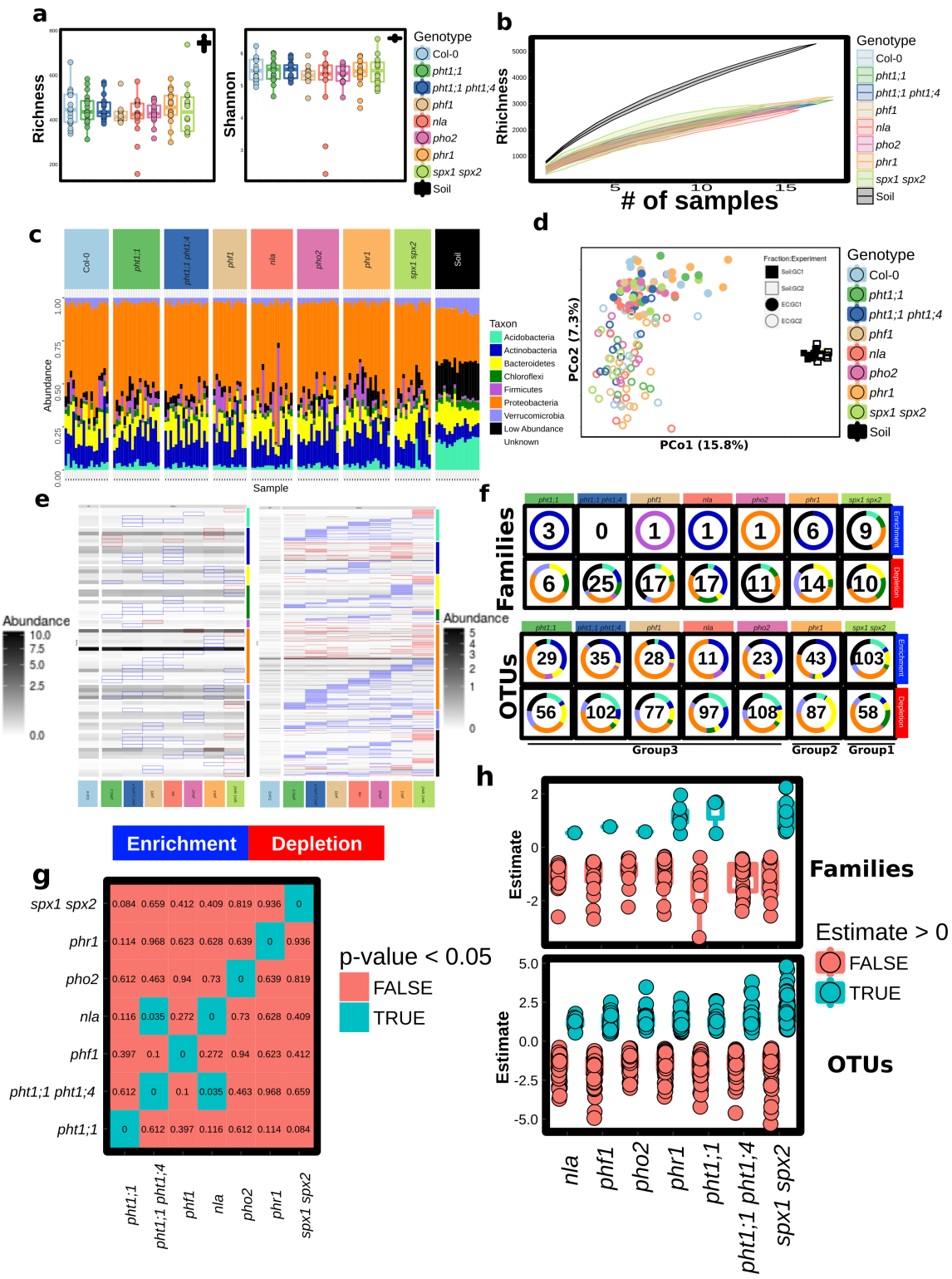


Figure 5.6: **The Arabidopsis PSR alters highly specific bacterial taxa abundances**. **a**, Alpha diversity of bacterial root microbiome in wild-type Col-0, PSR mutants and bulk

soil samples. We used ANOVA methods and no statistical differences were detected between plant genotypes after controlling for experiment. **b**, Additive beta-diversity curves showing how many OTUs are found in bulk soil samples or root endophytic (EC) samples of the same genotype as more samples (pots) are added. The curves show the mean and the 95 % confidence interval calculated from 20 permutations. **c**, Phylum-level distributions of plant root endophytic communities from different plant genotypes and bulk soil samples. **d**, Principal Coordinates Analysis based on Bray-Curtis dissimilarity of root and bulk soil bacterial communities showing a large effect of experiment on variation, as expected according to previous studies (Lundberg et al., 2012). For a-d the number of biological replicates per genotype and soil are: Col-0 (n=17), *pht1;1* (n=18), *pht1;1 pht1;4* (n=17), *phf1* (n=13), *nla* (n=16), *pho2* (n=16), *phr1* (n=18), *spx1 spx2* (n=14) and Soil (n=17). **e**, Bacterial taxa that are differentially abundant (DA) between PSR mutants and Col-0. Each row represents a bacterial Family (left) or OTU (right) that shows a significant abundance difference between Col-0 and at least one mutant. The heat-map grey scale shows the mean abundance of the given taxa in the corresponding genotype, and significant enrichments and depletions with respect to Col-0 are indicated with a red or blue rectangle, respectively. Taxa are organized by phylum shown on the right bar colored according to f. **f**, Doughnut plot showing Family-level (top) and OTUs- level (bottom) differences in endophytic root microbiome compositions between mutants (columns) and Col-0 plants. The number inside each doughnut indicates how many bacterial Families are enriched or depleted in each mutant with respect to Col-0, and the colors in the doughnut show the phylum level distribution of those differential abundances. **g**, Tables of p-values from Monte Carlo pairwise comparisons between mutants. A significant p-value (cyan) indicates that two genotypes are more similar than expected by chance. Results of Family level comparison are shown. This plot should be compared with the corresponding OTU-level plot in Fig. 5.2d. **h**, Distributions of plant genotypic effects on taxonomic abundances at the Family (up) or OTU (down) level. For each genotype, the value of the linear model coefficients for individual OTUs or Families is plotted grouped by

their sign. Positive values indicate that a given taxon has increased abundance in a mutant with respect to Col-0, while a negative value represents the inverse pattern. Only coefficients from significant comparisons are shown. The number of taxa (*ie.* points) on each box and whisker plots is indicated in the corresponding doughnut plot in f.

5.6.2 Control experiments pertinent to figures 2 and 3

For the design of these experiments, we used as a reference the PSR studied in agar under axenic conditions and long day (where the PSR was originally defined). In this setting, 1 mM - 2 mM Pi is considered full Pi. In our study of the PSR, the optimal Pi concentration in media for microbial growth is typically higher than 2 mM. To avoid excessive stress that could compromise the viability of our SynCom and/or exacerbate production of toxic secondary metabolites that damage the plant, we selected 50 μM Pi (20 times lower than 1 mM, but still a contact point to published data). Plants grown at this Pi concentration (in a media free of sucrose and in a short day regime) showed marginal activation of the PSR and a reduced Pi concentration in the shoot (Figs 5.3a-b and 5.7d) as compared with plants grown on 1 mM Pi plates. We reasoned that these conditions would facilitate a nutritional competition between plant and the SynCom in the absence of a steady state full induction of PSR, thus providing an excellent scenario for the study of PSR influenced by microbes. The plants showed visible symptoms of PSR only in the presence of the SynCom. We could therefore simultaneously trigger two different stresses (biotic and abiotic) with a single factor (SymCom) and ask how the plants reacted to it.

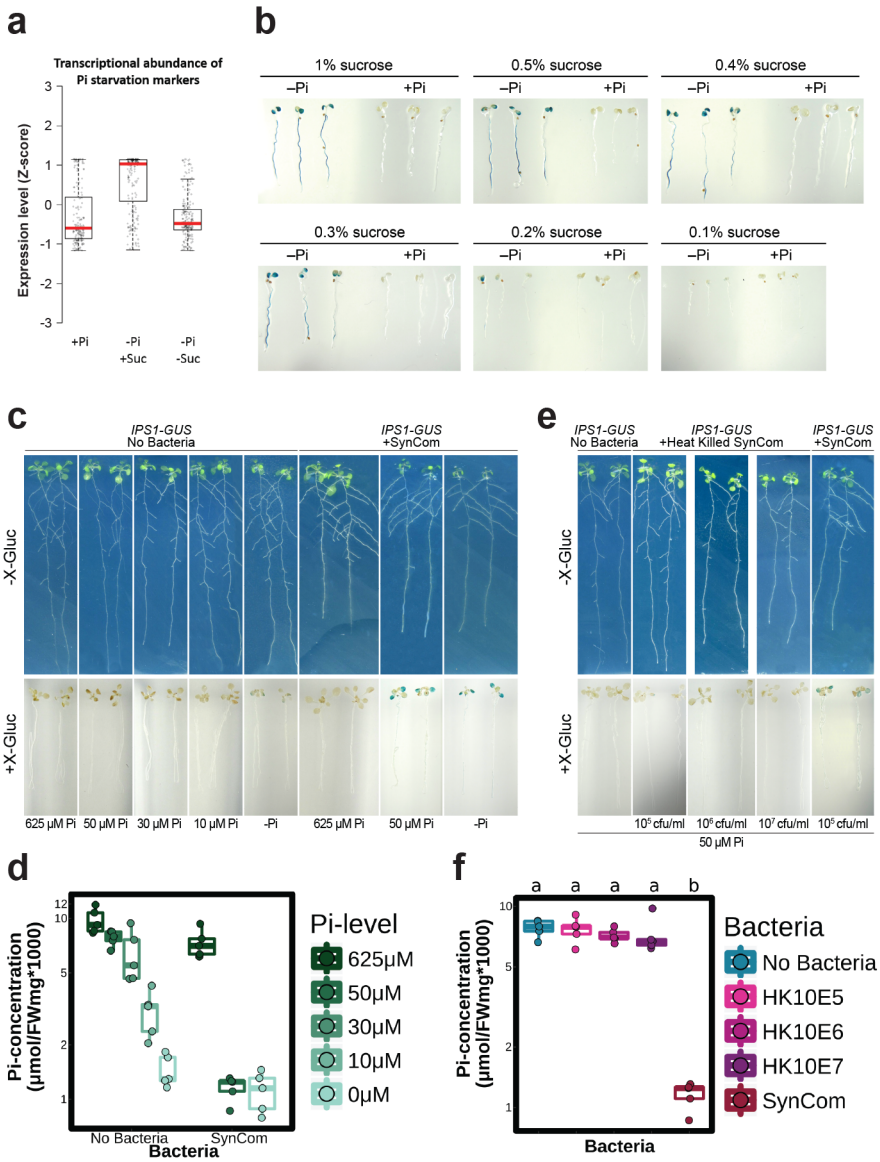


Figure 5.7: **The SynCom induces PSR independently of sucrose in Arabidopsis**. **a**, Expression analysis of a core of 193 PSR marker genes in an RNA-seq experiment using Col-0 plants. The RPKM expression values of these genes were z-score transformed and used to generate box and whiskers plots that show the distribution of the expression values of this gene set. Plants were grown in Johnson medium containing replete [1 mM Pi; (+Pi)] or stress [5 μM Pi; (-Pi)] Pi concentrations with (+Suc) or without (-Suc) 1 % sucrose. **b**, Expression analysis of the reporter constructs IPS1:GUS (n=20). In this construct, the promoter region of IPS1, highly induced by low Pi, drives the expression of GUS. Plants were

grown in Johnson medium +Pi or -Pi at different percentages of sucrose. These results show that sucrose is required for the induction of the PSR in typical sterile conditions. Images shown are representative of the 20 plants analyzed in each case. **c**, Top: Plants grown in sterile conditions at different Pi concentrations [left (No Bacteria)] or with a SynCom [right (+SynCom)]. Bottom: Histochemical analysis of Beta-glucoronidase (GUS) activity in overexpressing IPS1:GUS plants (n=20) from top panel. Images shown are representative of the 20 plants analyzed in each case. **d**, Pi concentration in plant shoots from c , in all cases n=5. Analysis of Variance indicated a significant effect of the Pi level in the media (F = 44.12, df = 1, p-value = 9.72e-8), the presence of SynCom (F = 32.61, df = 1, p-value = 1.69e-6) and a significant interaction between those two terms (F = 4.748, df = 1, p-value = 0.036) on Pi accumulation. **e**, Top: Plants grown in axenic conditions (No Bacteria), with a concentration gradient of heat-killed SynCom [2 h at 95°C, (+Heat killed SynCom)] or with SynCom alive. Bottom: Histochemical analysis of GUS activity in overexpressing IPS1:GUS plants (n=15) from top panel. All plants were grown at 50 M Pi. Images shown are representative of the 15 plants analyzed in each case. **f**, Quantification of Pi concentration in plant shoots from e, (in call cases n=5). The SynCom effect on Pi concentration requires live bacteria. Plants were germinated on Johnson medium containing 0.5 % sucrose, with 1 mM Pi for 7 d in a vertical position, then transferred to 0, 10, 30, 50, 625 μM Pi media (without sucrose) alone or with the Synthetic Community at 10^5 c.f.u/mL (only for the conditions 0, 50 and 625 μM Pi), for another 12 d. For the heat-killed SynCom experiments, plants were grown as above. Heat-killed SynComs were obtained by heating different concentrations of bacteria 10^5 c.f.u / mL, 10^6 c.f.u / mL and 10^7 c.f.u / mL at 95°C for 2 h in an oven. The whole content of the heat-killed SynCom solutions were add to the media. In all cases, addition of the SynCom did not change significantly the final Pi concentration or the pH in the media. Letters indicates grouping based on ANOVA and Tukey post-hoc test at 95 % confidence, conditions with the same letter are statistically indistinguishable.

We performed several sets of control experiments. First, to eliminate the possibility that the SynCom merely mediated sucrose fertilization to restore the PSR transcriptional response, we supplemented Col-0 plants with a concentration gradient of heat-killed SynCom (section 5.7.3). These treatments did not change either the induction of IPS1:GUS13 or the Pi concentration in the plant shoot (Fig 5.7e-f). Transcriptional PSR was triggered by our SynCom in low phosphate even without sucrose in the media, a nutritional situation more closely related to growth in wild soil (Fig. 5.3b). Second, transgenic reporter IPS1:GUS plants (Bustos et al., 2010) growing in reduced Pi conditions accumulated reduced shoot Pi concentrations, but the expression of the PSR reporter was not induced even in the absence of supplemental Pi, (Fig. 5.7c) where seedlings achieved Pi concentrations similar to plants grown in the presence of the SynCom (Fig. 5.7d). However, this marker was induced at low Pi levels in the presence of the SynCom (Fig. 5.7c-d). Therefore, nutritional competition between plant and microbes might explain the reduction in the Pi-concentration in the shoots of plants grown at 50 μM Pi to a level similar to plants grown without Pi supplementation, but it is not enough to explain the induction of the IPS1:GUS reporter or the fact that our bacterial SynCom enhanced the PSR (Fig. 5.3c). Additionally, the finding that PHR1 directly regulates a large proportion of the plant immune system during the PSR triggered by our bacterial SynCom (Fig. 5.4) argues against Pi exhaustion as the cause of microbial triggering of plant PSR. Third, Phosphite [KH_2PO_3; (Phi)] is a non-metabolizable analog of Pi and its accumulation delays the PSR (Jost et al., 2015). Col-0 plants pre-treated with Phi had low Pi content (Fig. 5.8a) but only weakly induced core PSR markers (Fig. 5.8b), even in the presence of our SynCom. We detected similar PHR1 PHL1-dependent induction of the core PSR markers using either replete (1 mM) or low (5 μM) Pi pre-treatments across genotypes, indicating that after 12 days the SynCom induces a long-lasting response to low Pi (Fig. 5.8a-b). Finally, plants colonized by the SynCom also mimicked developmental phenotypes of PSR: a shorter main root (Fig. 5.8c-d) and more lateral roots than non-inoculated plants (Fig. 5.8c, e and f). In sum, the transcriptional PSR responses we observed in the presence

of our SynCom were activated by canonical PSR mechanisms and we infer that plants have evolved a mechanism to coordinate defense and PSR.

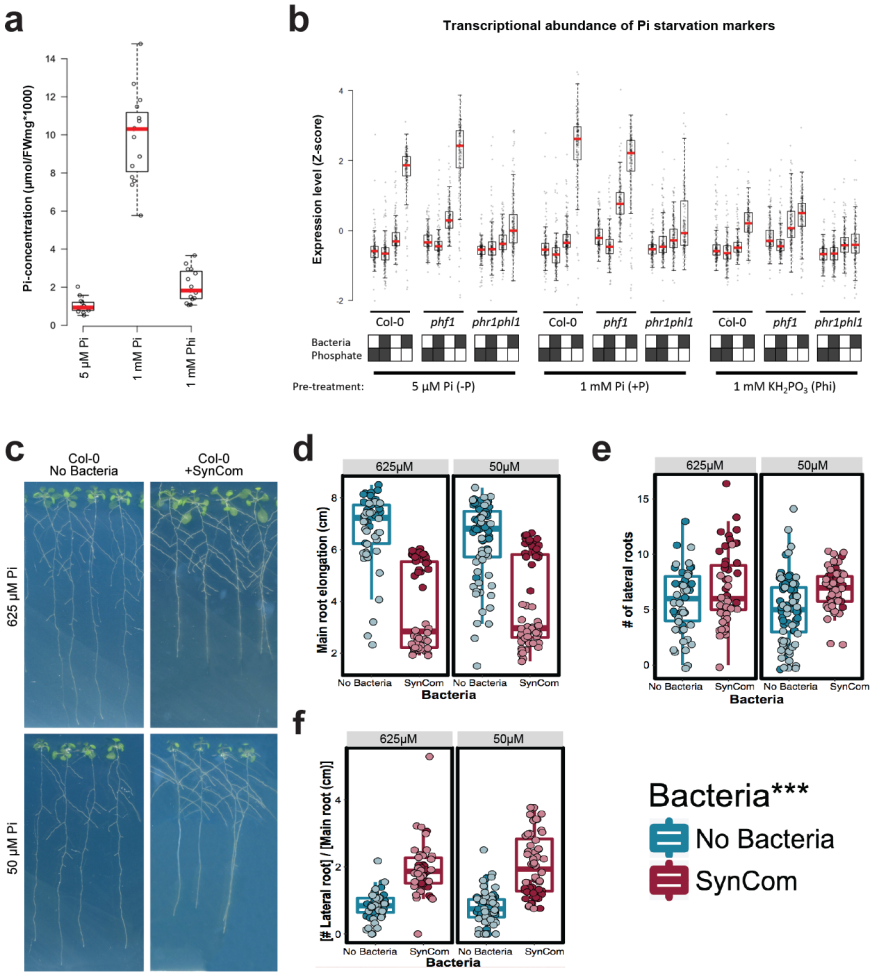


Figure 5.8: B**acteria induce the PSR using the canonical pathway in Arabidopsis**. **a**, Pi concentration in the shoot of Col-0 plants germinated in three different conditions, 5 μM Pi (-Pi) (n=14), 1 mM Pi (+Pi) (n=15) and 1 mM KH_2PO_3 (Phi) (n=15) for 7 days. Phi is a non-metabolizable analog of Pi; its accumulation delays the response to phosphate stress. **b**, Expression profile analysis of a core of PSR-marker genes in Col-0, *phf1* and *phr1 phl1*. The RPKM expression values of these genes were z-score transformed and used to generate box and whiskers plots that show the distribution of the expression values of this gene set. Plants were germinated in three different conditions, 5 phr1 phl1M Pi (-Pi), 1 mM Pi (+Pi) and 1 mM KH_2PO_3 (Phi) and then transferred to low Pi (50 μM Pi) and high

Pi (625 μM Pi) alone or with the SynCom for another 12 d. The figure shows the average measurement of four biological replicates. **c**, Phenotype of plants grown in axenic conditions at 625 μM Pi (Top) or at 50 μM Pi (Bottom) [left (No Bacteria)] or with a SynCom [right (+SynCom)]. Images showed here are representative of the total number of plants analyzed in each case as noted below. **d**, Quantification of the main root elongation, **e**, Number of lateral roots per plant, and **f**, Number of lateral roots per cm of main root in plants from c. For d, e and f the number of biological replicates are: 625 μM No Bacteria (n=48), 625 μM + SynCom (n=46), 50 μM No Bacteria (n=73), and 50 μM SynCom (n=56), distributed across two independent experiments indicated with different shades of color. Measurements were analyzed with ANOVA while controlling for biological replicate. Asterisks denote a significant effect (p-value $< 1e-16$) of treatment with SynCom for the three phenotypes in d, e and f. In all cases, neither the interaction between Pi and Bacteria, nor Pi concentrations alone had a significant effect and were dropped from the ANOVA model. In all cases, residual quantiles from the ANOVA model were compared with residuals from a Normal distribution to confirm that the assumptions made by ANOVA hold (see code on GitHub for details, see section 5.7.10).

5.6.3 General features of the SynCom colonization experiment in agar

After SynCom inoculation we also found that agar- and root-associated microbiomes were markedly different from the input and from each other (Figs. 5.3d and 5.11e). We also identified eight strains as robust root colonizers regardless of plant genotype or Pi levels (Fig. 5.3d, Supplementary Table 5).

5.6.4 Differentially expressed genes in plants growing in the presence of the SynCom

a

	Col-0	*phr1*	*phr1/phl1*
LowPi	1 (1 up; 0 down)	0 (0 up; 0 down)	0 (0 up; 0 down)
Bacteria	519 (126 up; 393 down)	552 (297 up; 255 down)	276 (63 up; 213 down)
LowPi*Bacteria	2537 (1579 up; 958 down)	471 (332 up; 139 down)	232 (158 up; 74 down)

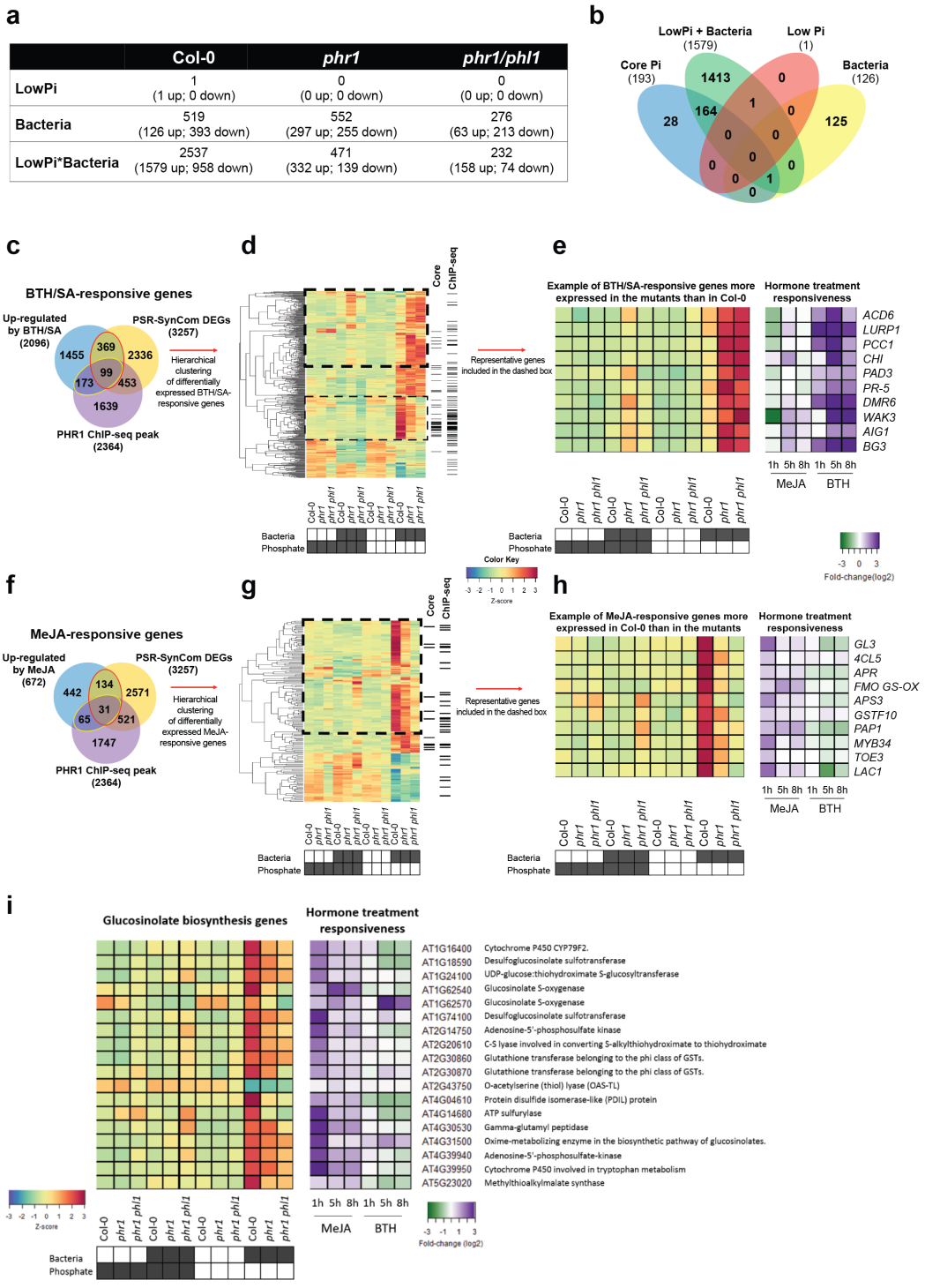


Figure 5.9: **PHR1 controls the balance between the SA and JA regulons during the PSR induced by a 35-member SynCom**. **a**, Total number of differentially expressed genes (FDR ≤ 0.01 and minimum of 1.5X fold-change) in Col-0, *phr1* and *phr1 phl1* with respect to Low Pi (50 μM Pi), bacteria presence and the interaction between low Pi and bacteria. **b**, Venn diagram showing the overlap between the PSR marker genes (Core Pi) and the genes that were up-regulated in Col-0 by each of the three variables analyzed. The combination of bacteria and low Pi induced the majority (85%) of the marker genes. **c**, Venn diagram showing the overlap among PSR-SynCom DEGs, genes up-regulated by BTH treatment of Arabidopsis seedlings, and the direct targets of PHR1 identified by ChIP-seq. The red ellipse indicates BTH/SA-responsive genes that were differentially expressed. The yellow ellipse indicates SA-responsive genes that were bound by PHR1 in a ChIP-seq experiment. **d**, Hierarchical clustering analysis of genes in c. Columns on the right indicate those genes that belong to the core PSR marker genes ('core' lane) or that contain a PHR1 ChIP-seq peak ('ChIP-seq' lane). **e**, Examples of typical SA-responsive genes are shown on the right along with their expression profiles in response to MeJA or BTH/SA treatment compared to Col-0. **f**, Venn diagram showing the overlap among DEG from this work (PSR-SynCom), genes up-regulated by MeJA treatment of Arabidopsis seedlings and the genes bound by PHR1 in a ChIP-seq analysis. Red ellipse indicates JA-responsive genes that were differentially expressed. Yellow ellipse indicates 96 JA-responsive genes that were bound by PHR1 in a ChIP-seq experiment. **g**, Hierarchical clustering analysis of genes in f. The columns on the right are the same as in d. **h**, Examples of well-characterized JA-responsive genes are shown on the right along with their expression profiles in response to BTH and MeJA treatments obtained in an independent experiment. **i**, Heatmap showing the expression profile of 18 genes that were differentially expressed in our experiment and participate in the biosynthesis of glucosinolates. The transcriptional response to BTH/SA and MeJA treatments is shown on the right and was determined in an independent experiment in which Arabidopsis seedlings were sprayed with either hormone. The gene IDs and the enzymatic activity of the encoded

proteins are shown on the right. Results presented in this figure are based on ten biological replicates for Col-0 and *phr1* and six for *phr1 phl1*. The color key (blue to red) related to d, e, g, h, i represents gene expression as Z-scores and the color key (green to purple) related to e, h, i represents gene expression as $\log_2$ fold-changes.

We identified 3257 differentially expressed genes that responded to either low Pi, presence of the SynCom, or the interaction of both (hereafter PSR-SynCom DEGs) (Fig. 5.9a-b and Supplementary Table 6). In agreement with the fact that PSR is not activated in Col-0 grown in the low Pi conditions we used, only one gene showed a significant change in transcript levels in response to low phosphate availability (Fig. 5.9a-b and Supplementary Table 6). In contrast, 1579 genes, including 164/193 (85%) of the core PSR marker set, were up-regulated and 958 genes were repressed in response to low phosphate when the SynCom was present (Fig. 5.9a-b). In this experiment, plants were grown for 7 days in Johnson medium containing 1 mM Pi, and then transferred for 12 days to low (50 μM Pi) and high Pi (625 μM Pi) conditions alone or with the SynCom. No sucrose was added to the medium.

PHR1 negatively regulates the expression of a set of SA-responsive genes during co-cultivation with the SynCom (Fig. 5.9c-e). 468 BTH/SA-responsive genes that were differentially expressed in response to the SynCom and low phosphate. A total of 99 of these genes (21%) are likely direct targets of PHR1. 272 SA-responsive genes were bound by PHR1 in a ChIP-seq experiment (see Figs. 5.4e). Approximately one-third of them (99/272) were differentially expressed in the SynCom experiment (Fig. 5.9c). Hierarchical clustering analysis showed that nearly half of the BTH/SA-induced genes that were differentially expressed in our experiment are more expressed in *phr1* or *phr1 phl1* mutants compared to Col-0 (5.9d, dashed box). A subset of the SA marker genes is less expressed in the mutant lines (5.9d, thin dashed box). This set of genes is also enriched in the core PSR markers and in PHR1 direct targets (p-value < 0.001; hypergeometric test), indicating that PHR1 can function as a positive activator of a subset of SA-responsive genes. Importantly, these genes are not typical

components of the plant immune system but rather encode proteins that play a role in the physiological response to low phosphate availability (e.g., phosphatases and transporters).

PHR1 activity is required for the activation of JA-responsive genes during co-cultivation with the SynCom (Fig. 5.9f-h). 165 JA-responsive genes that were differentially expressed by the presence of Syncom and low phosphate. Thirty-one of these (19%) were defined as direct targets of PHR1. 96 JA-responsive genes were bound by PHR1 in a ChIP-seq experiment. Approximately one-third of them (31/96) were differentially expressed in the SynCom experiment (Fig. 5.9f). Hierarchical clustering analysis showed that almost 75% of the JA-induced genes that were differentially expressed in our experiment are less expressed in the *phr1* mutants (5.9g, dashed box).

We found 18 genes that were differentially expressed in our experiment and participate in the biosynthesis of glucosinolates (Schweizer et al., 2013). In general, these genes showed lower expression in the *phr1* mutants indicating that PHR1 activity is required for the activation of a sub-set of JA-responsive genes that mediate glucosinolate biosynthesis/ MeJA induces the expression of these glucosinolate biosynthetic genes, whereas BTH represses many of them (Fig. 5.9i).

5.6.5 General features of Col-0 and phr1 phl1 plants exposed to flg22

To further accentuate the role of PHR1 in the direct regulation of response to microbes, we chose a chronic exposure to flg22. We observed that 251 of the 2690 (9.33 %) genes up-regulated during an acute exposure to flg22 (between 8 and 180 min) (Rallapalli et al., 2014) were also up-regulated in our experiment (Fig. 5.10a-b; Supplementary Table 11; Supplementary Table 13) and that this gene set contained more PHR1 direct targets than expected by chance (31 observed versus 22 expected, p-value = 0.0297).

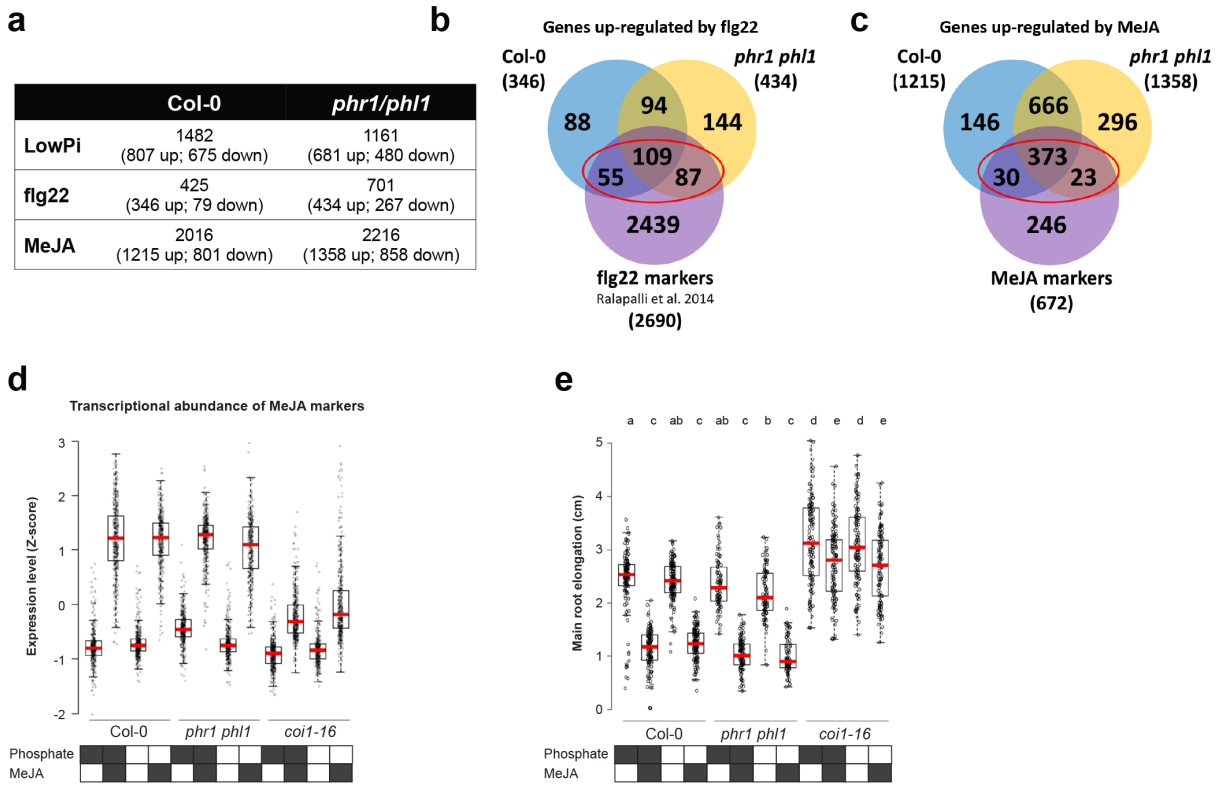


Figure 5.10: **PHR1 activity effects on flg22 and MeJA-induced transcriptional responses**. **a**, Total number of differentially expressed genes (FDR ≤ 0.01 and minimum of 1.5X fold-change) in Col-0 and *phr1 phl1* with respect to low Pi (50 μM Pi), flg22 treatment (1 μM) and MeJA (10 μM). In this experiment, plants were grown for 7 days in Johnson medium containing 1 mM Pi, and then transferred for 12 days to low (50 μM Pi) and high Pi (625 μM Pi) conditions alone, or in combination with each treatment. Sucrose was added to the medium at a final concentration of 1 %. **b**, Venn diagram showing the overlap among genes that were up-regulated by chronic exposure to flg22 in Col-0 and in *phr1 phl1* and a literature-based set of genes that were up-regulated by acute exposure (between 8 to 180 min) to flg22 (Rallapalli et al., 2014). The red ellipse indicates the 251 chronic flg22-responsive genes defined here. **c**, Venn diagram showing the overlap among genes that were up-regulated by chronic exposure to MeJA in Col-0 and in *phr1 phl1* in this work and a set of genes that were up-regulated by MeJA treatment of Arabidopsis seedlings (between 1 and 8 hours). The

red ellipse indicates the intersection of JA-responsive genes identified in both experiments. **d**, Col-0 and *phr1 phl1* exhibit similar transcriptional activation of 426 common JA-marker genes (c) independent of phosphate concentration. As a control we used coi1-16, a mutant impaired in the perception of JA. The gene expression results are based on six biological replicates per condition. e, Growth inhibition of primary roots by MeJA. Root length of wild-type Col-0 (n= 125 (+ Pi - MeJA), 120 (+ Pi + MeJA), 126 (- Pi - MeJA), 125 (- Pi + MeJA)), *phr1 phl1* (n=85, 103, 90, 80) and the JA perception mutant coi1-16 (n= 125, 120, 124, 119) was measured after 4 days of growth in the presence or not of MeJA with or without 1 mM Pi. Letters indicate grouping based on multiple comparisons from a Tukey post-hoc test at 95 % confidence. In agreement with the RNA-seq results, no difference in root length inhibition was observed between Col-0 and *phr1 phl1*.

5.7 Methods

5.7.1 Census study experimental procedures

For experiments in wild soil, we collected the top-soil (approx. 20 cm) from a site free of pesticide and fertilizer at Mason Farm (MF; North Carolina, USA; +35°53' 30.40", -79°1'5.37") (Lundberg et al., 2012). Soil was dried, crushed and sifted to remove debris. To improve drainage, soil was mixed 2:1 volume with autoclaved sand. Square pots (2 x 2 inch square) were filled with the soil mixture and used to grow plants. Soil micronutrient analysis is published by Lundberg et al. (2012).

All *Arabidopsis thaliana* mutants used in this study were in the Columbia (Col-0) background (Supplementary Table 16). All seeds were surface-sterilized with 70% bleach, 0.2% Tween-20 for 8 minutes, and rinsed 3X with sterile distilled water to eliminate any seed-borne microbes on the seed surface. Seeds were stratified at 4°C in the dark for 2 days.

To determine the role of phosphate starvation response in controlling microbiome composition, we analyzed five mutants related to the Pi-transport system (*pht1;1*, *pht1;1 pht1;4*, *phf1*, *nla*, and *pho2*) and two mutants directly involved in the transcriptional regulation

of the Pi-starvation response (*phr1* and *spx1 spx2*). All these genes are expressed in roots (Bustos et al., 2010; Shin et al., 2004; González et al., 2005; Huang et al., 2013; Lin et al., 2013; Puga et al., 2014).

Seeds were germinated in sterile square pots filled with MF soil prepared as described above. We also used pots without plants as "bulk soil" controls. All pots, including controls, were watered from the top with non-sterile distilled water to avoid chlorine and other tap water additives 2 times a week. Plants were grown in growth chambers with a 16-h dark/8-h light regime at 21°C day 18°C night for 7 weeks. In all experiments, pots with plants of different genotypes were randomly placed in trays according to true random numbers derived from atmospheric noise; we obtained those numbers from `www.random.org`. We positioned trays in the growth chamber without paying attention to the pots they contained, and we periodically reshuffled them without paying attention to the pot labels.

Plants and bulk soil controls were harvested and their endophytic compartment (EC) microbial communities isolated as described in Lundberg et al. (2012). DNA extraction was performed using 96-well format MoBio PowerSoil Kit (MOBIO Laboratories) following the manufacturers instruction.

The method of Ames (1966) was used to determine the phosphate concentration in the shoots of seedlings grown on different Pi regimens and treatments. Main root length elongation was measured using ImageJ software (Barboriak et al., 2005) and for shoot area and number of lateral roots WinRhizo software (Arsenault et al., 1996) was used.

5.7.2 Processing of 16S sequencing data

For wild soil experiment 16S sequencing, we processed libraries according to Caporaso et al. (2012). Three sets of index primers were used to amplify the V4 (515F-806R) region of the 16S rRNA gene of each sample. In each case, the reverse primer had a unique molecular barcode for each sample (Caporaso et al., 2012). PCR reactions with ∼20 ng template were performed with 5 Prime Hot Master Mix in triplicate using plates 2, 4 and 5 from the 16S rRNA Amplification Protocol $s \sim$. PCR blockers mPNA and pPNA (Lundberg et al., 2013)

were used to reduce contamination by plant host plastid and mitochondrial 16S amplicon. The PCR program used was:

1. 95°C for 180 seconds
2. 35 cycles of:
    (a) 95°C for 45 seconds
    (b) 78°C for 30 seconds (PNA annealing)
    (c) 50°C for 60 seconds
    (d) 72°for 90 seconds
3. 12°C for 5 minutes
4. 4°C for ever

Reactions were purified using AMPure XP magnetic beads and quantified with Quant IT Picogreen. Amplicons were pooled in equal amounts and then diluted to 5.5 pM for sequencing. Samples were sequenced on an Illumina MiSeq machine at UNC, using a 500-cycle V2 chemistry kit. The library was spiked with 25% PhiX control to increase sequence diversity. The raw data for the wild soil experiments is available in the EBI Sequence Read Archive (accession PRJEB15671).

For SynCom experiment 16S library, we amplified the V3-V4 regions of the bacterial 16S rRNA gene using primers 338F (5'-ACTCCTACGGGAGGCAGCA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3'). Libraries were created using a modified version of the Lundberg et al. (2013) protocol. Basically, the molecule-tagging step was changed to an exponential amplification to account for low DNA yields with the following reaction:

- 5μL of Kapa Enhancer
- 5μL of Kapa Buffer A

- 1.25μL of 5μM 338F
- 1.25μL of 5μM 806R
- 0.375μL of mixed PNAs (1:1 mix of 100μM pPNA and 100μM mPNA)
- 0.5μL Kapa dNTPs
- 0.2μ Kapa Robust Taq
- 5μL DNA

With the following temperature cycling:

1. 95°C for 60 seconds
2. 24 cycles of:
    (a) 95°C for 15 seconds
    (b) 78°C for 10 seconds (PNA annealing)
    (c) 50°C for 30 seconds
    (d) 72°for 30 seconds
3. 12°C for 5 minutes
4. 4°C for ever

The PCR product was cleaned with AMPure XP magnetic beads. Following PCR cleanup to remove primer dimers, the PCR product was indexed using the same reaction and 9 cycles of the cycling conditions described in Lundberg et al. (2013). Sequencing was performed at UNC on an Illumina MiSeq instrument using a 600-cycle V3 chemistry kit. The raw data for the SynCom experiments is available in the EBI Sequence Read Archive accession PRJEB15671.

For wild soil census analysis, sequences from each experiment were pre-processed following standard method pipelines from (Lebeis et al., 2015; Lundberg et al., 2012). Briefly, sequence pairs were merged, quality-filtered and de-multiplexed according to their barcodes. The resulting sequences were then clustered into Operational Taxonomic Unit (OTUs) using UPARSE (Edgar, 2013) implemented with USEARCH7.1090, at 97% percent identity. Representative OTU sequences (Supplementary Dataset 1) were taxonomically annotated with the RDP classifier (Wang et al., 2007) trained on the Greengenes database (4/February/2011; Supplemental Dataset 1). We used a custom script (`https://github.com/surh/pbi/blob/master/census/1.filter_contaminants.r`) to remove organellar OTUs, and OTUs that had no more than a kingdom-level classification, and an OTU count table was generated (Supplementary Table 1, Supplementary Dataset 1).

SynCom sequencing data were processed with MT-Toolbox (Yourstone et al., 2014). Categorizable reads from MT-Toolbox (i.e. reads with correct primer and primer sequences that successfully merged with their pair) were quality filtered with Sickle (Joshi and Fass, 2011) by not allowing any window with Q-score under 20, and trimmed from the 5 end to a final length of 270 bp. The resulting sequences were matched to a reference set of the strains in the SynCom generated from Sanger sequences, the sequence from a contaminant strain (47Yellow) that grew in the plate from strain 47 (Supplementary Table 2) and Arabidopsis organellar sequences. Sequence mapping was done with USEARCH7.1090 with the option usearch_global at a 98% identity threshold. 90% of sequences matched an expected isolate, and those sequence mapping results were used to produce an isolate abundance table. The remaining unmapped sequences were clustered into OTUs with the same settings used for the census experiment, the vast majority of those OTUs belonged to the same families as isolates in the SynCom, and were probably unmapped due to PCR and/or sequencing errors. We combined the isolate and OTU count tables into a single master table. The resulting table was processed and analyzed with the code at (`https://github.com/surh/pbi/blob/master/syncom/7.syncomP_16S.r`). Matches to Arabidopsis organelles were discarded. PCR blanks

were included in the sequencing and the average counts per strain observed on those blanks were subtracted from the rest of the samples following Nguyen et al. (2015). Figure 5.11 shows the number of usable reads across samples, and the remaining number after subtracting sterile controls (blanks).

5.7.3 *In vitro* plant growth conditions

For physiological, transcriptional analysis or pathology experiments, we used *phr1*, *phr1 phl1*, *phf1*, and *coi1-16*, *sid2-1* mutants, which are all in the Col-0 genetic background (Supplementary Table 16). For all physiological and transcriptional analysis *in vitro*, Arabidopsis seedlings were grown on Johnson medium [KNO_3 (0.6 g/L), $Ca(NO_3)_2$*$4H_2O$ (0.9 g/L), $MgSO_4$*$7H_2O$ (0.2 g/L), KCl (3.8 mg/L), H_3BO_3 (1.5 mg/L), $MnSO_4$*H_2O (0.8 mg/L), $ZnSO_4$*$7H_2O$ (0.6 mg/L), $CuSO_4$*$5H_2O$ (0.1 mg/L), H_2MoO_4 (16.1 g/L), $FeSO_4$*$7H_2O$ (1.1 mg/L), Myo-Inositol (0.1 g/L), MES (0.5 g/L), pH 5.6 - 5.7] solidified with 1% bacto-agar (BD, Difco). Media were supplemented with Pi (KH_2PO_4) at distinct concentrations depending on the experiment; 1 mM Pi was used for complete medium and approximately 5 μM Pi (traces of Pi in the agar) was the Pi concentration in the medium not supplemented with Pi. Unless otherwise stated, plants were grown in a growth chamber in a 15-h dark/9-h light regime (21°C day /18°C night).

For Synthetic Community experiments, plants were germinated on Johnson medium containing 0.5% sucrose, with 1 mM Pi, 5 μM Pi or supplemented with KH_2PO_3 (phosphite) at 1 mM for 7 d in a vertical position, then transferred to 50 μM Pi or 625 μM Pi media (without sucrose) alone or with the Synthetic Community at 10^5 c.f.u/mL, for another 12 d. For the heat-killed SynCom experiments, plants were grown as above. Heat-killed SynComs were obtained by heating different concentrations of bacteria: 10^5 c.f.u/mL, 10^6 c.f.u/mL and 10^7 c.f.u/mL at 95°C for 2 h in an oven. The whole content of the heat-killed SynCom solutions were added to the media.

For the functional activation of the PSR by the SynCom, plants were germinated on Johnson medium containing 0.5% sucrose, 1 mM Pi for 7 d in a vertical position, then

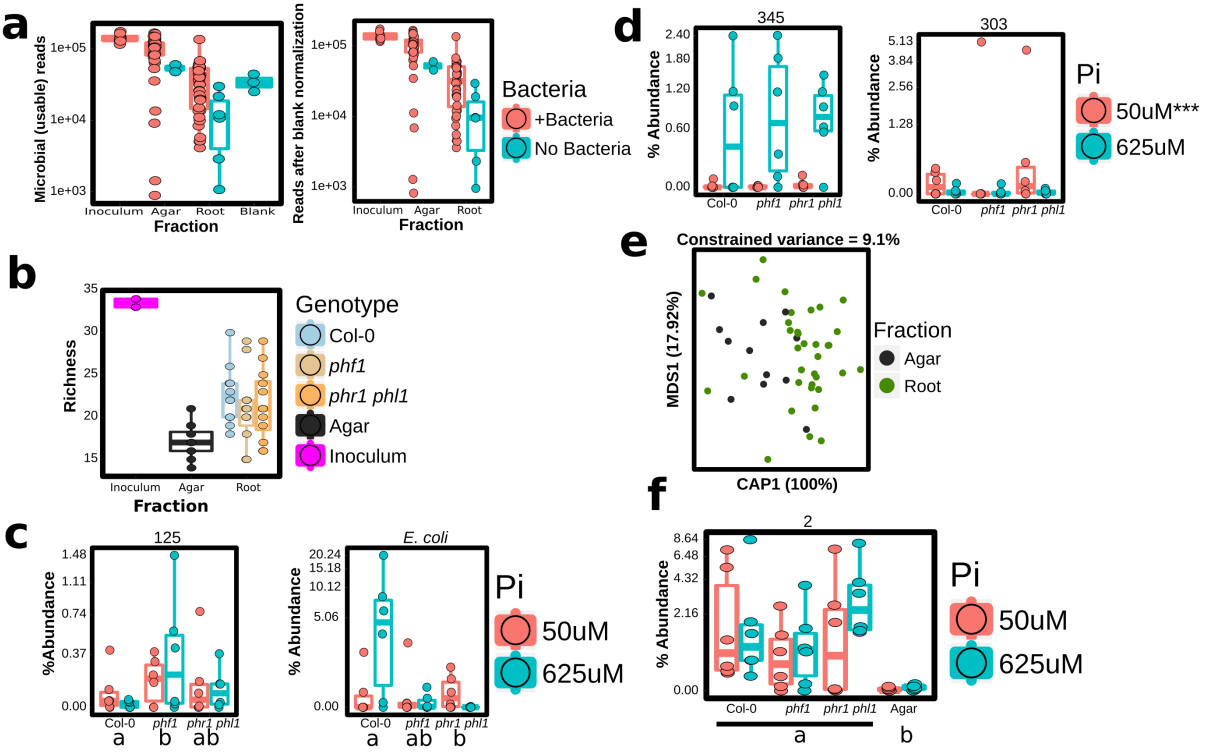


Figure 5.11: **Plant genotype and Pi concentration alter SynCom strain abundances**. **a**, Number of bacterial reads in samples of different types (left) and number of reads after blank normalization (right, see section 5.7.2). The number of biological replicates are: Inoculum (n=8), Agar + SynCom (n=41), Agar No Bacteria (n=2), Root + SynCom (n=36), Root No Bacteria (n=6) and Blank (n = 3), across two independent experiments. **b**, Richness (number of isolates detected) in SynCom samples. No differences were observed between plant genotypes. The number of biological replicates per group is n=12 except for Inoculum (n=4) and phf1 (n=11). **c**, Exemplary SynCom strains that show quantitative abundance differences between genotypes. Genotypes with the same letter are statistically indistinguishable. **d**, Exemplary SynCom strains that show quantitative abundance differences depending on Pi concentration in the media. Asterisks note statistically significant differences between the two Pi concentrations. **e**, CAP analysis of Agar vs Root difference in SynCom communities. These differences explained 9.1% of the variance. The number of biological replicates per fraction is: Agar (n=12) and Root (n=35), distributed across two independent experiments. **f**, Exemplary SynCom strain that shows a statistically significant differential abundance between Root and Agar samples. Statistically significant differences are defined as FDR < 0.05. For c, d and f the number of biological replicates for every combination of genotype and Pi level is always n=6, evenly distributed across two independent experiments.

transferred to 0, 10, 30, 50 and 625 μM Pi alone, or to 0, 50 and 625 μM Pi with the Synthetic Community at 105 c.f.u / mL, for another 12 d. At this point, we harvested our time zero (3 replica per conditions, each replica was 5 shoots harvested across all plates used). The remaining plants were transferred again to 1 mM Pi to evaluate the capacity of the plants for Pi accumulation in a time series analysis. We harvested plant shoots every 24 h for 3 days and Pi-concentration was determined. Pi increase was calculated as:

$$\frac{Pi_i - Pi_0}{Pi_0} \tag{5.1}$$

where Pi_i is the Pi concentration on the i-th day.

Relative increase in Pi concentration is plotted in Fig. 5.3c. Both relative and absolute Pi concentration values are provided in Supplementary Table 4.

We repeated this experiment twice. For the first experiment, we used 6 plates with 10 plants per condition (48 plates and 480 plants in total). We harvested three replicas per time point with 5 shoots each. In all cases, shoots were harvested across all plates used. For the second experiment, we used 11 plates with 10 plants per condition (88 plates and 880 plants). In this case, we harvested 6 replicas for 1, 2 and 3 days after the re-feeding with Pi, and three replicas for time zero. Each replica contains 5 shoots harvested across all the plates used.

For the demonstration that sucrose is required for the induction of PSR in sterile conditions, plants overexpressing the PSR reporter construct IPS1:GUS13 were grown in Johnson medium containing 1 mM Pi or 5 μM Pi supplemented with different concentrations of sucrose. After 12 days, the expression of the reporter constructs IPS1:GUS, highly induced by low Pi, was followed by GUS staining. Plants were grown in a growth chamber in a 15-h light/9-h dark regime (21°C day /18°C night).

For the ChIP-seq experiment, *phr1* harboring the PromPHR1:PHR1-MYC construct (Puga et al., 2014) and Col-0 seedlings were grown on Johnson medium 1 mM Pi, 1% sucrose for 7 days and then transferred to a media not supplemented with Pi for another 5 days. Plants were grown in a growth chamber in a 15-h light/9-h dark regime (21°C day /18°C

night). A total of 2364 genes were identified as regulated by PHR1. The ChIP-seq data will be fully presented in de Lorenzo and Paz-Ares (2017).

For the transcriptional analysis under conditions typically used to study PSR (axenic growth with sucrose present; no microbiota involved), with Methyl Jasmonate (MeJA) and the 22-amino acid flagellin peptide (flg22), plants were germinated on Johnson medium (1% sucrose) containing 1 mM Pi for 7 d in a vertical position and then transferred to 1 mM Pi and 5 μM Pi media containing 1% sucrose either alone or supplemented with 10 μM MeJA (Sigma) or 1 μM flg22 (Sigma) for 12 d.

For growth inhibition assays, seedlings were grown on Johnson medium (1% Sucrose) in 1 mM Pi and 5 μM Pi conditions for 5 d, transferred to 1 mM Pi and 5 μM Pi media supplemented or not with 10 μM MeJA for 5 d. Main root length was then measured using ImageJ software (Barboriak et al., 2005).

5.7.4 Bacterial isolation and culture

For Synthetic Community experiments, we selected 35 diverse bacterial strains. 32 of them were isolated from roots of Arabidopsis and other Brassicaceae species grown in two wild soils (Lundberg et al., 2012). Two strains came from Mason Farm unplanted soil (Lundberg et al., 2012), and Escherichia coli DH5α was included as a control (Supplementary Table 2). More than half (19/35) of the strains belonged to families enriched in the EC of plants grown in Mason Farm soil (Supplementary Table 2) (Lundberg et al., 2012; Lebeis et al., 2015). The strains were chosen from a larger isolate collection in a way that maximizes SynCom diversity while retaining enough differences in their 16S rRNA gene to allow for easy and unambiguous identification.

A single colony of bacteria to be tested was inoculated in 4 mL of 2xYT medium (16 g/L Tryptone, 10 g/L Yeast Extract, 5 g/L NaCL, $\sim$5.5 mM Pi) in a test tube. Bacterial cultures were grown while shaking at 28°C overnight. At this point, the Pi concentration was reduced to by dilution to 5 mM Pi average in the supernatants (10 cultures used for the quantification). Cultures were then rinsed with a sterile solution of 10 mM $MgCl_2$ followed by

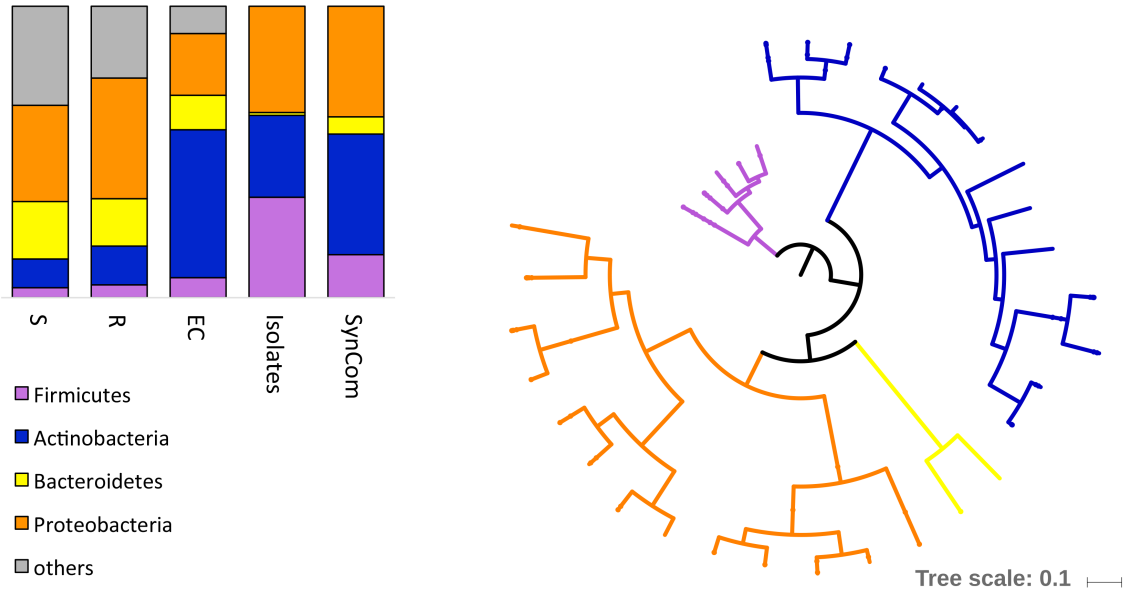


Figure 5.12: **Phylogenetic composition of the 35-member synthetic community (SynCom)**. Left: Comparison of taxonomic composition of soil (S), rhizosphere (R) and endophyte (EC) communities from (Lundberg et al., 2012), with the taxonomic composition of the isolate collection obtained from the same samples and the SynCom selected from within it and used in this work. Right: Maximum likelihood phylogenetic tree of the 35-member SynCom based on a concatenated alignment of 31 single copy core proteins.

a centrifugation step at 2600 g for 8 min. This process was repeated twice. The concentration of Pi in the supernatant after the first wash with $MgCL_2$ was 0.06 mM Pi and after the second wash it was reduced to 0.005 mM Pi. In the suspension of SynCom member cells in $MgCl_2$, the average concentration of Pi was 0.08 mM. The OD600nm was measured and assuming that 1 OD600nm unit is equal to 10^9 c.f.u/mL we equalized individual bacterium concentration to a final value of 10^5 c.f.u/mL of medium. The concentration of Pi in the final SynCom was 0.09 μM Pi. Thus, based on these results, we were not Pi fertilizing the plant by adding the SynCom. Medium was cooled down (to 40-44°C) near the solidification point and then the bacteria mix was added to the medium with agitation. We monitored the pH in the media after adding 1, 5, 10 mL of 10 mM $MgCl_2$ which represents almost ten times the volume we used to add the SynCom. After adding $MgCl_2$ the pH in the media remained stable. We also analyzed the pH after adding the SynCom at 10^5, 10^6 and 10^7 c.f.u/ml of media and found no pH changes. Therefore, we considered that the MES buffer we used was appropriate for this experiment.

To isolate and quantify bacteria from plant roots in the SynCom experiment, plant roots were harvested, and rinsed 3 times with sterile distilled water to remove agar particles and weakly associated microbes. Plant material was then freeze-dried. Root pulverization and DNA extraction was conducted as described above.

To isolate and quantify bacteria from agar samples, a freeze and squeeze protocol was used. Syringes with a square of sterilized miracloth at the bottom were completely packed with agar and kept at -20°C for a week. Samples were thawed at room temperature and syringes were squeezed gently into 50 mL tubes. Samples were centrifuged at max speed for 20 min and most of the supernatant discarded. The remaining 1-2 mL of supernatant, containing the pellet, was moved into clean microfuge tubes. Samples were centrifuged again, supernatant was removed, and pellets were used for DNA extraction. DNA extraction was performed using 96-well format MoBio PowerSoil Kit (MOBIO Laboratories).

5.7.5 Pathology studies

For oomycete pathology studies, *Hyaloperonospora arabidopsidis* (Hpa) isolate Noco2 was propagated on the susceptible Arabidopsis ecotype Col-0. Spores of Hpa were suspended in deionized sterile water at a concentration of 5104 spores/mL. The solution containing spores was spray-inoculated onto 10-d-old seedlings of Arabidopsis grown in fertilized potting soil. Inoculated plants were grown at 21°C under a 9-h light regime. Asexual sporangiophores were counted 5 d post-inoculation on at least 100 cotyledons for each genotype.

For bacterial pathology studies, *Pseudomonas syringae* pv. tomato DC3000 was suspended in 10 mM $MgCl_2$ to a final concentration of 10^5 c.f.u/mL. 35-40-d-old plants of Arabidopsis grown on soil were hand-infiltrated using a needle-less syringe on the abaxial leaf surface. Leaf discs (10 mm diameter) were collected after 1 h and 3 d post inoculation, and bacterial growth was measured as described before (Hubert et al., 2009).

5.7.6 Genome-wide gene expression analyses

We performed 3 different sets of RNA-seq experiments in this study. (I) The first set (Figs. 5.3b, 5.4 and 5.13b) evaluated the effect of the SynCom on the phosphate starvation response of Arabidopsis seedlings. In addition to wild-type Col-0 (4 replicates), *phf1* (4 replicates) and *phr1 phl1* (4 replicates) were included in the experiment shown in Fig. 5.3b, whereas Col-0 (10 replicates), *phr1* (10 replicates) and *phr1 phl1* (6 replicates) were used in the experiment shown in Figs. 5.4 and 5.13b. (II) The second experiment (Fig. 5.12a-b) is an expansion of the first and was designed to evaluate whether different pre-treatments (1 mM Pi, 5 μM Pi, 1 mM Phosphite [Phi]) influence the phosphate starvation response triggered by the SynCom. We used Col-0 (4 replicates), *phf1* (4 replicates) and *phr1 phl1* (4 replicates) in this experiment. (III) Finally, the third experiment evaluated the effect of MeJA and flg22 on the phosphate starvation response (Fig. 5.5 and 5.10) of Arabidopsis seedlings. The genotypes Col-0 (6 replicates) and *phr1 phl1* (6 replicates) were used. The experiments listed above were repeated between two and five independent times and each repetition (defined as "batch" in the generalized linear model, see RNA-seq data analysis,

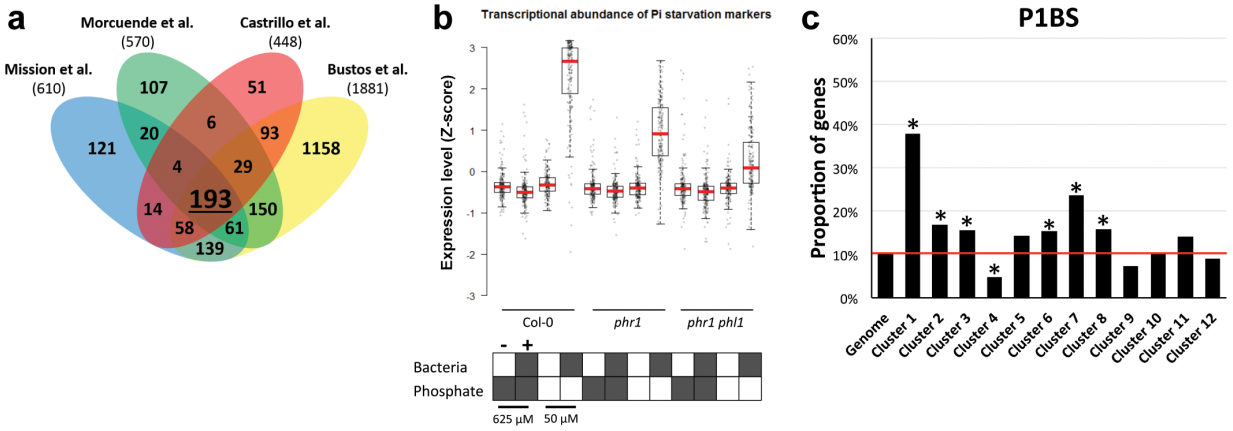


Figure 5.13: **Induction of the PSR triggered by the SynCom is mediated by PHR1 activity**. **a**, Venn diagram with the overlap among genes found up-regulated during phosphate starvation in four different gene expression experiments (Misson et al., 2005; Morcuende et al., 2007; Bustos et al., 2010; Castrillo et al., 2013). The intersection (193 genes) was used as a robust core set of PSR for the analysis of our transcriptional data (Supplementary Table 3). **b**, Expression profile of the 193 core PSR genes indicating that the SynCom triggers phosphate starvation under Low Pi conditions in a manner that depends on PHR1 activity. The RPKM expression values of these genes were z-score transformed and used to generate box and whiskers plots that show the distribution of the expression values of this gene set. Col-0, the single mutant *phr1* and the double mutant *phr1 phl1* were germinated at 1 mM Pi with sucrose and then transferred to low Pi (50 μM) and high Pi (625 μM Pi) alone or with the SynCom. The figure shows the average measurement of ten biological replicates for Col-0 and *phr1* and six for *phr1 phl1*. c, Percentage of genes per cluster (from Fig. 5.4) containing the PHR1 binding site (P1BS, GNATATNC) within 1000 bp of their promoters. The red line indicates the percentage of Arabidopsis genes in the whole genome that contain the analyzed feature. Asterisk denotes significant enrichment or depletion (p-value ≤ 0.05; hypergeometric test).

below) included two biological replicates per genotype per condition. Supplementary Table 15 contains the metadata information of all RNA-seq experiments. Raw reads and read counts are available at the NCBI Gene Expression Omnibus under accession number GSE87339.

5.7.7 RNA isolation and RNA-seq library construction

Total RNA was extracted from roots of Arabidopsis according to Logemann et al. (1987). Frozen seedlings were pulverized in liquid nitrogen. Samples were homogenized in 400 μl of Z6-buffer; 8 M guanidinium-HCl, 20 mM MES, 20 mM EDTA pH 7.0. Following the addition of 400 μl phenol:chloroform:isoamylalcohol; 25:24:1, samples were vortexed and centrifuged (20000 g, 10 min) for phase separation. The aqueous phase was transferred to a new 1.5 ml

tube and 0.05 volumes of 1N acetic acid and 0.7 volumes 96% ethanol were added. The RNA was precipitated at -20°C overnight. Following centrifugation, (20000 g, 10 min, 4°C) the pellet was washed with 200 μl sodium-acetate (pH 5.2) and 70% ethanol. The RNA was dried, and dissolved in 30 μl of ultrapure water and stored at -80°C until use.

Illumina-based mRNA-seq libraries were prepared from 1000 ng RNA. Briefly, mRNA was purified from total RNA using Sera-mag oligo(dT) magnetic beads (GE Healthcare Life Sciences) and then fragmented in the presence of divalent cations (Mg2+) at 94°C for 6 min. The resulting fragmented mRNA was used for first-strand cDNA synthesis using random hexamers and reverse transcriptase, followed by second strand cDNA synthesis using DNA Polymerase I and RNAseH. Double-stranded cDNA was end-repaired using T4 DNA polymerase, T4 polynucleotide kinase and Klenow polymerase. The DNA fragments were then adenylated using Klenow exo-polymerase to allow the ligation of Illumina Truseq HT adapters (D501D508 and D701D712). All enzymes were purchased from Enzymatics. Following library preparation, quality control and quantification were performed using a 2100 Bioanalyzer instrument (Agilent) and the Quant-iT PicoGreen dsDNA Reagent (Invitrogen), respectively. Libraries were sequenced using Illumina HiSeq2500 sequencers to generate 50 bp single-end reads.

5.7.8 RNA-seq data analysis

Initial quality assessment of the Illumina RNA-seq reads was performed using the FASTX-Toolkit. Cutadap (Martin, 2011) was used to identify and discard reads containing the Illumina adapter sequence. The resulting high-quality reads were then mapped against the TAIR10 Arabidopsis reference genome using Tophat (Trapnell et al., 2009), with parameters set to allow only one mismatch and discard any read that mapped to multiple positions in the reference. The Python package HTSeq (Anders et al., 2015) was used to count reads that mapped to each one of the 27,206 nuclear protein-coding genes. Fig 5.14 shows a summary of the uniquely mapped read counts per library. Raw sequencing data and read counts are available at the NCBI Gene Expression Omnibus accession number GSE87339.

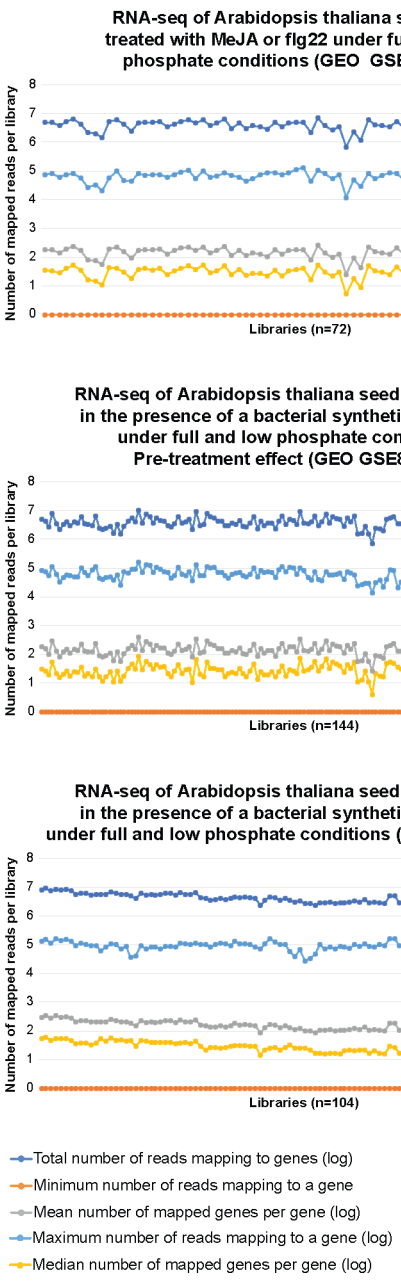


Figure 5.14: **Number of mapped reads for each RNA-seq library used in this study**. The figure shows the maximum, minimum, average and median number of reads mapping per gene for all RNA-seq libraries generated. The total number of reads mapping to genes is also shown for each library. With the exception of the minimum number of mapped reads, which is zero for all libraries, all values are shown in a log scale.

Differential gene expression analyses were performed using the generalized linear model (glm) approach (McCarthy et al., 2012) implemented in the edgeR package (Robinson et al., 2010). This software was specifically developed and optimized to deal with over-dispersed count data, which is produced by RNA-seq. Normalization was performed using the trimmed mean of M-values method (TMM (Robinson et al., 2010); function calcNormFactors in edgeR). The glmFit function was used to fit the counts in a negative binomial generalized linear model with a log link function (McCarthy et al., 2012). For the SynCom experiment (Fig. 5.4), the model includes the covariates: phosphate content (High or Low), bacteria (present or absent) and batch effect. A term for the interaction between Phosphate and Bacteria was included as represented below:

$$Expression = Phosphate + Bacteria + (Phosphate * Bacteria) + Batch \tag{5.2}$$

The model used to analyze the effect of MeJA and flg22 (Fig. 5.5) included the following covariates: phosphate content (High or Low), MeJA (present or absent), flg22 (present or absent) and batch effect.

$$Expression = Phosphate + MeJA + flg22 + Batch \tag{5.3}$$

In each model, the term "Batch" refers to independent repetitions of the experiment (see section 5.7.6). Data from the different genotypes were fitted independently with the same model variables. The Benjamini-Hochberg method (False Discovery Rate; FDR) (Benjamini and Hochberg, 1995) was applied to correct the p-values after performing multiple comparisons. Genes with FDR below or equal to 0.01 and fold-change variation of at least 1.5X were considered differentially expressed.

Transcriptional activation of the phosphate starvation response was studied using a literature-curated set of phosphate starvation marker genes (Fig. 5.13a, Supplementary Table

3). This core set consists of 193 genes that were up-regulated by phosphate starvation stress across four different gene expression experiments (Bustos et al., 2010; Morcuende et al., 2007; Misson et al., 2005; Castrillo et al., 2013). The RPKM (Reads Per Kilobase of transcript per Million mapped reads) expression values of these 193 genes were z-score transformed and used to generate box and whiskers plots to show the distribution of the expression values of this gene set.

Hierarchical clustering analyses were performed with the heatmap.2 function in R from the gplots package (Warnes et al., 2016) using the sets of differentially expressed genes identified in each experiment. Genes were clustered based on the Euclidean distance and with the complete-linkage method. Genes belonging to each cluster were submitted to Gene Ontology (GO) enrichment analyses on the PlantGSEA platform (Yi et al., 2013) in order to identify over-represented biological processes.

5.7.9 Defining markers of the MeJA and SA responses

Genes whose transcription is induced by MeJA (672 genes), BTH/SA (2096 genes) or both hormones (261 genes) were used as markers of the activation of these immune response output sectors in Arabidopsis (Supplementary Table 10) (Yang et al., 2017). These gene sets were defined using two-week old Col-0 seedlings grown on potting soil and sprayed with MeJA (50 μM; Sigma), BTH (300 μM; Actigard 50WG) or a mock solution (0.02% Silwet, 0.1% ethanol). Samples were harvested 1 h, 5 h and 8 h after the treatment in two independent experiments. Total RNA was extracted with the RNeasy Plant Mini kit (Qiagen) and then used to prepare Illumina mRNA-seq libraries. The bioinformatics pipeline to generate count tables and the criteria used to define differentially expressed genes between conditions (Hormone treatment vs. Mock treatment) was the same as described above. Raw sequencing data are available at the NCBI Gene Expression Omnibus under the accession number GSE90077.

5.7.10 Statistical analyses

Most statistical analyses were performed in the R statistical environment (R Core Team, 2014) and follow methods previously described (Lebeis et al., 2015). As described in the following subsections, a number of packages were used, and many were called through AMOR-

0.0-14 (Herrera Paredes, 2016), which is based on code from Lebeis et al. (2015). All scripts and knitr (Xie, 2016) output from R scripts are available upon request. Most plots are ggplot2 (Wickham, 2009) objects generated with functions in AMOR (Herrera Paredes, 2016). For all linear modeling analyses (ANOVA, ZINB, GLM), terms for batch and biological replicate were included whenever appropriate. Code for both census and SynCom analysis is available at `https://github.com/surh/pbi`.

For wild soil and SynCom experiments, the number of samples per genotype and treatment was determined based on our previously published work, which showed that seven and five samples are enough to detect differences in wild soils and SynCom experiments, respectively (Lundberg et al., 2012; Lebeis et al., 2015). For RNA-seq experiments, we used at least four replicates per condition, which is sufficient for parameter estimation with the edgeR software (Robinson et al., 2010).

Alpha and beta diversity were calculated on count tables that were rarefied to 1000 reads. Samples with less than this number of usable reads (i.e. high quality non-organellar reads) were discarded. Alpha diversity (Shannon index, richness) metrics were calculated using the "diversity" function in vegan (Oksanen et al., 2016), and differences between groups were tested with ANOVA (Fig. 5.6a). Site diversity (Fig. 5.6b) was calculated with the "sitediv" function in AMOR (Herrera Paredes, 2016). Unconstrained ordination was performed with vegan (Bray-Curtis), and Principal Coordinate Analysis (PCoA) was performed with AMOR (Fig. 5.6d) (Herrera Paredes, 2016). Canonical Analysis of Principal Coordinates (CAP) is a form of constrained ordination (Anderson and Willis, 2003) and was performed using the "capscale" function of the vegan package in R (Oksanen et al., 2016). CAP was performed on the full counts of the EC samples only, using the "Cao" distance. Constraining was done separately on plant genotype while conditioning on sequencing depth and biological replicate. This approach allowed us to focus on the portion of variation that is associated with plant genotype, conditionally, independent of other factors.

For the SynCom experiments, richness was directly calculated in R. Principal Coordinate

Analysis was performed with the "PCO" function of AMOR (Herrera Paredes, 2016) using the "Cao" distance which was calculated with vegan (Oksanen et al., 2016) on an abundance table rarefied to 1500 reads per sample. Canonical Analysis of Principal Coordinates (CAP) was performed using the "capscale" function of the vegan package (Oksanen et al., 2016) in R. CAP was performed on the full counts of the root samples only, using the "Cao" distance. Constraining was done separately on Fraction, Pi level and plant genotype while conditioning on sequencing depth and the other covariates.

Differentially abundant bacterial taxa across fraction and genotype in the wild soil experiments were identified using the same approach as in Lebeis et al. (2015). Briefly, we used a Zero-Inflated Negative Binomial (ZINB) framework that allowed us to test for the effect of specific variables, while both controlling for the other covariates and accounting for the excess of zero entries in the abundance tables. These zero-entries likely represented under-sampling and not true absences. The same analysis was performed at the family and OTU-level on the measurable OTUs (taxa that have an abundance of at least 25 counts in at least five samples) (Lundberg et al., 2012). Results are in Fig. 5.6e-h and Supplementary Table 1. Fig. 5.6h shows the distribution of significant genotypic effects on bacterial abundances at both taxonomic levels; in both cases the behavior is similar, indicating small and even effects of all genotypes.

For the comparison of enrichment profiles between genotypes, we followed the same Monte-Carlo approach described in Lebeis et al. (2015). Briefly we looked at the enrichment/depletion profile of bacterial taxa for each mutant compared to wild-type Col-0, and asked, for each pair of mutants, if they were more similar than expected by chance and assed significance by random permutation. Results are in Fig. 5.2d and 5.6g.

To define differentially abundant strains in SynCom experiments, we found that a Negative Binomial GLM approach gave more stable results than the ZINB approach. We used the edgeR package (Robinson et al., 2010) to fit a quasi Negative Binomial GLM model with the glmQLFit function, and significance was tested with the glmQLFtest function (Lun et al.,

2016). Results of all relevant pairwise comparisons are in Fig. 5.11 and Supplementary Table 5.

For the definition of robust colonizers in synthetic community experiments, we calculated the average relative abundance of *E. coli* on all root samples and counted, for each strain, how many times it was more abundant than *E. coli*s average on the same set of root samples. Then we used a one-sided binomial test to ask if the probability of a given strain to be more abundant than the average *E.coli* was significantly higher than a coin toss (50%). Strains that passed the test were labeled as robust-colonizers, the rest of the strains were labeled as Sporadic or Non-Colonizers. The results are indicated in Fig. 5.3e and Supplementary Table 2.

5.7.11 Data and software accessibility

All data generated from this project is publicly available. Raw sequences from soil census and SynCom colonization are available at the EBI Sequence Read Archive under accession PRJEB15671. Count tables, metadata, taxonomic annotations and OTU representative sequences from the Mason Farm census and Syncom experiments are available as Supplementary Datasets 1 and Supplementary Datasets 2 respectively. Custom scripts used for statistical analysis and plotting are available at (`https://github.com/surh/pbi`). Raw sequences from transcriptomic experiments are available at the NCBI Gene Expression Omnibus under the accession number GSE87339. The corresponding metadata information is provided in Supplementary Table 15. All code is available upon request.

CHAPTER 6

Bacterial consortia predictably modulate plant phenotypes[1]

Microbes can alter phenotypes in their hosts. Long standing evidence for this fact exists mainly in the form of numerous reports of plant growth-promoting bacteria (Glick, 2012). Evidence for analogous effects of bacteria on animal hosts is more recent (Goodrich et al., 2014; Geva-Zatorsky et al., 2017). Despite the large number of plant growth-promoting bacteria identified in laboratory conditions, the vast majority have failed to produce robust effects in wild or agricultural settings (Bulgarelli et al., 2013). This indicates that typical binary association assays (with just one type of microbe and one type of plant) performed in the laboratory fail to capture critical aspects of more complex systems.

An interesting possibility that has been raised recently is the use of microbial consortia, as opposed to single strains, to produce more robust changes in host phenotypes. Microbiome transplant experiments in mammals (Smith et al., 2013) and inoculation with defined consortia in plants (Bai et al., 2015) have shown that such consortia can produce robust changes in their hosts. However, it remains unknown whether the observed changes are the product of

[1]The contents of this chapter has not been peer-reviewed. It is the draft of a manuscript that will be co-first authored by myself (Sur Herrera Paredes), Dr. Gabriel Castrillo from Jeff Dangl's lab, and PhD student Tianxiang Gao from Vladimir Jojic's group. Other members of the Dangl lab also made significant contributions and will be recognized with authorship in the manuscript. Including but not limited to Terry Law. For this chapter, the specific contributions of different people are as follow: GC, SHP, TG and JD designed the experiments. GC and TL set up the experiments, collected and processed the samples for *in vitro* growth curves and binary associations. SHP analyzed the data from the *in vitro* growth curves and binary associations, designed bacterial blocks and the first synthetic communities. GC and TL set up the synthetic community association experiments, collected phenotypic data and processed samples for transcriptomics. SHP analyzed the phenotypic and transcriptiomic data. TG and VJ designed and implemented the neural network and generated candidate block swaps for validation. GC and TL set up the validation experiments and collected phenotypic data. TG analyzed the validation phenotypes. SHP, GC, TG and JD analyzed data and designed figures. SHP, GC, TG and JD wrote the manuscript with input from TL.

stacking many alternative strains with the same effect, and thus maximizing the chance that at least one will work; or if the changes in host phenotypes are due to emergent properties from the simultaneous presence of a microbial consortia and their host.

Here we systematically evaluate the performance of bacterial *in vitro* screening, and plant-bacteria binary association assays in their ability to predict the function of a bacterial consortia in a complex bacterial background. We take advantage of a large collection of root isolates from Brassicaceae species to design well-defined but complex and partially overlapping bacterial synthetic communities. Our design allowed us to directly estimate the contributions of different bacterial sets to a number of plant phenotypes, and to make predictions about *never-seen-before* bacterial communities, thus establishing the ability to make causal inferences directly on bacterial consortia.

Overall, we found that binary association assays, but not *in vitro* bacterial screenings, can inform the design of bacterial consortia. We observed that most bacterial communities led to a similar overall activation of defense, but a dramatically different activation of the phosphate starvation response, as well as specific differences in the response to jasmonic acid and auxin. Our results indicate that the effect of bacterial consortia can be explained mostly by *stacking* of redundant bacterial functions. However, we observed the emergence of unexpected outcomes in a few instances, and we show that statistical methods capable of capturing such nuances are better at predicting novel communities, highlighting the importance of systematic exploration and the development of appropriate analytical frameworks. Our synthetic community design approach is a blueprint for screening bacterial collections and designing multi-strain cocktails that maximize the chance of success in a more complex setting. Despite being reductionist in essence, our approach is flexible and powerful enough to dissect complex plant-microbiota interactions.

6.1 *In vitro* isolate screening

We focus on the phosphate starvation response in Arabidopsis. After nitrogen, phosphorus is the second most important plant macronutrient (Vitousek et al., 2010). Although

phosphorous is relatively abundant in soils, plants can only absorb inorganic orthophosphate (hereafter phosphate) which is limited in soil. Plants respond to low phosphate with a stress response that is a combination of developmental, physiological and molecular adaptations. Some of those adaptations involve the exudation of enzymes such as phytases and organic acids (Narang et al., 2000) which acidify the rhizosphere,increasing phosphate solubilization and making it available to the plant. Many bacterial isolates can solubilize phosphate, which could potentially help the plants; however the inoculation of plants with phosphate solubilizing bacteria has found limited success (Leggett et al., 2010; Sharma et al., 2013). A potential complication is that the plant transcription factor PHR1, the master regulator of the phosphate starvation response, is also a negative regulator of defense (Castrillo et al., 2017).

We hypothesized that if plants recruit bacteria to help cope with phosphate starvation, they might do so via molecular cues in the root exudates, and thus we would expect that bacteria responding to those cues coul be identified via their *in vitro* growth patterns in the presence of root exudates. We collected root exudates from *A. thaliana* seedlings that had been grown in two phosphate conditions (section 6.7.2). We then performed *in vitro* growth curves of ∼600 individual bacterial strains, isolated from roots of Brassicaceae plants grown in two previously characterized soils (Lundberg et al., 2012), in those two exudates, as well as in the two phosphate conditions without plant exudate (section 6.7.3). We found a variety of bacterial behaviors in response to root exudates and phosphate concentrations, and as expected, those behaviors showed a strong phylogenetic signal (Fig. 6.1).

The patterns in the bacterial growth-curves, and the metabolomic profiles from the exudates (Fig. 6.13) demonstrate that plants release a different molecular set in different phosphate starvation conditions, and that the concentrations of those exudates are sufficient to influence bacterial growth.

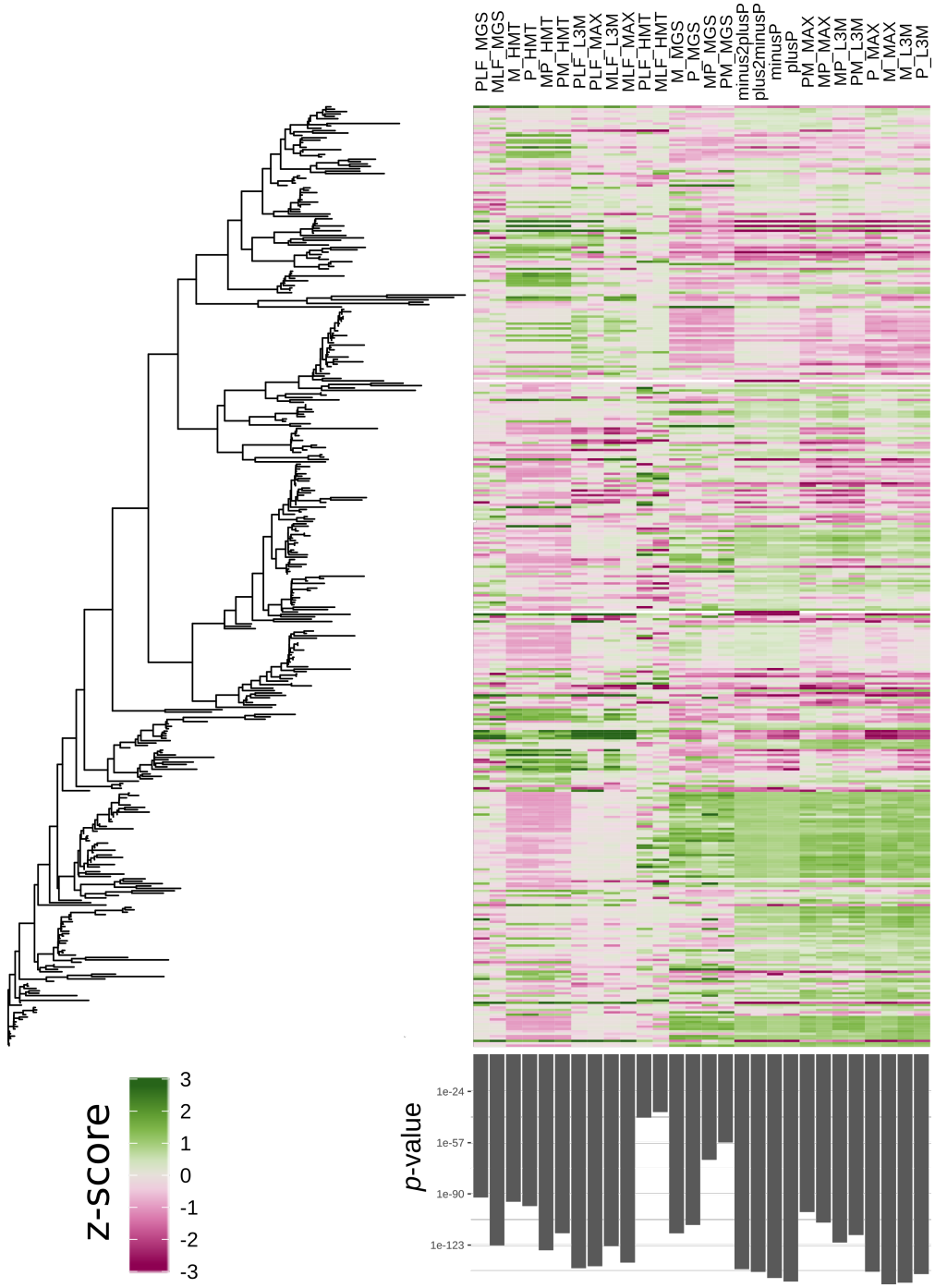


Figure 6.1: **Phylogenetic signal in bacterial growth curves**. **Left**: Phylogenetic tree of 395 bacterial strains. **Right**: Heatmap of growth curve features. Feature names are at the top and have the following meanings: minus2plusP, plus2minus, minusP and plusP are the area under the growth curve for each of the *in vitro* conditions. For the rest, the prefix indicates the conditions as follows: minusP (M), plusP (P), minus2PlusP (MP), plus2MinusP (PM), $\log_2$(MP/P) (PLF) and $\log_2$(PM/M) (MLF). Suffix indicates the measurement: maximum density (MAX), mean density over last 3 measurements (L3M), mean time to reach half maximum density (HMT), and the maximum growth rate (MGS). **Bottom**: p-values from Pagel's λ test for phylogenetic signal.

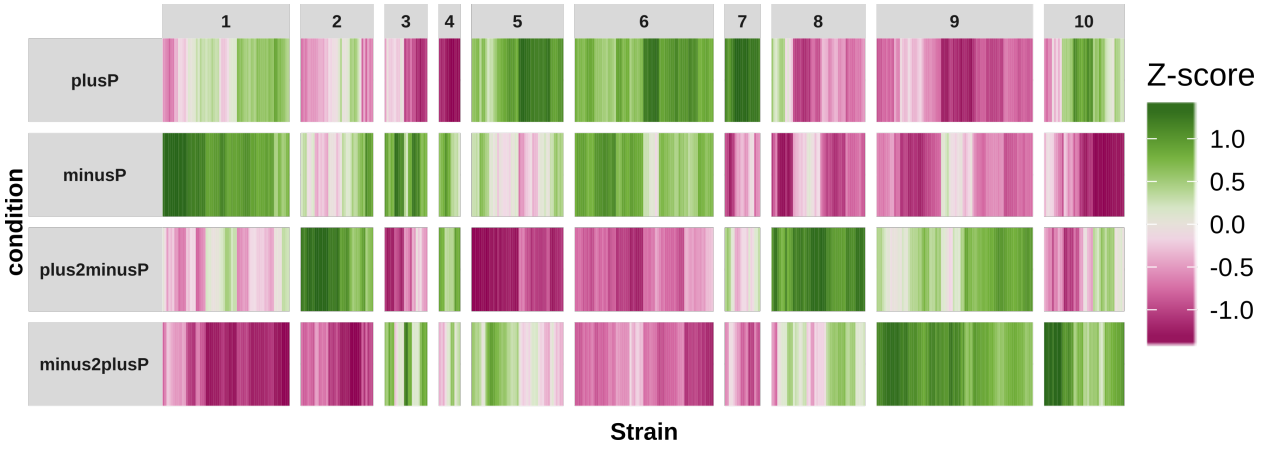


Figure 6.2: **Bacterial classification according to in vitro performance**. Heatmap showing the log-transformed and standardized media area under the curve for 440 strains that were grown in four *in vitro* conditions. The top two are Johnson media with or without 1mM phosphate added. The bottom two conditions are Johnson media with exudates from plants that where transferred between media containing, or not containing, 1mM phosphate. Strains were grouped by hierarchical clustering, using the Euclidean distance and the complete linkage method.

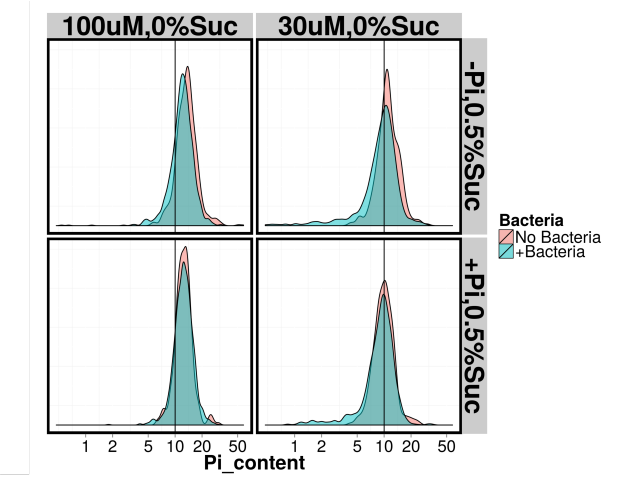
6.2 Individual strains modulate plant phosphate accumulation

We used the area under the curve (AUC) as an aggregate measure of bacterial performance in the different conditions, and we used hierarchical clustering to classify the bacteria into ten groups that represent different classes of response to root exudates and phosphate concentrations (section 6.7.4). In order to determine if the bacterial *in vitro* response to root exudates is indicative of function for the plant, we selected the most responsive isolates from each of the groups (section 6.7.4). We then tested the change in plant shoot phosphate accumulation in response to the presence of ∼180 individual strains when compared with plants grown axenically. We evaluated this behavior in four phosphate starvation conditions that represent a *two-by-two* combination matrix of two phosphate level pretreatments prior to bacterial inoculation (*pre-treatment*), and two phosphate levels that were applied concomitant with each bacteria (*post-treatment*, section 6.7.6).

Overall, we found that most bacteria have a slightly negative effect on plant shoot phosphate accumulation (Figs. 6.3 and 6.4a), probably because of competition for the nutrient. This effect was stronger when the phosphate concentration was lower (Fig. 6.3 right

vs left column), consistent with our previous finding that a bacterial synthetic community drives a context-dependent competition with the plant for Pi (Castrillo et al., 2017). Contrary to our expectations, we found no correlation between the effect of individual isolates on shoot phosphate accumulation, and their performance *in vitro* (Fig. 6.4b). We also found a very small phylogenetic signature (Fig. 6.4a), suggesting that the ability of individual strains to modulate plant phosphate levels can be caused by multiple mechanisms and/or is an evolutionarily *flexible* trait. Overall, bacterial strains were more likely to have a stronger negative effect on plant shoot phosphate accumulation in the more limiting phosphate condition (30μM Pi post-treatment), consistent with previous results (Fig. 6.3 bottom table) (Castrillo et al., 2017). Conversely, individual strains were more likely to have a positive effect on shoot phosphate accumulation in the more phosphate-rich condition (Fig. 6.3 bottom table). Importantly, the effect of individual strains on plant phosphate accumulation was independent of bacterial titers from different plant organs, and colonization did not require an intact phosphate starvation response (Fig. 6.14). The scale of our survey of plant-bacteria binary associations, and its resulting distribution of bacterial effects on plant phosphate accumulation argues that the majority of the plant-bacteria interaction are competitive, at least in the context of phosphate starvation.

We have previously shown that a 35-member bacterial synthetic community can lead to PHR1-dependent activation of the phosphate starvation response in Arabidopsis, and that PHR1 negatively regulates plant immunity (Castrillo et al., 2017). We therefore asked whether activation of the Arabidopsis phosphate starvation response is required for bacterial modulation of shoot phosphate accumulation. We found that inactivation of the Arabidopsis PSR, via the non-metabolizable phosphate analogue phosphite, dramatically reduced the effect of bacteria on plant phosphate accumulation, for both positive and negative modulators (Fig. 6.5). This suggests that plant signaling is required by bacteria to activate the modulation of plant phosphate accumulation in either direction, and thus, that the negative effect of bacteria is not merely due to a bacterial autonomous response to low phosphate, but to a



	minusP_100uM	minusP_30uM	plusP_100uM	plusP_30uM
increase	3	4	17**	9
decrease	39	66**	6	39
total	42	70**	23	48

**p-value < 0.01

Figure 6.3: **Bacterial effect on shoot phosphate accumulation**. Top: Distribution shoot phosphate concentrations in plants co-incubated with individual bacterial strains (+Bacteria) or in axenic condition (No Bacteria), in the four phosphate conditions. Bottom: Number of strains that significantly increase or reduce plant shoot phosphate accumulation with respect to no bacteria. Asterisks indicate a significantly higher number of strains with an effect than expected (hypergeometric test).

perception by the plant and activation of its phosphate starvation response.

In summary, we performed a large-scale survey for bacterial-induced plant phosphate accumulation in binary associations. We found a majority of competitive interactions. The ability of bacteria to modulate plant phosphate accumulation is mostly independent of bacterial phylogeny and performance *in vitro*, but is dependent on the plant phosphate starvation response.

6.3 Bacterial blocks act additively on plant phosphate accumulation

We found no correlation between *in vitro* bacterial assays, and the bacterial ability to modulate plant phosphate accumulation in binary associations (Fig. 6.4). This prompted us to ask whether the results from those binary associations are indicative of the bacterial effects when a more complex bacterial community is present. We decided to use a microcosm reconstitution system, in which we inoculate plants with complex but well-defined bacterial synthetic communities (Lebeis et al., 2015; Castrillo et al., 2017). We chose a subset of 78

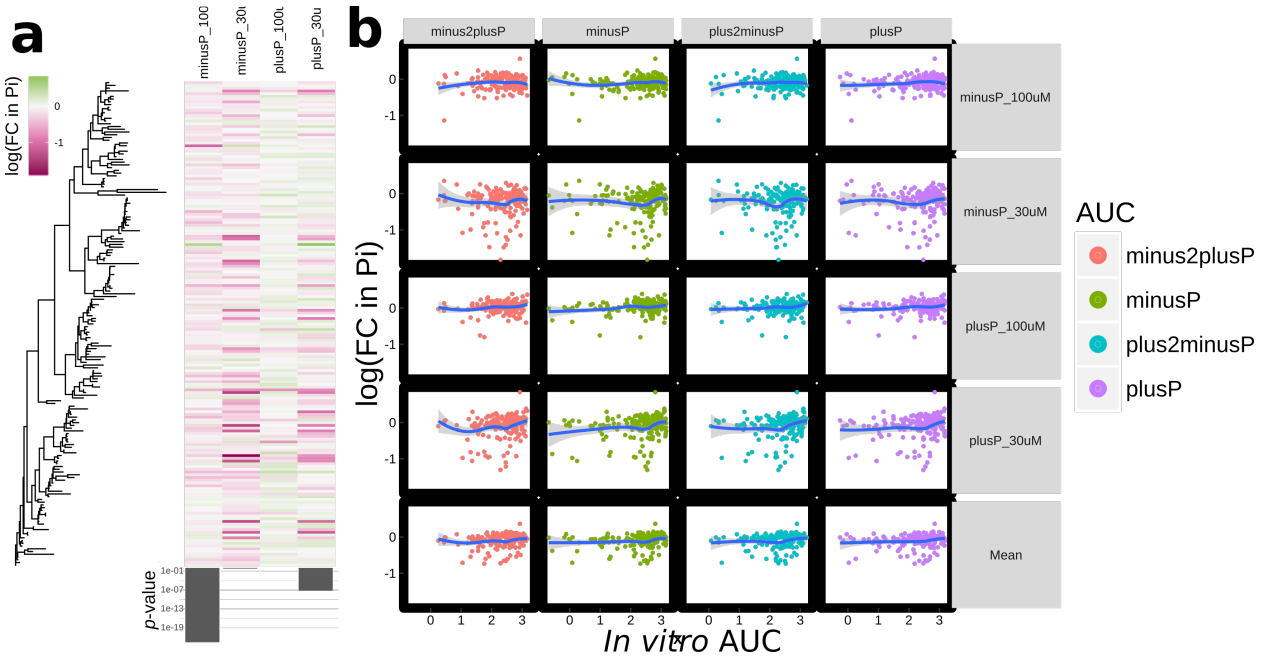


Figure 6.4: **Bacterial modulation of plant phosphate accumulation is independent of bacterial phylogeny and *in vitro* performance**. **a** Heatmap of log fold-change in shoot phosphate accumulation, between plants inoculated with individual bacterial strains and axenically grown seedlings. Bacteria are sorted according to their phylogeny as indicated by the tree on the left. Bottom bar plot shows the p-value from Pagel's λ test for phylogenetic signal. **b** Scatter plots showing the area under the curve (AUC) for bacterial growth curves in four media conditions (x-axis) and the change in phosphate accumulation due to bacteria from a, as well as the mean log fold-change in phosphate accumulation across all four conditions. Dots are color coded by their in vitro growth condition, and the blue line shows the LOESS smoother.

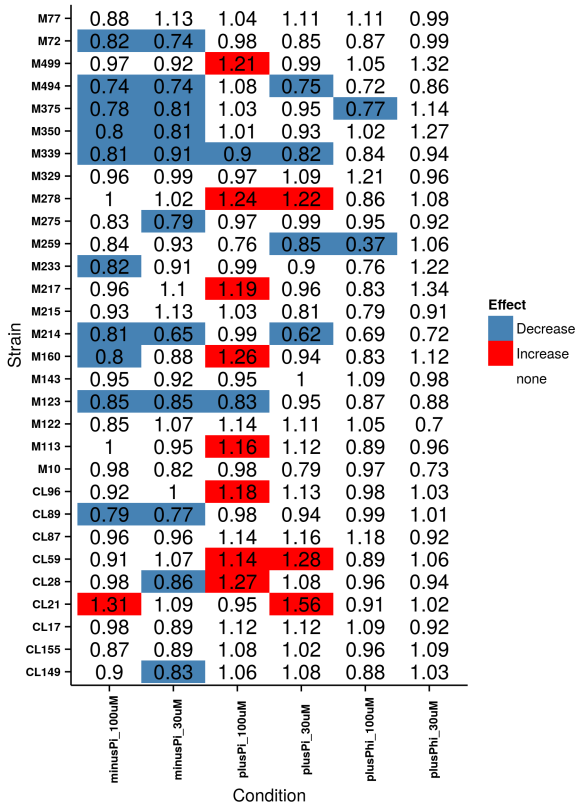


Figure 6.5: **Activation of the plant phosphate starvation response is required for bacterial modulation of plant phosphate accumulation**. Table showing the fold-change in shoot phosphate accumulation between plants inoculated with an individual strain, and plants grown axenically. Six phosphate conditions were used (section 6.7.6. Cells with a color block indicate statistically significant changes from no bacteria (q-value < 0.05; ANOVA and Tukey test). The last two columns are from plants pre-treated with phosphite which inhibits the activation of the phosphate starvation response.

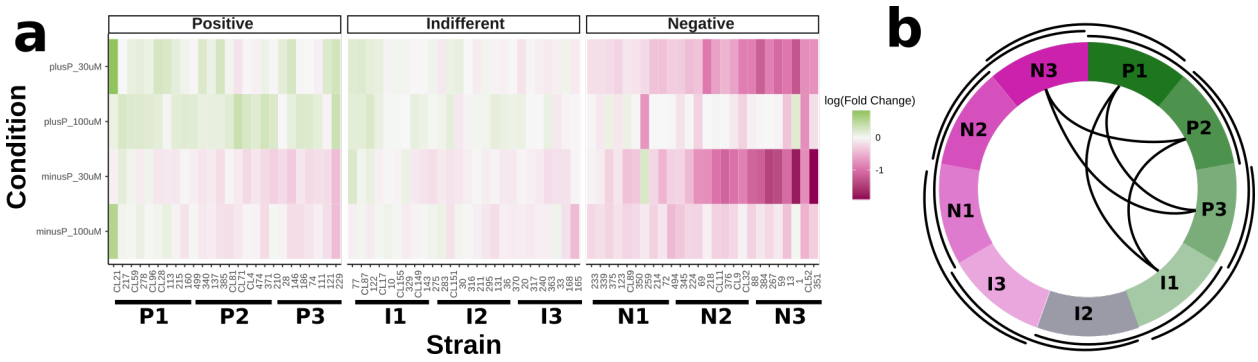


Figure 6.6: **Designing synthetic communities from binary association data**. **a** Heatmap of 78 strains tested in binary association that have positive, negative or indifferent effects on plant phosphate accumulation. Strains are sorted within each group according to their mean effect on phosphate accumulation. Color scale shows log fold-change in shoot phosphate accumulation with respect to axenically grown plants. Bars and labels at the bottom show the nine bacterial blocks used in the following experiments. **b** Fourteen synthetic communities constructed from pairs of blocks. Sections in the circle are the bacterial blocks from a, and black curved segments represent communities. Outer segments represent communities made of adjacent blocks, and curves inside the circle represent communities between non-adjacent bacterial blocks.

strains from those tested in binary association (section 6.7.9) and grouped them into nine blocks of 8-9 strains, each according to their effect on shoot phosphate accumulation (Fig. 6.6a; section 6.7.9). We then designed 14 partially overlapping synthetic communities by combining pairs of blocks (Fig. 6.6b). We selected those communities to maximize the chance to observe extreme plant phenotypes and to obtain the most information from the most extreme blocks (section 6.7.9).

We then performed community-associations with Arabidopsis by growing seedlings in the same conditions as before (section 6.7.7), but inoculated with each of the 14 synthetic communities (Fig. 6.6) instead of with individual strains. Besides shoot phosphate accumulation, we also measured main root elongation, shoot size and total root network, which are phenotypes that have been extensively studied in the context of plant phosphate starvation in axenic conditions (REFS).

We observed that the synthetic communities affected plant phenotypes in ways that were mostly consistent with the effect of the individual strains that compose them. Most synthetic

communities had a negative effect on plant phosphate accumulation, compared with axenic controls. This effect was stronger in communities made mostly of negative (N) blocks, and in the most nutritionally limiting conditions (Fig. 6.7a). Most communities also decreased plant main root elongation (Fig. 6.7b). This observation is not unexpected since this phenotype is a known marker for MAMP-triggered immunity (Ranf et al., 2011). Moreover, main root elongation correlated strongly with phosphate content, with more phosphate starved plants showing reduced main root elongation (Fig. 6.7a-b). This is consistent with a common strategy followed by Arabidopsis plants, since the topsoil is usually more phosphate-rich (Lambers et al., 2015). Surprisingly, a number of communities led to increased rosette size with respect to axenically grown plants, despite the reduced shoot phosphate accumulation. This effect was stronger in the less nutritionally challenging conditions (Fig. 6.7c). This indicates that these communities can independently modulate development and the phosphate starvation response. Finally, synthetic communities had more variable effects in the plant total root network, with many causing an increase (Fig. 6.7d). However, in the condition where bacteria had the strongest negative effect on shoot phosphate accumulation (i.e. post-treatment of 30μM phosphate, right panels), the greatest increase in root network was also observed, in particular by synthetic communities containing negative (N) blocks. This is consistent with the increase of lateral roots that is a hallmark of the phosphate starvation response in Arabidopsis (Lambers et al., 2015). Overall, plants inoculated with synthetic communities from the negative (N)-blocks (*i.e.* those made of strains that decrease shoot phosphate accumulation) had phenotypes consistent with an activated phosphate starvation response. Thus the binary association assays are informative with regards to the behavior of bacteria in a more complex biotic background.

A key question is whether the effect of bacteria on plant phenotypes is consistent across different microbial backgrounds, or whether it is context dependent. Our experimental design places each block in at least two bacterial backgrounds, and so it is possible to estimate the common (additive) effect of each block across bacterial backgrounds, and to

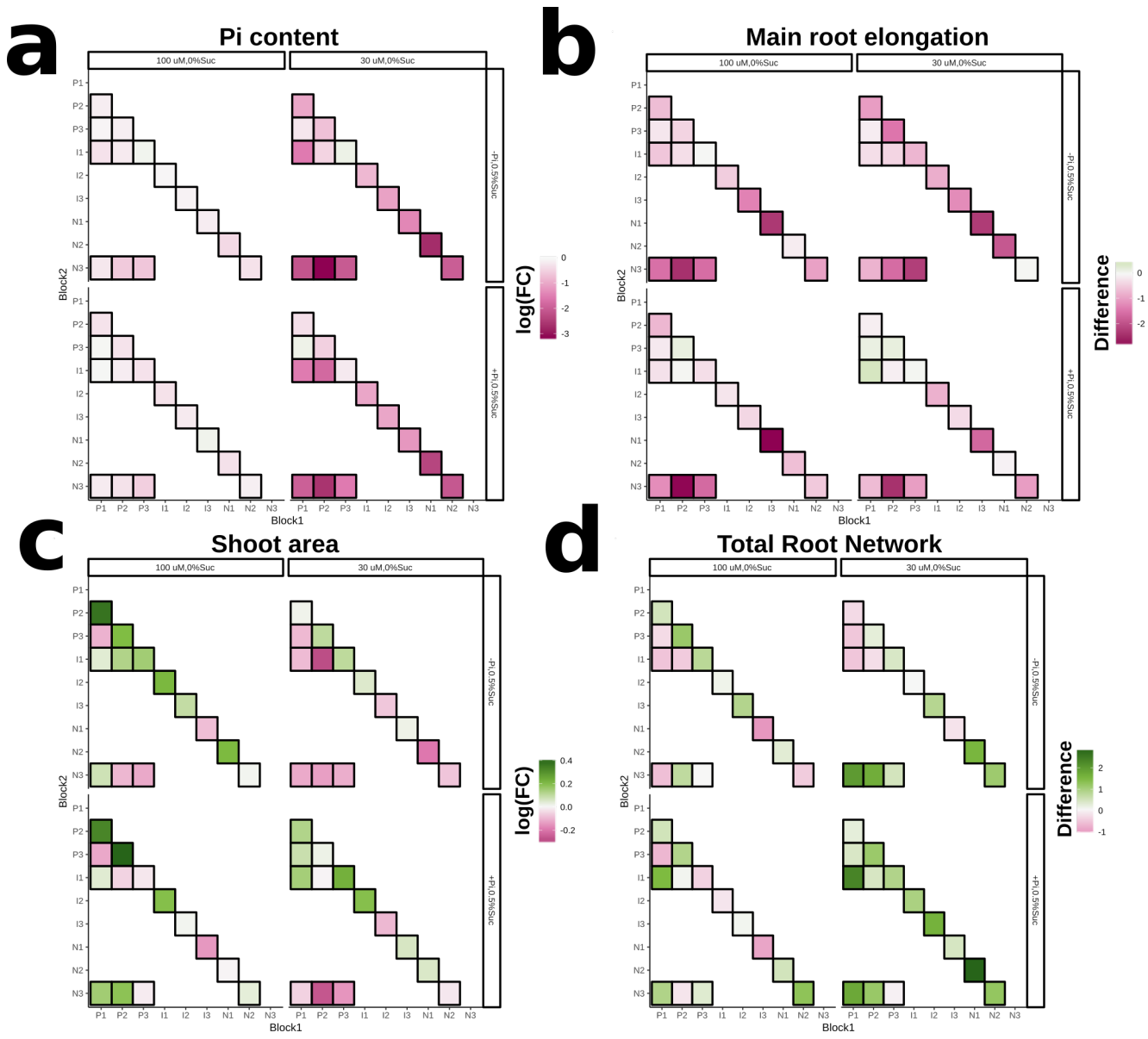


Figure 6.7: **Synthetic communities alter plant phenotypes**. Change in plant phenotypes induced by synthetic communities compared with axenically grown seedlings. In each plot, the four panels represent the four media conditions tested with pre-treatment as rows and post-treatment as columns. X- and Y-axes show the 9 bacterial blocks and the lower triangle cells in each panel show the phenotype change induced by a synthetic community composed of the two blocks indicated by its X and Y coordinates. In all plots zero (white) represents no change in the corresponding phenotype with respect to axenically grown plants, and the color scale indicates more (green) or less (magenta) than axenically grown plants. The phenotypes shown are: shoot phosphate accumulation (**a**), main root elongation (**b**), shoot area (**c**) and total root network (**d**). The values for Pi-content and shoot area (a and c) indicate log fold-change with respect to axenically grown plants. The values for main root elongation and total root network (b and d) represent difference with respect to axenically grown plants. Combinations of blocks tested have a black outline.

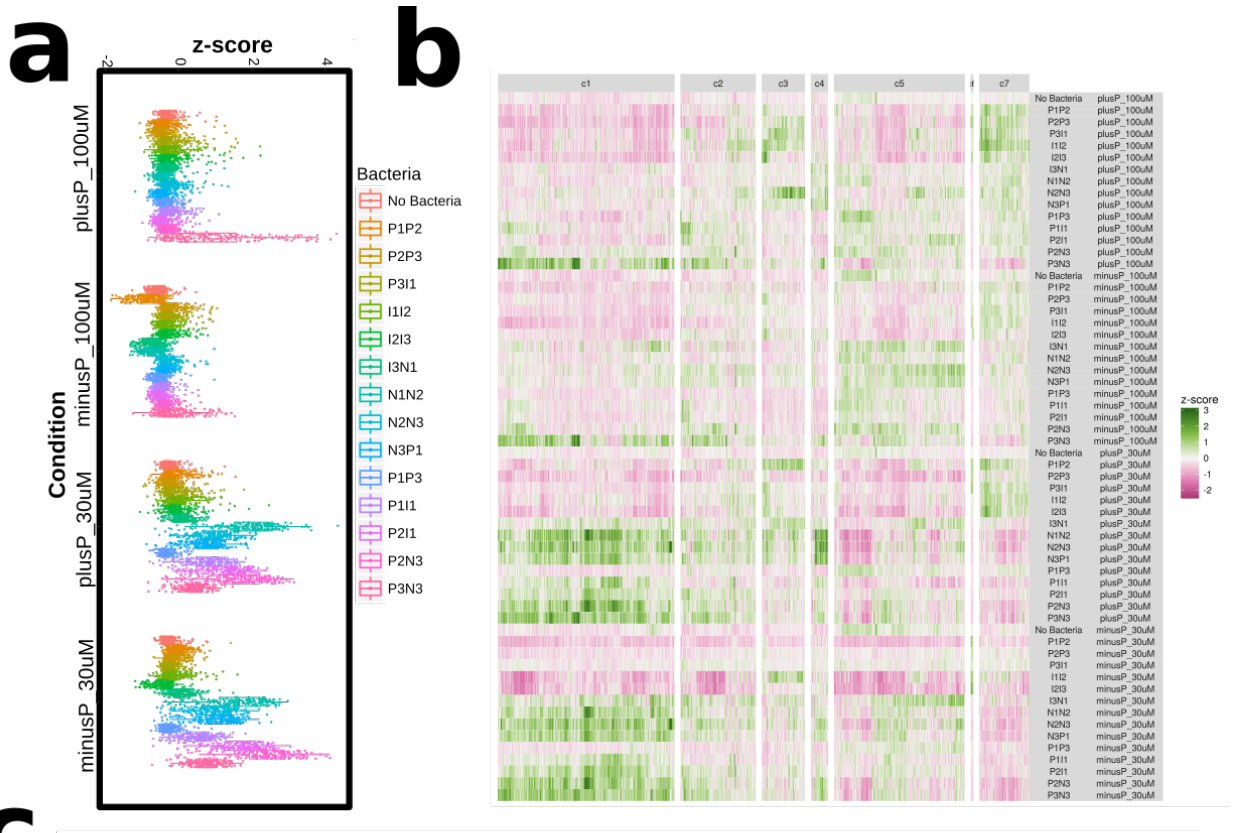
ask if those common effects are sufficient to explain the phenotypic variation (section 6.7.12). Surprisingly, we found that for all phenotypes, the additive contributions of the bacterial blocks is sufficient to explain most of the plant phenotypic variation (Fig. 6.15), suggesting that intra-block bacterial interactions on these plant phenotypes are at least as strong as inter-block interactions.

Together with our observation that synthetic communities behave in line with the expectations derived from binary-association (Fig. 6.10). The striking sufficiency of additive contributions to explain the effect of synthetic communities indicates that, at least in the context of phosphate starvation, the knowledge obtained in binary-association experiments is partially transferrable to a more complex environment where a bacterial community is present. Our results also indicate that while bacterial abundance may be highly dynamic, and bacteria-bacteria direct interactions are probably important for microbiome assembly, the effect of the microbiota on host phenotypes is robust to variation in bacterial composition.

6.4 Bacterial modulation of plant transcriptional responses

We previously defined a core set of phosphate starvation markers that are transcriptionally up-regulated upon phosphate starvation in Arabidopsis (*core-Pi*), and we found that, in the absence of sucrose, the presence of a diverse synthetic community of 35 strains led to teir activation (Castrillo et al., 2017). We asked whether our 14 synthetic communities were also capable of rescuing the induction of the phosphate starvation response. In line with our previous results, we did not observe activation of the phosphate starvation response marker genes in the absence of bacteria, even in the most extreme phosphate deficiency conditions (Fig. 6.8a; top set of points on each condition). Furthermore, in the most phosphate-depleted condition we observed induction of the phosphate starvation response marker genes by some communities only (Fig. 6.8a; bottom two conditions). The observed specificity of the plant PSR transcriptional induction suggests that a bacterial biological activity is responsable.

We observed that a large number of genes follow a pattern of expression similar to that of the *core-Pi* set (Fig. 6.8b; cluster c1). Gene ontology enrichment analysis revealed that this



Cluster c1

GO Name	Total	Found	%	FDR
DEFENSE_RESPONSE_BY_CALLOSE_DEPOSITION	62	50	81	0.000000003
REGULATION_OF_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY	32	25	78	0.000237
D-XYLOSE_METABOLIC_PROCESS	35	25	71	0.00071
GALACTOLIPID_METABOLIC_PROCESS	100	69	69	1.24E-10
GALACTOLIPID_BIOSYNTHETIC_PROCESS	99	68	69	2.04E-10
COUMARIN_METABOLIC_PROCESS	55	38	69	0.00000878
COUMARIN_BIOSYNTHETIC_PROCESS	54	37	69	0.0000145
ZINC_ION_HOMEOSTASIS	24	16	67	0.038
GLYCOLIPID_METABOLIC_PROCESS	115	75	65	1E-10
POLYSACCHARIDE_LOCALIZATION	94	61	65	1.23E-08
CALLOSE_LOCALIZATION	93	60	65	2.04E-08
REGULATION_OF_ANION_TRANSPORT	40	26	65	0.00149
GLYCOLIPID_BIOSYNTHETIC_PROCESS	106	68	64	1.75E-09
POSITIVE_REGULATION_OF_FLAVONOID_BIOSYNTHETIC_PROCESS	103	66	64	3.38E-09
ER-NUCLEUS_SIGNALING_PATHWAY	192	120	63	2.44E-16
RESPIRATORY_BURST_INVOLVED_IN_DEFENSE_RESPONSE	121	76	63	2.58E-10
AMINO_ACID_IMPORT	75	47	63	0.00000282
CELLULAR_RESPONSE_TO_BIOTIC_STIMULUS	195	120	62	5.73E-16
ENDOPLASMIC_RETICULUM_UNFOLDED_PROTEIN_RESPONSE	184	114	62	2.69E-15
CELLULAR_RESPONSE_TO_IRON_ION_STARVATION	116	72	62	1.39E-09
NUCLEOTIDE-SUGAR_METABOLIC_PROCESS	39	24	62	0.00555
AMMONIUM_TRANSPORT	29	18	62	0.0339
BASIC_AMINO_ACID_TRANSPORT	29	18	62	0.0339
CELLULAR_RESPONSE_TO_PROTEIN_STIMULUS	192	118	61	1.17E-15
RESPONSE_TO_UNFOLDED_PROTEIN	186	114	61	4.74E-15
CELLULAR_RESPONSE_TO_UNFOLDED_PROTEIN	186	114	61	4.74E-15
CELLULAR_RESPONSE_TO_PHOSPHATE_STARVATION	139	85	61	4.2E-11
NEGATIVE_REGULATION_OF_PROGRAMMED_CELL_DEATH	170	102	60	5.3E-13
ER_TO_GOLGI_VESICLE-MEDIATED_TRANSPORT	105	63	60	6.12E-08

Cluster c2

GO Name	Total	Found	%	FDR
SALICYLIC_ACID_METABOLIC_PROCESS	222	59	27	0.000000009
DEFENSE_RESPONSE,_INCOMPATIBLE_INTERACTION	535	126	24	8.84E-16
DEFENSE_RESPONSE_TO_FUNGUS	342	78	23	4.82E-09
NEGATIVE_REGULATION_OF_DEFENSE_RESPONSE	273	64	23	9.61E-08
RESPONSE_TO_BACTERIUM	572	125	22	1.39E-13
SYSTEMIC_ACQUIRED_RESISTANCE	444	98	22	8.68E-11
INNATE_IMMUNE_RESPONSE	926	190	21	1.34E-18
SALICYLIC_ACID_MEDIATED_SIGNALING_PATHWAY	349	72	21	0.000000755
CELLULAR_RESPONSE_TO_SALICYLIC_ACID_STIMULUS	351	72	21	0.000000906
IMMUNE_RESPONSE	980	193	20	2.31E-17
IMMUNE_SYSTEM_PROCESS	980	193	20	2.31E-17
REGULATION_OF_RESPONSE_TO_STRESS	544	101	19	0.000000116
RESPONSE_TO_FUNGUS	497	94	19	0.000000194
REGULATION_OF_DEFENSE_RESPONSE	529	98	19	0.000000208
RESPONSE_TO_SALICYLIC_ACID_STIMULUS	470	87	19	0.00000166
RESPONSE_TO_OTHER_ORGANISM	1411	251	18	5.71E-18
RESPONSE_TO_BIOTIC_STIMULUS	1675	287	17	9.68E-19
DEFENSE_RESPONSE	1644	269	16	2.24E-15

Cluster c4

GO Name	Total	Found	%	FDR
RESPONSE_TO_JASMONIC_ACID_STIMULUS	471	92	20	6.78E-52
RESPONSE_TO_WOUNDING	340	75	22	5.96E-45
JASMONIC_ACID_BIOSYNTHETIC_PROCESS	135	54	40	6.34E-43
JASMONIC_ACID_METABOLIC_PROCESS	159	65	41	2.49E-52

Figure 6.8: **Overall transcriptional response to synthetic communities**. **a** Activation of the phosphate starvation response by specific synthetic communities. Average expression of 193 phosphate starvation response markers (*core-Pi*) in all conditions and with all synthetic communities. **b** Clustering of ∼17000 most variably expressed genes in our experiments. Rows represent the average from all samples with a given bacterial treatment in each condition, and columns represent genes. Genes are clustered according to their expression profiles. **c** Gene ontology enrichments for clusters c1, c2 and c4 from b.

cluster corresponds to a combination of defense and stress response genes, including response to phosphate starvation (Fig. 6.8c). Genes in cluster c3, which are more highly expressed on the 30μM phosphate post-treatment, were enriched by numerous membrane phospholipid metabolic genes, potentially indicating activation of the plant phosphate recycling pathways. Clusters c2 and c3 were mostly implicated in plant immunity. Genes in cluster c2 were more highly expressed in most of bacteria inoculated plants than in no bacteria, and highly enriched for genes involved in defense and salicylic acid signaling (Figs. 6.8b-c. Genes in c3 more highly activated by specific communities and enriched in the jasmonic acid sector of immunity (Fig. 6.8b-c).

We specifically asked which genes respond to different combinations of bacteria, condition and combinations of both (section 6.7.18). Consistent with our clustering analysis, we saw that genes that were more expressed by plants in association with bacteria than axenically grown seedlings, were also associated with genes relating to defense to multiple pathogens, and to salicylic acid, indicating activation of PAMP/MAMP-triggered immunity (Fig. 6.9a). We also asked whether bacterial positive (P) blocks activated a different set of genes than negative (N) blocks. As expected, negative (N) blocks were associated with higher expression of genes annotated as response to abiotic stress, including phosphate starvation and abscisic acid responses (Fig. 6.9b). Moreover, we also found increased expression of jasmonic acid response genes (Fig. 6.9b), consistent with our previous finding that phosphate starvation response down-regulates the salicylic acid sector of immunity (Castrillo et al., 2017). The specificity in the induction of the plant phosphate starvation response by some but not all synthetic communities (Fig. 6.8a) prompted us to ask whether different negative (N) blocks activate the same set of phosphate starvation response genes. We found almost no differences between the two most extreme negative blocs (N2, N3), but we found over 200 genes that were differentially regulated by blocks N1 and N2 (Fig. 6.6) in the lower (3OμM) phosphate conditions. Almost all of those genes, were implicated in defense, and were more highly expressed in the least extreme (N1) block (Fig. 6.9c). This result indicates that all negative

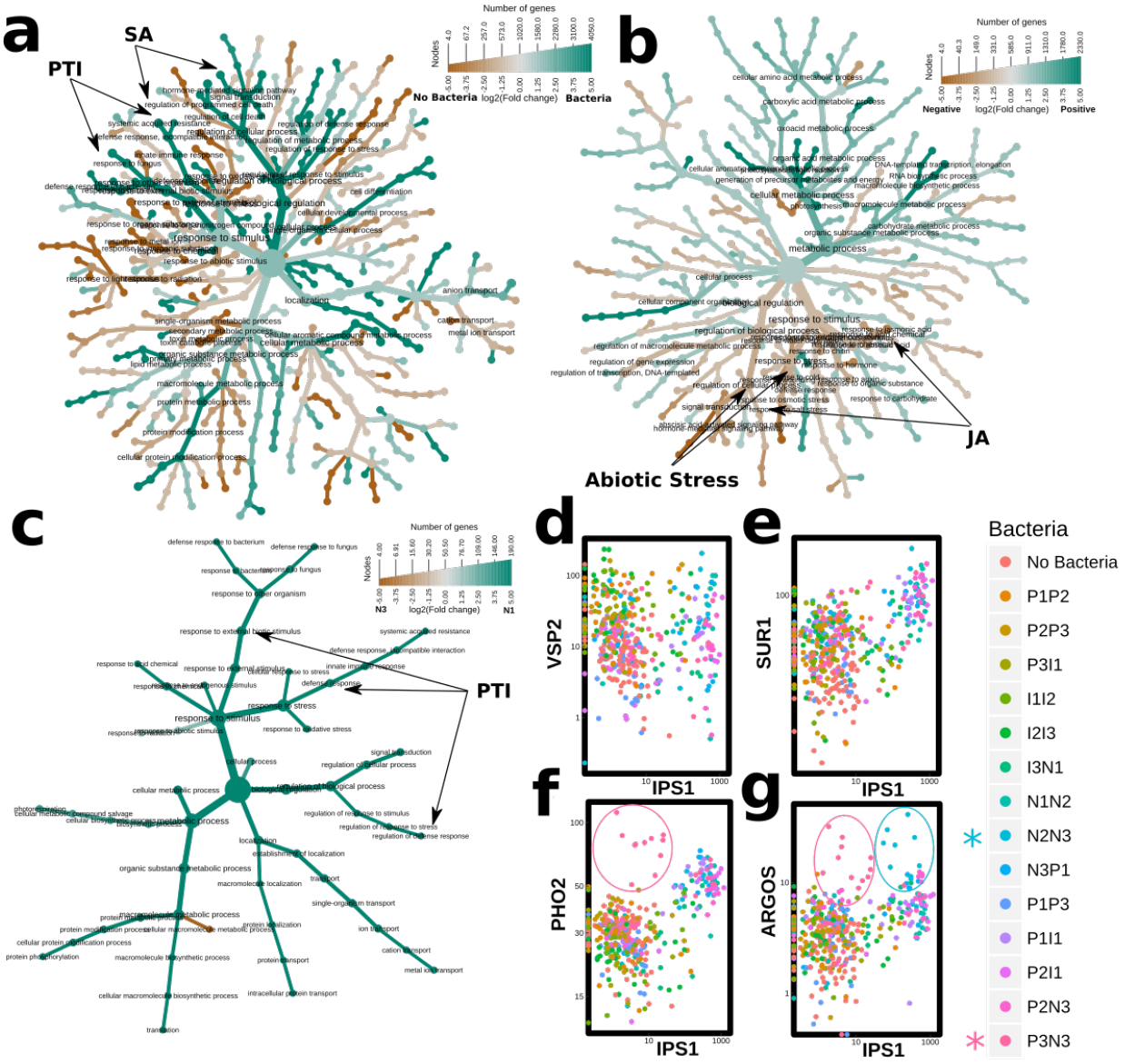


Figure 6.9: **Modulation of the plant transcriptome by bacteria**. **a** Bacteria activate defense. Genes that are differentially expressed in response to bacteria are associated with PAMP/MAMP-triggered immunity (PTI), and the salicylic acid sector of immunity (SA). **b** Negative blocks induce response to abiotic stress and the jasmonic acid (JA) sector of immunity. **c** In the low (30μM) phosphate concentration post-treatment the most negative block (N3) shows more reduced expression of PAMP/MAMP-triggered immunity (PTI). For a-c, the tree represents the relationships between gene ontology annotations of all differentially expressed genes between the groups indicated on the legend. Color scale shows the average $\log_2$ fold-change in expression among genes in each gene ontology term, and the size of each node represents the number of differentially expressed genes in that class. Panels d-e compare the expression of the phosphate starvation response marker IPS1, with jasmonic acid response marker VSP2 (**d**), glucosinolate biosynthesis marker SUR1 (**e**), phosphate transporter ubiquitin-conjugating enzyme PHO2 (**f**), and auxin regulated gene ARGOS (**g**). Expression values are RPKM in a $\log_{10}$ scale. Circles in f-g highlight samples inoculated with communities P3N3 (pink) and N2N3 (blue).

blocks activate the same set of phosphate starvation response genes at similar levels, but that that the most extreme bacterial blocks (N3) lead to a stronger suppression of defense.

Overall, we observed that the plant transcriptomic profiles followed expected general expression patterns based on our binary association assays, with plants treated with bacteria showing variable levesl of defense activation, and plants inoculated with (N) blocks displayin an activated phosphate starvation response in the lower phosphate conditions. However, given the apparent uncoupling of some traditional phosphate starvation phenotypes (*i.e.* increased shoot size even when phosphate accumulation is reduced), we sought to identify the transcriptional signature that underlies these apparent discrepancies. By using well-defined marker genes for different plant responses, we identified that an increase in phosphate starvation response does not lead to a general increase of the jasmonic acid response (Fig. 6.9d), but to the activation of a specific sector involved in glucosinolate biosynthesis (Fig. 6.9e), in line with recent results that implicate this pathway with plant-microbe interactions in the context of phosphate starvation (Hiruma et al., 2016). Interestingly, we found that PHO2, an ezyme responsible for the degradation of phosphate transporters,was more highly expressed when the phosphate starvation response is active, but we also found that synthetic community P3N3 constitutively induces this gene, potentially explaining the strong low phosphate content of these plants, despite the presence of a positive block (Figs. 6.9f 6.10). We also identified the auxin regulated gene ARGOS which has a weak positive correlation with the induction of the phosphate starvation response, but is constitutively induced by synthetic communities P3PN3, and N2N3 (Fig. 6.9). The ARGOS gene is known to control organ size in Arabidopsis, and transgenic expression of this gene results in enlarged aerial organs (Hu, 2003), which could serve to counter balance the negative effect on size that low phosphate typically has and that is weak when either of these two communities are present (Fig. 6.10).

In summary, we showed that synthetic communities activate defense, but that different blocks of bacteria activate different sectors and at different levels, without breaking the balance

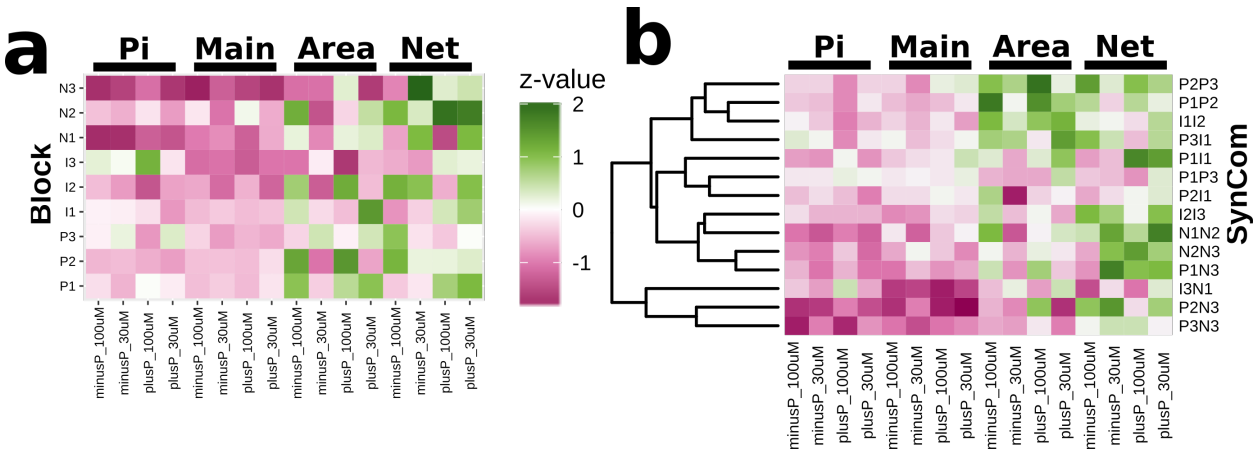


Figure 6.10: **Comparing individual block effects with community effects**. a Shows the scaled effect that each individual bacterial block has on each of the four plant phenotypes tested on each of the four conditions. b Similar to a, but the effect of each individual community is shown. Synthetic communities have been clustered according to similarity of their effects on plant phenotypes. In all cases, the values correspond to the scaled coefficients from a linear model. The values have been scaled by dividing by the standard deviation of all coefficients for the same phenotype and condition (each column in the plots). In all cases, zero (white) represents no change with respect to axenically grown plants.

between jasmonic acid and salicylic acid responses. We showed that the general phosphate recycling from phospholipids is activated by plants in the presence of most communities, but most of the transcriptional response to phosphate starvation is activated only by specific communities. We showed that the transcriptional profiles of synthetic communities are in line with expectations based on binary association assays, and plant phenotypes induced by the synthetic communities, and that we can use the transcriptional profiles to investigate the uncoupling between phosphate starvation response and plant development.

6.5 Designing novel bacterial consortia

Despite the strong explanatory power of the additive contributions of bacterial blocks, there are some limitations. For example, all bacterial blocks are estimated to have a negative effect on root elongation, but several synthetic communities end up with an increased main root elongation with respect to axenically grown plants (Fig. 6.10). This indicates the presence of interacting effects that are not appropriately captured by a simple linear model. Deep learning methods are ideally suited to this type of problem (LeCun et al., 2015; Angermueller

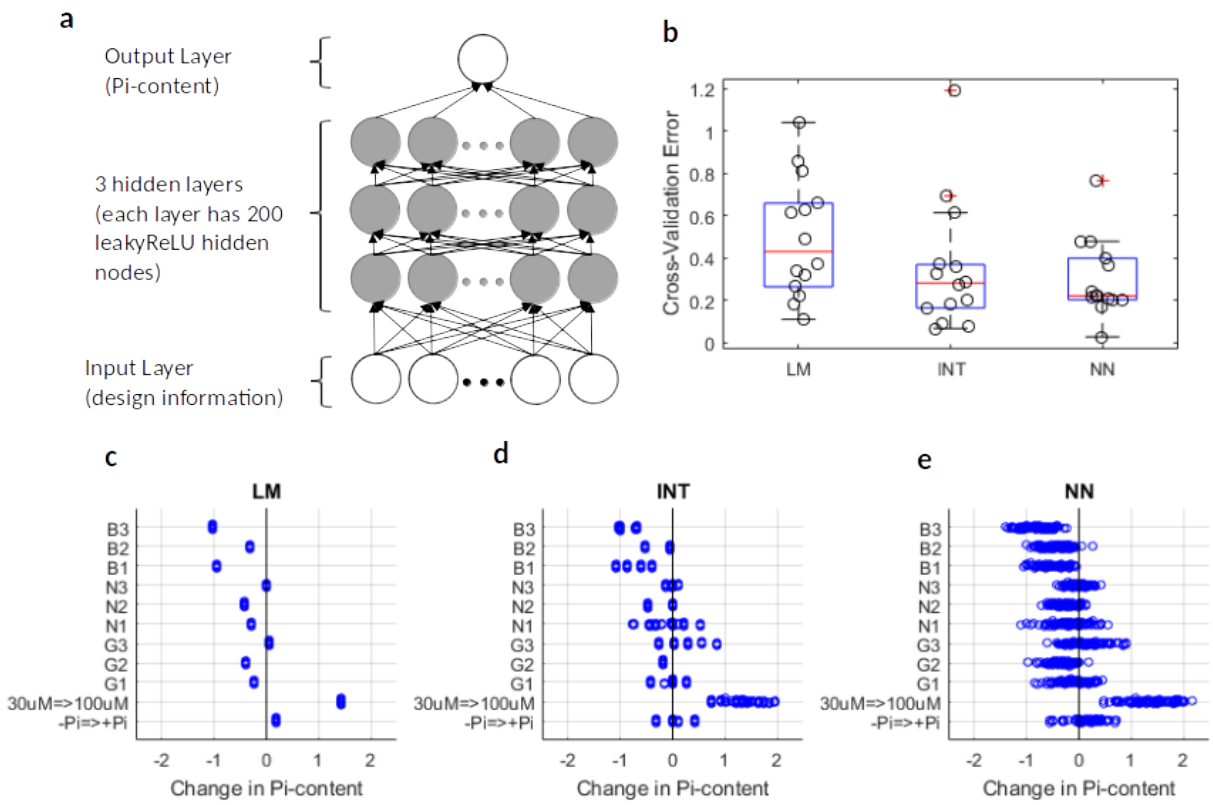


Figure 6.11: **Complex tri-partite interaction captured by a neural network**. **a** Schematic representation of our neural network. **b** Cross-validation error from 3 types of models in their ability to predict plant phosphate accumulation. **c-e** Sensitivity of phosphate accumulation with respect to each biological variable for each type of model. For b-e LM stands for linear model, INT for linear model with interactions, and NN for neural network.

et al., 2016; Min et al., 2016). Therefore, we built a neural network model which is able to capture complex non-linear relationships between the bacterial blocks and abiotic conditions that affect plant phenotypes (Fig. s6.11a; section 6.7.19). This is demonstrated by a lower cross-validation error of the neural network compared to linear models (Fig. 6.11b); section 6.7.19). Sensitivity analysis confirmed that the presence of positive (P) or indifferent (I) blocks; and higher phosphate concentrations in the pre- and post-treatments correlated with increased plant phosphate accumulation (Fig. 6.11c-e; section 6.7.20).

The ultimate test of a predictive model is its ability to predict the behavior of a system in novel circumstances. Therefore, we decided then to test novel synthetic communities that had never been seen by the model or the scientists performing the experiments and analysis.

We designed bacterial block *swaps* that would maximize the increase in shoot phosphate accumulation (Fig. 6.12a; section 6.7.21). We tested the 25 strongest predictions from the neural network in one condition (section 6.7.7). We observed a significant correlation ($\rho = 0.42$) between predicted and observed shoot phosphate accumulation change caused by the block *swap* (Fig. 6.12b), and the neural network had the lowest prediction error (Fig. 6.12c). In total, 23/25 block *swaps* tested had changes in phosphate accumulation in the predicted direction (p-value = $9.7x10^{-6}$; one-sided binomial test; $p = 0.5$). Moreover, 16/25 bacterial *swaps* showed a statistically significant increase in plant phosphate content (p-value = $2.021x10^{-15}$; one-sided binomial test; $p = 0.05$). Only 1/25 bacterial *swap* led to a statistically significant decrease in phosphate accumulation. Therefore, we successfully demonstrated our ability to predict the function of novel synthetic communities on the plant.

6.6 Conclusion

While it is clear that bacteria influence phenotypes of their hosts, it has proven difficult to determine their relevance in complex environments. Here, we leverage a large collection of bacterial root isolates to systematically evaluate how manipulation of the plant microbiome results in changes in plant phenotypes. We showed that we can use knowledge derived from binary association assays to construct synthetic communities that differentially modulate plant phenotypes, and showed that those communities differentially modulate the plant transcriptional immune and phosphate starvation responses. Plant responded to phosphate level and bacteria presence by activating the phosphate starvation and defense responses, as was apparent from developmental phenotypes and transcriptional profiles. However, ther was variation in both responses dependent of the bacteria present, indicating fine-tuning in response to a complex environment. We observed interesting cases where the presence of bacteria produced unexpected uncoupling of plant phenotypes. Transcriptomic analysis revealed potential molecular pathways that explain these results.

We observed high consistency between our expectations based on binary association assays and the results from synthetic community experiments. However, a number of cases could not

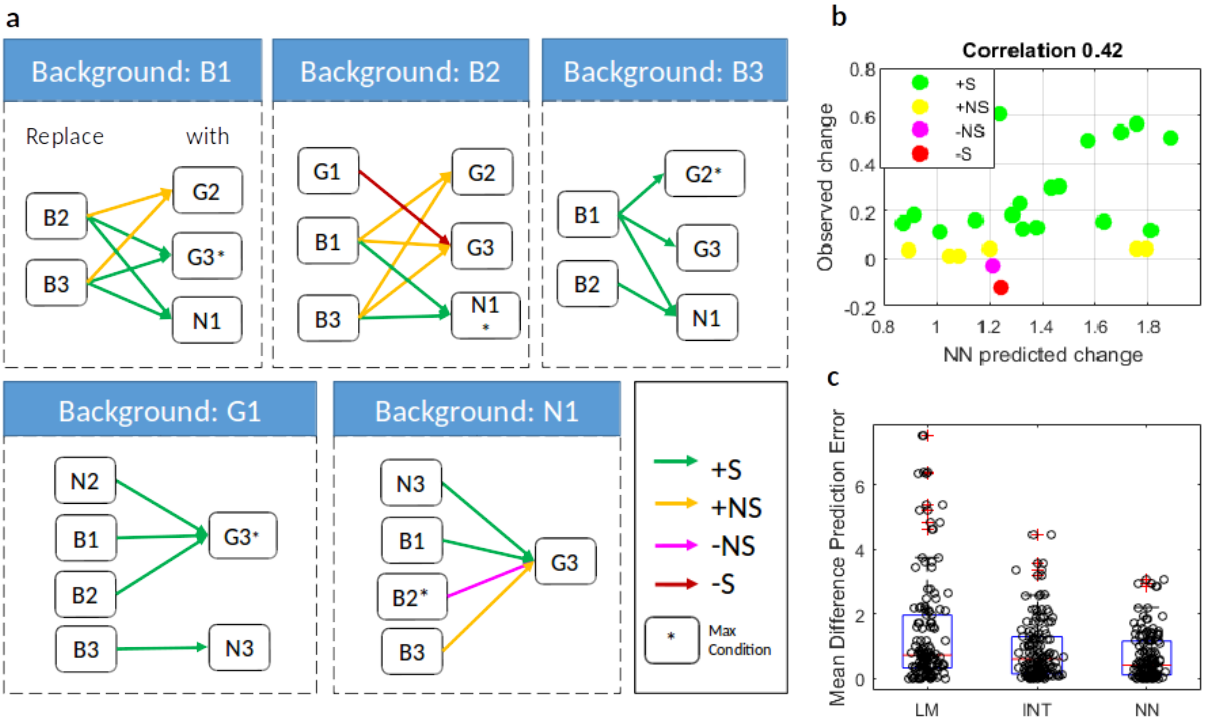


Figure 6.12: **Prediction *never-seen-before* synthetic communities**. **a** most significant 25 hypotheses generated by the neural network. These hypotheses cover 20 synthetic communities. Each box represents the selected *swaps* under a particular background block. Each arrow represents a replacement of the bacterial block on the left for the one on the right. Asterisk indicates the synthetic community that leads to maximal plant phosphate accumulation. **b** correlation between shoot phosphate accumulation change predicted by the neural network (x-axis), and changed obtained experimentally. For a-b color represents the experimental result: significant increase (green), non-significant increase (red), non-significant decrease (pink) and significant decrease (red). **c** prediction error on all tested *swaps* for the linear model (LM), linear model with interaction (INT) and neural network (NN).

be explained by simple analytical methods. We used state of the art deep learning techniques to capture complex relationships among bacteria, plant phenotypes and abiotic conditions. We confirmed our ability to estimate causality by successfully predicting plant phenotypes for *never -before-generated* synthetic communities.

Tri-partite interactions involving the relationship between two types of organisms and their environment are a hallmark of host-microbe systems (Ewald, 1988; Hooper, 2001). However, they have been hard to dissect systematically given the experimental and analytical challenges that they pose. In the context of host-associated microbiome, the standard approaches involve association of microbial features with host phenotypes (Gilbert et al., 2016), and the use of binary association assays to establish causality (Geva-Zatorsky et al., 2017). The first approach is correlative, while the second lacks generality. Exhaustive and simultaneous variation of multiple variables, together with network models, has been used to show that it is possible to dissect complex interactions (Ristova et al., 2016). However, as the number of variables increases this approach becomes impractical very quickly. Our experimental design approach, based on partially overlapping synthetic communities, is able to achieve high accuracy despite exploring only a subset of all possible combination. By design, each block is tested in multiple backgrounds which, together with our validation results, shows that we can attain both causality and generality.

6.7 Methods

6.7.1 Seed sterilization

All seeds were surfaced-sterilized with 70% bleach, 0.2% Tween-20 for 8 minute, and 3 rinses with sterile distilled water. This treatment eliminates any seed-borne microbes on the seed surface. Seeds were stratified at 4°C in the dark for 2 days.

6.7.2 Exudate preparation and profiling

For root exudate preparation, Col-0 seeds were germinated on Johnson medium 0.5% sucrose, solidified with 0.6% agar and supplemented or not with 1 mM Pi, in a horizontal position (approximately 160 plants per plate). After 7 days of growth, seedlings were transferred to a 12-well plate. Each well was filled up with 3 mL of liquid Johnson medium

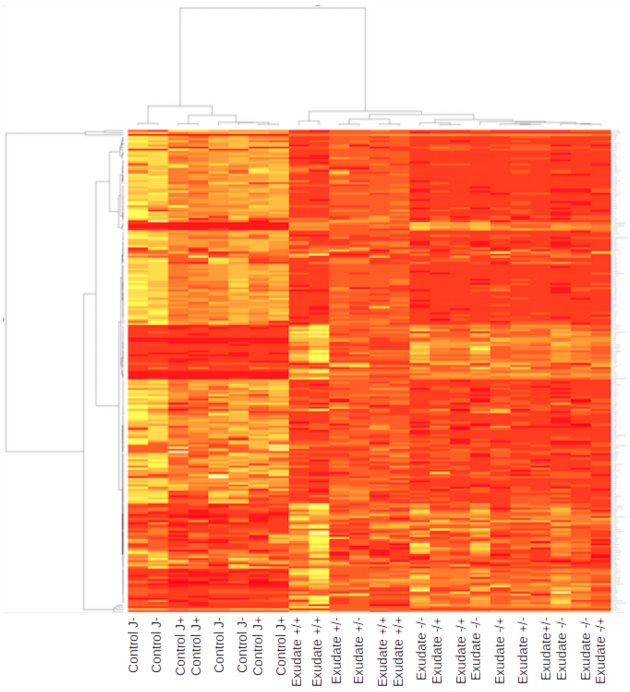


Figure 6.13: **Root exudates primary metabolite analysis**. Heatmap showing primary metabolite analysis of the two Johnson media utilized, and the exudates from the two conditions.

and between 50-60 seedlings. For this experiment, we transferred the seedlings to the opposite concentration of Pi from the solid growth conditions (i.e. plants that were initially grown in 1 mM Pi were transferred to liquid medium with no supplementation of Pi and vice versa). Plants were grown in liquid media with agitation for 24h in a growth chamber in a 16-h light/8-h dark regime (24°C/21°C).

Liquid supernatants, containing root exudates, were collected, filtered (0.22 m) and used for next experiments. Figure 6.13 shows the primary metabolite analysis of the collected exudates.

Primary metabolites profile was performed using ALEX-CIS GCTOF MS in the NIH West Coast Metabolomics Center (University of California, Davis). Plant root exudates and control samples were extracted following Fiehn et al. (2010). 30μL aliquots of each samples were extracted by 1 ml of degassed acetonitrile:isopropanol:water (3:3:2, v/v/v) at -20°C, centrifuged and decanted with subsequent evaporation of the solvent to complete dryness.

A clean-up step with acetonitrile/water (1:1) removed membrane lipids and triglycerides. The cleaned extracts were aliquoted into two equal portions and the supernatants were dried down again. Internal standards C08-C30 FAMEs were added and the samples were derivatized by methoxyamine hydrochloride in pyridine and subsequently by N-methyl-N-trimethylsilyltrifluoroacetamide for trimethylsilylation of acidic protons. Data was acquired using the chromatographic parameters published in Fiehn et al. (2008). A column Restek corporation rtx5Sil-MS (30m length x 0.25mm internal diameter with 0.25μm film made of 95% dimethyl/5%diphenylpolysiloxane was used. Helium was used as mobile phase with a column temperature of 50-330°C and a flow rate of 1 mL min^{-1}. 0.5μL of sample was injected with 25 splitless time into a multi-baffled glass liner at 50°C ramped to 250°C by 12°C s^{-1}.

Mass spectrometry parameters was used as follows: a Leco Pegasus IV mass spectrometer is used with unit mass resolution at 17 spectra s^{-1} from 80-500 Da at -70 eV ionization energy and 1800 V detector voltage with a 230°C transfer line and a 250°C ion source.

6.7.3 Bacterial *in vitro* growth curves

For the screening of the bacteria collection in different plant root exudates, bacteria from -80°C glycerol stocks were grown on LB plates at 28°C. A single colony was then inoculated in 200μL of 2xYT medium (16 g/L Tryptone, 10 g/L Yeast Extract, 5 g/L NaCL, ∼5.5 mM Pi) in a 96 well polystyrene plate (Costar) and covered with a breathable Aeraseal (Excel) to prevent contamination. Bacteria cultures were grown with agitation at 28°C. After 24h, all cultures were diluted 1/10 in the different plant exudates and control conditions and grown at 28°C with agitation. The Optical Density at 600 nm was measured every 3 hours during the day and every 14 hours during the night for 5 days using a microplate reader.

6.7.4 Isolate growth-curve clustering and selection for *in planta* assays

First the growth curves were quality filter by removing strains that had profiles that were highly similar to *blank* samples. All the following operations were done with functions available via the PGCA R package (`https://github.com/surh/PGCA`). For each strain and condition, the median growth curve was obtained by calculating the median OD600nm per time point. The four resulting growth curves (for four conditions) were concatenated and

grouped by hierarchical clustering based on their correlation distance according to the formula $d_{xy} = 1 - \rho_{xy}$, where ρ_{xy} is the Pearson correlation coefficient between strains x and y. The resulting clustering dendrogram was cut at a height of 0.5, which was decided based on visual inspection. Clusters that had more than 40% *blank* samples were discarded together with any strains that fall on them. The remaining *blanks* were also discarded.

We extracted a number of features for each bacterial growth pattern which can be seen in Fig. 6.1. The area under the curve (AUC) for each strain and condition was calculated by adding the median OD600nm per time point using PGCA. We also extracted the maximum optical density for all samples of a given strain and condition (MAX), the average optical density over the last 3 days (L3M), the average time that it takes a strain to reach half of its maximum density (GSP). For the last three features, we calculated them for each condition as well as the $\log_2$ ratio from the condition that had the same ending phosphate concentration.

For grouping strains, according to their in vitro performance, their AUC was log-transformed and then standardized per condition. Hierarchical clustering was used with the Euclidean distance and the complete linkage method in R (R Core Team, 2014). The resulting groups are shown in Fig. 6.2. For selecting strains to test in in planta plate assays, an ANOVA model was fit on each strain using the AUC values as dependent variable and condition (media) as the only independent variable. We calculated the R^2 which indicated which proportion of the variation in in vitro performance (AUC) is attributable to media. We prioritized testing multiple strains per cluster (Fig. 6.2) that had the highest R^2 values. The code and data to perform these analysis is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

6.7.5 Phylogenetic signal analyses

For all strains with an available Sanger generated 16S rRNA gene sequence (395/440), we used MUSCLE (Edgar, 2004) to perform a multiple sequence alignment with default parameters. We then filtered out positions that had more than 99% gaps as well as the top 10% most entropic sequences using QIIME (Caporaso et al., 2010). The resulting filtered

alignment was used to build a maximum likelihood tree with FastTree (Price et al., 2009) using midpoint rooting.

We standardized all the phenotypes to allow for simultaneous visualization and easier comparison. We used the phylosig function from the phytools R package (Revell, 2012) to test Pagel's lambda (Pagel, 1999) for phylogenetic signal. The results are shown in Fig. 6.1 for the growth curve features and Fig. 6.4a for the binary association plate assays. Results were visualized with the ggtree R package (Yu et al., 2016). The code and data to perform these analysis is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

6.7.6 Plant-bacteria binary association assays

For binary-association experiments, plants were germinated in axenic condition on Johnson medium [KNO_3 (0.6g/L), $Ca(NO_3)_2$*$4H_2O$ (0.9g/L), $MgSO_4$*$7H_2O$ (0.2g/L), KCl (3.8mg/L), H_3BO_3 (1.5mg/L), $MnSO_4$*H_2O (0.8mg/L), $ZnSO_4$*$7H_2O$ (0.6mg/L), $CuSO_4$*$5H_2O$ (0.1mg/L), H_2MoO_4 (16.1μg/L), $FeSO_4$*$7H_2O$ (1.1mg/L), Myo-Inositol (0.1g/L), MES (0.5g/L), pH 5.6-5.7, 1% bacto-agar (BD, Difco),] 0.5% sucrose with 1mM Pi, $\sim$5 μM Pi [traces of Pi from the agar, (Difco)] or supplemented with 1mM phosphite in a vertical position for 7d. Seedlings were then transferred to 30μM Pi and 100μM Pi media (without sucrose) alone or with the monoculture at 10^5 c.f.u/mL of medium, for another 7d. Arabidopsis plants were grown in a growth chamber in a 16-h light/8-h dark regime (24°C/21°C).

For the demonstration that plant germinated in different Pi regimens or with phosphite differential activated the PSR in sterile conditions, plants overexpressing the PSR reporter construct IPS1:GUS13 were grown in Johnson medium containing 1mM Pi, 1mM phosphite or traces of Pi $\sim$5 μM Pi. After 7 days, the expression of the reporter constructs IPS1:GUS, highly induced by low Pi, was followed by GUS staining.

6.7.7 Synthetic community experiments

For synthetic community experiments, plants were germinated in axenic condition on Johnson medium 0.5% sucrose with 1mM Pi or $\sim$5μM Pi [traces of Pi from the agar, (Difco)] in a vertical position for 7 days; then transferred to 30μM Pi or 100μM Pi media (without

sucrose) alone or with the Synthetic Community at 10^5 c.f.u/mL of medium, for another 7 days. Arabidopsis plants were grown in a growth chamber in a 16-h light/8-h dark regime (24°C/21°C). Plant material was collected for transcriptional analysis (section 6.7.16) and for 16S profiling (section 6.7.13).

For the validation experiments, plants were germinated in axenic condition on Johnson medium 0.5% sucrose without supplementation of Pi in a vertical position for 7 days; then transferred to 30μM Pi medium (without sucrose) alone or with the synthetic communities at 10^5 c.f.u/mL of medium, for another 7 days. Arabidopsis plants were grown in a growth chamber in a 16-h light/8-h dark regime (24°C/21°C). Plant material was collected for 16S profiling (section 6.7.13).

6.7.8 Bacterial growth for binary association and synthetic community experiments

For mono-association and synthetic community experiments a single colony was inoculated in 4mL of 2xYT medium (16 g/L Tryptone, 10 g/L Yeast Extract, 5 g/L NaCL, ∼5.5mM Pi) in a test tube. Bacteria cultures were grown at 28°C with agitation over-night. Cultures were then rinsed with a sterile solution of 10mM $MgCl_2$ followed of a centrifugation step at maximal speed (2600g) for 8min. This process was repeated twice to eliminate any additional nutrient supplementation in the media. The OD600nm was measured and assuming that 1 OD600nm unit is equal to 10^9 c.f.u/mL we equalized individual bacterium concentration to a final value of 10^5 c.f.u/mL of medium. Medium was cooled down (to 40-44°C) near the solidification point and then the bacterium or the bacteria mix was added to the medium with agitation.

6.7.9 Block and synthetic community design

Before deciding which strains to include in the synthetic community experiments, we first identified which strains had a statistically significant effect on shoot phosphate accumulation. To that end, we compared the phosphate content between plants treated with a focal individual strain, and axenically grown plants on the same experiment as the focal strain. To determine significance, we log-transformed the phosphate content and used an ANOVA model in R (R

Core Team, 2014) with terms for bacterial treatment and biological replicate. Each of the four phosphate conditions was analyzed independently. By testing the bacterial treatment term in the model, we determined whether the effect was significant. We corrected all the obtained p-values using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). The code and data to perform this statistical analysis is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

The majority of the strains have negative effects, so we first identified strains that had a significant positive effect (q-vale < 0.1 from ANOVA) in the two conditions that end at 100μM phosphate concentration. We found 26 such strains and we labelled them as positive strains (Fig. 6.6a). Many strains are negative in two conditions or more, so we first identified strains that had a statistically significant (q-value < 0.1 from ANOVA) negative effect on shoot phosphate accumulation in at least three of the four conditions. We then identified strains that had a statistically significant negative effect in at least two conditions but with higher statistical confidence (q-value < 0.05). We removed two strains that did not come from our Brassicaceae cultivation efforts in two natural soils (strains Pseudomonas fluorescens WCS417r and R219). We combined the two sets of negative strains and obtained 26 strains that we termed negative strains (Fig. 6.6a). We finally identified strains that had no statistically significant effect (q-value > 0.1 from ANOVA) on phosphate accumulation in all conditions, and we randomly and programmatically selected 26 strains that we termed as the Indifferent groups (Fig. 6.6a). We finally calculated the mean effect that each bacterium on shoot phosphate accumulation by averaging the coefficients from the ANOVA in all conditions. We sorted the strains within each group according to that mean, and we divided each group into three blocks of bacteria by taking groups of 9, 9 and 8 strains ($9 + 9 + 8 = 26$; Fig. 6.6a). The bacterial blocks defined this way are the basis for the synthetic community design (Fig. 6.6b). The code and data to perform this statistical analysis is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

We decided to test a set of fourteen partially overlapping synthetic communities. Each

of the synthetic communities was made of a combination of two bacterial blocks. We made nine synthetic communities by combining adjacent blocks (i.e. blocks that are next to each other when sorted by their mean effect). Each of these nine communities is represented as an outer arc in Fig. 6.6b, and they are made mostly of strains that have similar effects on shoot phosphate accumulation when tested in binary association, but they represent the widest possible range of mean effects, and so we expect they will produce different plant phenotypes. We constructed another five synthetic communities which are represented as inner arcs in Fig. 6.6b that represent extra combinations of the most extreme blocks (*i.e.* P1 and N3) in order to test how strong their effects will be in a variety of backgrounds.

6.7.10 Shoot colonization experiments

To study the colonization of the plant shoot by root-inoculated bacteria, we germinated Col-0 and three different phosphate starvation response mutants: *pho1*, *phf1* and *phr1 phl1* (all mutants are in Col-0 background) on Johnson medium 1mM Pi, not supplemented with Pi or 1mM Phosphite for 7 days. Seedlings were then transferred to two-compartment plates for a week. In this system, root and shoot were placed in different compartments separated by a plastic barrier to prevent microbe diffusion through the medium. The root compartment was previously filled with Johnson medium 30μM Pi containing bacteria and the shoot compartment was filled with a solution of agar (water + 1% agar). Arabidopsis plants were grown in a growth chamber in a 16-h light/8-h dark regime (24°C/21°C). Bacteria accumulation in shoots and roots was analyzed in mock-inoculated plants and plants colonized by bacteria.

Roots, shoots and agar samples were harvested and weighted. Roots and shoots were rinsed 3 times with sterile distilled water to removed agar particles and no roots associated microbes, then placed in a sterile tube with 1mL of 10mM of $MgCl_2$. Plant material and agar samples were then crushed. These samples were serial diluted, plated in LB and colony-forming units (CFU) per ml of original culture were determined.

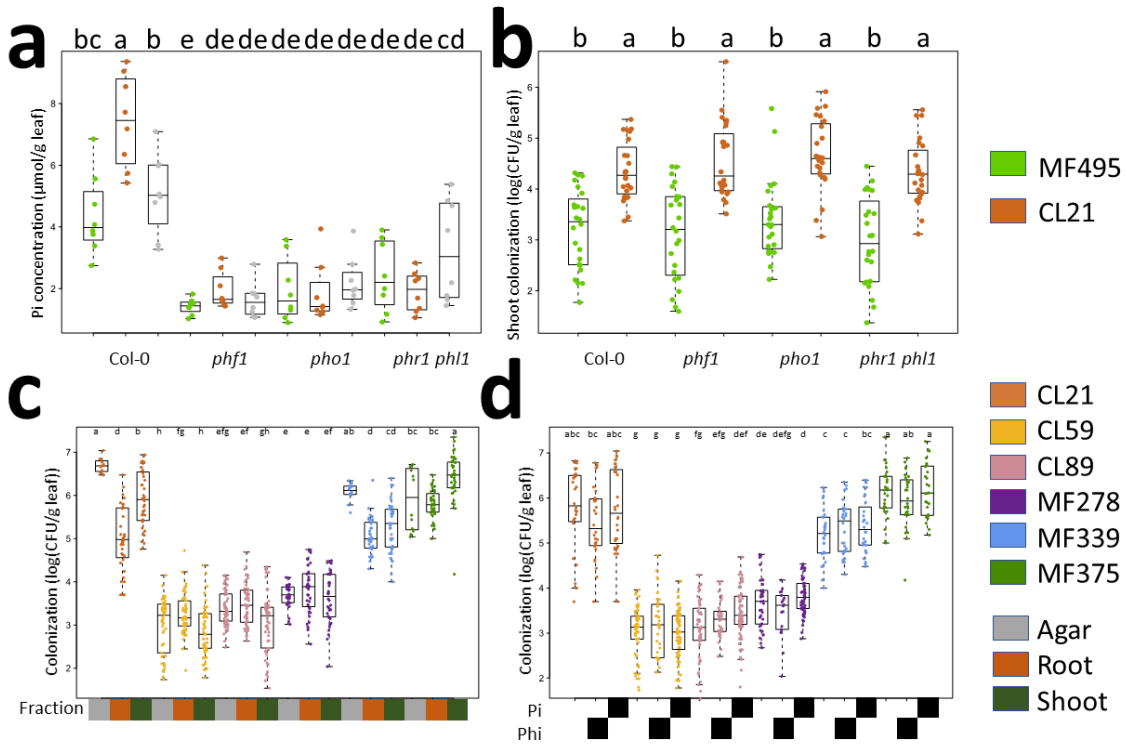


Figure 6.14: **Bacterial colonization and their effect on phosphate starvation are independent**. a shoot phosphate accumulation caused by to bacterial strains in Col-0 and three mutants deficient on the plant phosphate starvation response. b Bacterial shoot colonization for the strains and mutants in a. Effect on plant phosphate is independent of colonization levels. c plant colonization of 6 bacterial strains in different plant organs. d plant colonization of 6 bacterial strains according to different pre-treatments. Colonization is independent of activation of the plant phosphate starvation response.

6.7.11 Plant phenotyping

The method of Ames was used to determine the shoot free phosphate concentration of seedlings grown on different Pi regimens and treatments. Main root elongation was measured using ImageJ (Abramoff et al., 2004), and shoot area and total root network were measured with WinRhizo (Arsenault et al., 1996).

6.7.12 Estimating block additivity

To determine the degree of consistency of bacterial block effects on different plant phenotypes, we first compared plants inoculated with each community versus their axenically grown controls. We then estimated the main effects of each block using multiple regression and we compared the coefficients obtained from both methods. Phosphate content and rosette size measurements were log-transformed to reduce heteroscedasticity, and thus the coefficients from this analysis should be interpreted as log(fold-change) between inoculated plants and axenic controls. Measurements from main root elongation and total root network were adequately represented by our linear models, and so coefficients for these two phenotypes should be interpreted as the difference between inoculated plants and axenic controls. Only the fourteen original synthetic communities were included in the analysis.

To estimate synthetic community effects, we fit a linear model per phenotype and condition, only with the samples of one synthetic community at a time, plus the axenic controls performed on the same experiments. We had only bacterial treatment and experiment variables according to the following formula:

$$Phenotype = SynCom + Experiment \tag{6.1}$$

The resulting $SynCom$ coefficient was denominated the 'measured' synthetic community effect:

To find the expected phenotypic effect of each synthetic community, we first estimated each block's additive (main) effect. We did this by fitting all the data from each media condition into one linear model containing only terms for each block and experiment as

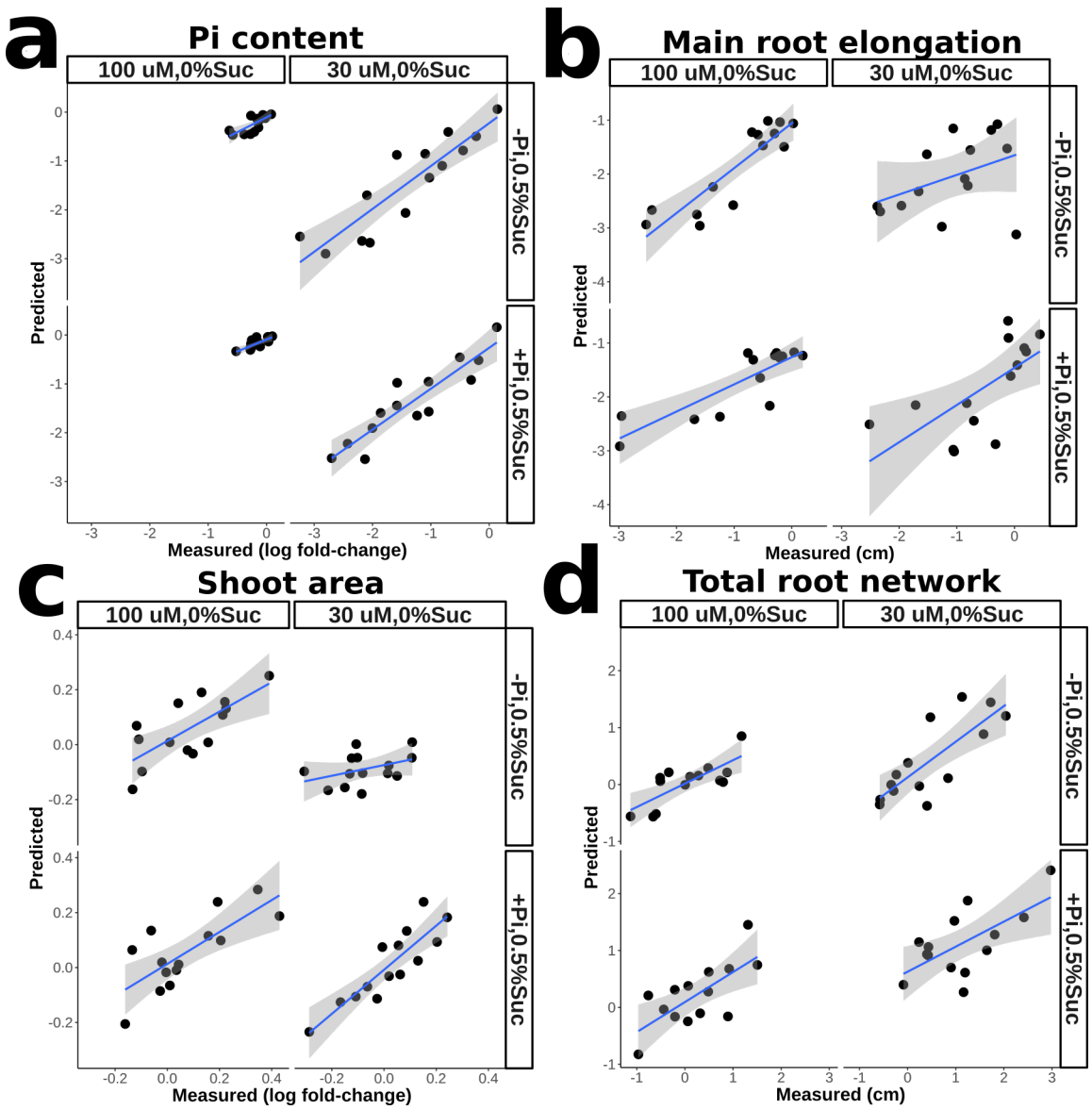


Figure 6.15: Additive contributions of bacterial blocks explain synthetic community phenotypes. Comparison between measured changes (x-axis) in plant phenotypes caused by synthetic communities with respect to axenically grown plants, and expected changes (y-axis) from purely additive effects of each block. In each plot, the four panels represent the four media conditions tested with pre-treatment as rows and post-treatment as columns. X- axis corresponds to the color-scale in Fig. 6.7, and Y-axis shows the result from adding the individual main effect estimated for each block (section 6.7.12. The blue line represents the least squares regression on the points from each panel. In all axes zero represents no change with respect to axenically grown plants. The phenotypes shown are shoot phosphate accumulation (a), main root elongation (b), shoot area (c) and total root network (d). The values for Pi-content and shoot area (a and c) indicate log fold-change with respect to axenically grown plants. The values for main root elongation and total root network (b and d) represent difference with respect to axenically grown plants.

independent variables using the following formula:

$$Phenotype = P1 + P2 + P3 + I1 + I2 + I3 + N1 + N2 + N3 + Experiment \qquad (6.2)$$

Each of the block variables is encoded as an indicator variable where they have the value of '1' when the corresponding block is present, and '0' when the corresponding block is absent.

We finally obtained the 'predicted' community effect by arithmetically adding the coefficients for the two blocks that make each community. The comparison between both 'measured' and 'predicted' synthetic community effects is shown in Fig. 6.15.

To compare the effects of both synthetic communities and blocks together, we re-scaled them by dividing them by the standard deviation of all coefficients from the same phenotype in the same condition (*i.e.* by column). Results are shown in Fig. 6.10. The code and data to perform this statistical analysis is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

6.7.13 DNA extraction for 16S analysis

For bacterial colonization analysis using 16s in synthetic communities experiments, roots were surface sterilized with freshly made 10% bleach with 0.1% Triton-X100 for 12 minutes. Following the bleaching, roots were rinsed once in sterile distilled water, then placed in 2.5% sodium thiosulfate to neutralize the bleach for 2 minutes, and rinsed once more with sterile distilled water. Roots were then freeze-dried and powdering in a 2mL tube with glass beads using the MPBio FastPrep for 20s at 4.0 m/s. These samples were used for DNA extraction using 96-well format MoBio PowerSoil kit (SDS/mechanical lysis) following the manufacturer's instructions.

To quantify bacteria from agar samples in the synthetic community experiments, freeze and squeeze protocol was used. Syringes, with a square of sterilized miracloth on the bottom, were completely packed with agar samples and kept at -20°C for a week. After that, samples were thawed at room temperature and syringes were squeezed gently into 50mL tubes. Samples

were centrifuged at max speed for 20 min and most of the supernatants were discarded. The remaining 1-2mL of supernatants, containing the pellets, was moved into clean microfuge tubes. Samples were centrifuged again, supernatants were removed and pellets were used for DNA extraction with 96-well format MoBio PowerSoil kit (SDS/mechanical lysis).

DNA was extracted simultaneously for both agar and root samples. We perfumed randomization of sample order using a mechanical method.

6.7.14 Synthetic community experiments 16S library preparation

We amplified the V3-V4 regions of the bacterial 16S rRNA gene using primers 338F (5'-ACTCCTACGGGAGGCAGCA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3'). Libraries were created using a modified version of the method by Lundberg et al. (2013), which is described in Castrillo et al. (2017). Basically, the molecule-tagging step was changed to an exponential amplification to account for low DNA yields with the following reaction:

- 5μL of Kapa Enhancer
- 5μL of Kapa Buffer A
- 1.25μL of 5μM 338F
- 1.25μL of 5μM 806R
- 0.375μL of mixed PNAs (1:1 mix of 100μM pPNA and 100μM mPNA)
- 0.5μL Kapa dNTPs
- 0.2μ Kapa Robust Taq
- 5μL DNA

With the following temperature cycling:

1. 95°C for 60 seconds
2. 24 cycles of:

(a) 95°C for 15 seconds

(b) 78°C for 10 seconds (PNA annealing)

(c) 50°C for 30 seconds

(d) 72°for 30 seconds

3. 12°C for 5 minutes

4. 4°C for ever

Following PCR cleanup to remove primer dimers, the PCR product was indexed using the same reaction and 9 cycles of the cycling conditions described in Lundberg et al. (2013). Sequencing was performed at UNC on an Illumina MiSeq instrument using a 600-cycle V3 kit. The raw data from these sequencing experiments is available in the EBI Sequence Read Archive (accession number will be made available upon submission).

6.7.15 16S profiling sequence processing and analysis

Synthetic community sequencing data were processed with MT-Toolbox (Yourstone et al., 2014). Categorizable reads from MT-Toolbox (i.e. reads with correct primer and primer sequences that successfully merged with their pair) were quality filtered with Sickle by not allowing any window with Q-score under 20, and trimmed from the 5' end to a final length of 350 bp. The resulting sequences were matched to a reference set of the strains in the synthetic community generated from Sanger sequences and Arabidopsis organellar sequences. Sequence mapping was done with USEARCH7.1090 with the option "-usearch_global" at a 98.5% identity threshold (which translates to four mismatches for our sequence length). XX% of sequences matched an expected isolate, and those sequence mapping results were used to produce an isolate abundance table.

The code and data to perform these analysis is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

6.7.16 RNA isolation for transcriptomics

Total RNA was extracted from roots of Arabidopsis according to Logemann et al. (1987). Frozen seedlings were pulverized in liquid nitrogen. Samples were homogenized in 400μl of Z6-buffer; 8M guanidinium-HCl, 20mM MES, 20mM EDTA pH 7.0. Following the addition of 400μl phenol:chloroform:isoamylalcohol; 25:24:1, samples were vortexed and centrifuged (20000g, 10 min) for phase separation. The aqueous phase was transferred to a new 1.5ml tube and 0.05 volumes of 1N acetic acid and 0.7 volumes 96% ethanol were added. The RNA was precipitated at -20°C overnight. Following centrifugation, (20000g, 10 min, 4°C) the pellet was washed with 200μl sodium-acetate (pH 5.2) and 70% ethanol. The RNA was dried, and dissolved in 30μl of ultrapure water and stored at -80°C until use.

6.7.17 RNA-seq library construction

Illumina-based mRNA-seq libraries were prepared from 1μg RNA. Briefly, mRNA was purified from total RNA using Sera-mag oligo(dT) magnetic beads (GE Healthcare Life Sciences) and then fragmented in the presence of divalent cations (Mg2+) at 94°C for 6 min. The resulting fragmented mRNA was used for first-strand cDNA synthesis using random hexamers and reverse transcriptase, followed by second strand cDNA synthesis using DNA Polymerase I and RNAseH. Double-stranded cDNA was end-repaired using T4 DNA polymerase, T4 polynucleotide kinase and Klenow polymerase. The DNA fragments were then adenylated using Klenow exo-polymerase to allow the ligation of Illumina Truseq HT adapters (D501D508 and D701D712). All enzymes were purchased from Enzymatics. Following library preparation, quality control and quantification were performed using a 2100 Bioanalyzer instrument (Agilent) and the Quant-iT PicoGreen dsDNA Reagent (Invitrogen), respectively. Libraries were sequenced using Illumina HiSeq2500 sequencers to generate 50bp single-end reads.

6.7.18 RNA-seq sequence processing and analysis

Initial quality assessment of the Illumina RNA-seq reads was performed using the FASTX-Toolkit. Cutadapt was used to identify and discard reads containing the Illumina adapter

sequence. The resulting high-quality reads were then mapped against the TAIR10 Arabidopsis reference genome using Tophat, with parameters set to allow only one mismatch and discard any read that mapped to multiple positions in the reference. The Python package HTSeq was used to count reads that mapped to each one of the 27,206 nuclear protein-coding genes. Raw sequencing data and read counts are available at the NCBI Gene Expression Omnibus (accession number will be made available upon submission).

For expression of the phosphate starvation response core markers, and the clustering analysis. We converted the count table into a table of reads per kilobase per million (RPKM) table, and standardized these values per gene, by subtracting the mean gene expression and dividing by the standard deviation of each gene. Hierarchical clustering was performed with the R function hclust using the complete linkage method. Gene ontology enrichment analysis was performed on the PlantGSEA online platform.

For the specific hypothesis tests, we used edgeR to fit a quasi-likelihood negative binomial model with the function glmQLFit, after estimating tagwise dispersion parameters. We then applied a quasi-likelihood ratio test with the function glmQLFTest in edgeR, using different sets of contrast. Our model specification included indicator terms for each bacterial block, as well as terms for phosphate condition, biological replicate, and interaction between phosphate condition and each block. This was defined according to the following formula:

$$\begin{aligned} Expression = {} & P1 + P2 + P3 + I1 + I2 + I3 + N1 + N2 + N3 + \\ & Phosphate + Experiment + \\ & P1 * Phoshpate + P2 * Phoshpate + P3 * Phoshpate + \\ & I1 * Phoshpate + I2 * Phoshpate + I3 * Phoshpate + \\ & N1 * Phoshpate + N2 * Phoshpate + N3 * Phoshpate \end{aligned} \tag{6.3}$$

The first 9 terms are indicator variables for the bacterial blocks of the synthetic communities that take the value of 1 when that block is present. Phosphate has 4 levels that correspond to the *two by two* phosphate condition experimental design. Experiment has 5 levels that

correspond to independent biological replicates of the synthetic community experiments (each community was in two biological replicates). A total of 41 coefficients (including intercept are generated from this design). The definition of the contrasts used for the different hypothesis is bundled in the R package wheelP, which will be made public when this manuscript is submitted for publication.

To plot the gene ontology hierarchy, we used code from METACODER PAPER, which adapts code from the metacoder R package.

The code and data to perform these analysis is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

6.7.19 Neural network construction

We focus on shoot phosphate accumulation, which had the lowest experimental variance of all phenotypes, and serves as a bedrock test for our approach. We used a multilayer feed forward neural network, a typical framework in deep neural network structure family, where input data are combined and transformed non-linearly through multiple layers of hidden neurons and nodes (Hornik et al., 1989).

First, we optimized neural network model over diverse architectures (width and depth) and hyper-parameters such as training iterations and regularizations, which prevents over-fitting. Secondly, we estimated the prediction error associated with our neural network with a *leave-SynCom-out* cross-validation experiment; in this case the model is only trained on all but one synthetic community and tested on that held out synthetic community in each fold. Based on the cross-validation experiment, we choose the best neural network model architecture, which was with 3 hidden layers and 200 hidden nodes in each layer, as shown in Fig. 6.11a).

In line with our expectations, we found that the neural network (NN) has the lowest prediction error on held-out synthetic community samples as compared with simple linear model (LM), and a linear model with manually constructed interaction features (INT) (Fig. 6.11b). Finally, to visualize the learned model of the neural network we investigate the

"sensitivity" of the network, which shows how much the predicted output changes after a perturbation in a particular input feature (Fig. 6.11c-e). This approach is similar to derivatives analysis of the network in many computer vision tasks (Simonyan et al., 2013; Wang et al., 2016). We found that, while sensitivity is constant in the linear model for a feature across different contexts, it changed in both the linear model with interactions and the neural network (Fig. 6.11c-e). Nevertheless, the changes in the neural network were bigger as compared with the linear model with interactions. This result indicates that the neural network can capture more complex context-dependent sensitivity than the rest of the models tested.

We define each input contains a biological replicate ID $b \in \{1,2\}$, a technical replicate ID $r \in \{1,2,3\}$, a pre-treatment $p \in \{\text{-Pi},\text{+Pi}\}$, post-treatment $q \in \{30\mu\text{M},100\mu\text{M}\}$ and a synthetic community $S \subseteq \{\text{P1,P2,P3,I1,I2,I3,N1,N2,N3}\}$. We call the combination of p and q as **phosphate condition**. A "design" is a combination of phosphate condition and synthetic community and an "input condition" is a combination of design and biological replicate ID. Let $z_{b,p,q,S,r}$ be the standardized Pi-content measurement for biological replicate b, technical replicated r, pre-treatment p, phosphate-level q, and synthetic community input S. The mean of Pi-content across three technical replicates is $y_{b,p,q,S} = \frac{1}{3}\sum_{r=1}^{3} z_{b,p,q,S,r}$ and its variance is $v_{b,p,q,S} = \text{Var}(z_{b,p,q,S,1}, z_{b,p,q,S,2}, z_{b,p,q,S,3})$. An input $\mathbf{x}_{b,p,q,S}$ is constructed as a binary vector of length 12. A description of how to construct the input vector $\mathbf{x}$ is showed in Table 6.1. A model learns a function $f(\mathbf{x}_{b,p,q,S}) = \hat{y}_{b,p,q,S}$ that takes binary input vector $\mathbf{x}$ as input and outputs the prediction for mean Pi-content $\hat{y}_{b,p,q,S}$.

Table 6.1: Description of input features $\mathbf{x}_{b,p,q,S}$

Feature Order i	1	2	3	4–12
Feature Name	BioRep ID b	Pre-treatment p	Phosphate q	$\mathcal{B} = \{P1, P2, P3,\ I1, I2, I3, N1, N2, N3\}$
$x_{b,p,q,S,i} = 0$	$b = 1$	$p =$ -Pi	$q = 30$uM	$\mathcal{B}_{i-3} \notin S$
$x_{b,p,q,S,i} = 1$	$b = 2$	$p =$ +Pi	$q = 100$uM	$\mathcal{B}_{i-3} \in S$

A linear model (LM) has the following form:

$$f_{\text{LM}}(\mathbf{x}) = b + \mathbf{x}\mathbf{w},$$

where b is the bias term. $\mathbf{w}$ is a vector of length p to indicate the linear effect of each feature. A linear model with interaction (INT) features has the following form:

$$f_{\text{INT}}(\mathbf{x}) = b + \mathbf{x}\mathbf{w} + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} x_i x_j \Theta_{ij},$$

where $\boldsymbol{\theta}$ is an upper triangular matrix where diagonal entries are zero. Comparing to LM, INT is able capture conditional specific behaviors. For example, a synthetic community can have different impact on Pi-content under different phosphate conditions. Elastic net regularization (Zou et al., 2007) is used to learn the parameters for both linear model and linear model with interaction features. An elastic net regularization has the following optimization objective:

$$\begin{aligned}
\text{LM:} \quad \mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_i (y_i - f_{\text{LM}}(\mathbf{x}_i))^2 + \lambda_1 |\mathbf{w}| + \lambda_2 ||\mathbf{w}||_2^2 \\
\text{INT:} \quad [\mathbf{w}^*, \boldsymbol{\theta}^*] &= \underset{\mathbf{w},\boldsymbol{\theta}}{\operatorname{argmin}} \sum_i (y_i - f_{\text{INT}}(\mathbf{x}_i))^2 + \lambda_1 (|\mathbf{w}| + |\boldsymbol{\theta}|) + \lambda_2 (||\mathbf{w}||_2^2 + \|\boldsymbol{\theta}\|_2^2)
\end{aligned}$$

where λ_1, λ_2 represent the regularization penalty for $l1$-norm and $l2$-norm of the parameters in the model. We optimize this objective to learn the parameters. λ_1 and λ_2 are chosen from a 10-fold cross-validation.

A Multilayer Feed forward neural network (NN) is used in our experiment. An NN contains an input layer, an output layer and L hidden layers. The "depth" of the network is the number of hidden layers and the "width" is the number of nodes in each hidden layer. For convenience, the input layer is defined as $\mathbf{h}_0(\mathbf{x}) = \mathbf{x}$, and the output of lth hidden layer is defined as $\mathbf{h}_l(\mathbf{x})$. The number of nodes in layer l is m_l. The activation that goes into lth

hidden layer is defined as:

$$\mathbf{a}_l(\mathbf{x}) = \mathbf{h}_{l-1}(\mathbf{x})\mathbf{W}_l + \mathbf{b}_l,$$

where $\mathbf{W}_l$ is a real value weight matrix of m_{l-1} by m_l and $\mathbf{b}_l$ is a bias vector of length m_l. The output of lth hidden layer is:

$$\mathbf{h}_l(\mathbf{x}) = \text{leakyReLU}(\mathbf{a}_l(\mathbf{x})),$$

where leaky-Rectified Linear Unit (Glorot et al., 2011) activation function is:

$$\text{leakyReLU}(\mathbf{x}) = \begin{cases} x & \text{, if } x \geq 0 \\ 0.01x & \text{, otherwise} \end{cases}$$

Finally, a linear output layer is on top of the last hidden layer:

$$f_{\text{NN}}(\mathbf{x}) = \mathbf{h}_L(\mathbf{x})\mathbf{W}^{L+1} + b_{L+1}$$

To train the NN, RMSprop (Tieleman and Hinton, 2012) with momentum 0.9 is used to tune the parameters. Only structures with equal width in each layer are considered in this experiment. The final network is chosen by best cross-validation error from the following hyper-parameter settings: depth of 1 to 4, width of 100,200,300,400,500, weight-decay penalty of 0.001,0.005,0.01,0.05,0.1,0.5, and epochs of 100,200,300,400,500. The final network has depth of 3, width of 200 and weight-decay penalty of 0.05 and epochs of 100.

The code to fit the neural network and the two linear models will be made available upon submission.

6.7.20 Sensitivity in different models

The sensitivity ρ_i of a feature i under a certain input context $\mathbf{x}$ is defined as:

$$\rho_i(\mathbf{x}) = f(\mathbf{x}_{(x_i=1)}) - f(\mathbf{x}_{(x_i=0)}),$$

where $\mathbf{x}_{(x_i=a)}$ means change the ith feature in $\mathbf{x}$ to a and keep other features fixed. As our inputs are all binary features, sensitivity is the difference in the output between the "on" (1) and "off" (0) state of a particular feature.

In the linear model, $\rho_i(\mathbf{x}) = w_i$. Hence, the sensitivity in linear model is independent of input context.

In the linear model with interaction features, $\rho_i(\mathbf{x}) = w_i + \sum_{j \neq i} x_j \Theta_{ij}$. Therefore, it has a input context dependent sensitivity.

In neural networks, exact form of $\rho_i(\mathbf{x})$ is complicated and hard to calculate. In our experiment, we calculate sensitivity $\rho_i(\mathbf{x})$ numerically.

In order to compare the three models, all possible inputs with synthetic community size $|S| = 2$ are generated as contexts to calculated the sensitivity.

The code to estimate sensitivity of the different models will be made available upon submission.

6.7.21 Generation of block swaps

We sought to identify cases where by replacing (*swapping*) one of the bacterial block in a reference synthetic community $S_1 = \{A, B\}$ for a different bacterial block resulting in the perturbed synthetic community $S_2 = \{A, C\}$ under a certain phosphate condition (pre-treatment p, post-treatment q), where A, B, C are different bacterial blocks, would induce significant improvement in Pi-content. To compare two synthetic communities, we can use trained model to estimate the mean and variance of the output in two synthetic communities. Mean Pi-content prediction for any input of interest can be calculated as $f_{b,p,q,S} = f(\mathbf{x}_{b,p,q,S})$. A worst-case variance estimate was used for our prediction, where the largest residual variance (difference between observed value and predicted value) related to a bacteria block is transferred from the training data to all related predictions: $\hat{z}_{b,p,q,S,r} = f_{b,p,q,S} + z_{b,p,q,M,r} - f_{b,p,q,M}$, where

$$M = \{e_1^*, e_2^*\} = \underset{e_1 \in S \text{ or } e_2 \in S}{\operatorname{argmax}} v_{b,p,q,\{e_1,e_2\}}.$$

6 predicted samples can be generated for any design: $\hat{\mathbf{z}}_{p,q,S} = \{\hat{z}_{b,p,q,S,r} | b \in \{1,2\} \wedge r \in \{1,2,3\}\}$. Given any two synthetic communities S_1, S_2 under a certain phosphate condition p, q, the mean difference is $\bar{\hat{z}}_{p,q,S_2} - \bar{\hat{z}}_{p,q,S_1}$ and the p-value is calculated from a two-sample t-test on $\hat{\mathbf{z}}_{p,q,S_1}$ and $\hat{\mathbf{z}}_{p,q,S_2}$.

The code to generate candidate block swaps will be made available upon submission.

6.7.22 Data and software accessibility

All data generated from this project is publicly available. Raw sequences from 16S profiling are available at the EBI Sequence Read Archive under accession XXXXXXX (will be made available upon submission). Raw sequences from transcriptomic experiments are available at the NCBI Gene Expression Omnibus under the accession number XXXXXX (will be made available upon submission).

The code and processed data from the *in vitro* growth curves, plant-bacteria binary associations, synthetic community 16S profiling and transcriptomics. As well as function and scripts for all analysis from *in vitro* experiments, binary association assays, 16S and transcriptomic analysis and block additive effects is bundled in the R package (wheelP) which will be made public when this manuscript is submitted for publication.

The code to fit the neural network, estimate sensitivity and generate hypothesis is will be made available upon submission.

CHAPTER 7

Root microbiome members act in isolation and in concert to modulate plant phenotypes [1]

Identifying microbes that influence host health remains a major challenge, but most methods rely on identifying corelations between bacterial abundances and host phenotypes (Gilbert et al., 2016). While these approaches are powerful, they cannot identify causal microbial effects on their host. Recently, the study of host-microbe interactions has benefited by an increasing availability of bacterial isolates from multiple hosts and environments (Nelson et al., 2010; Bai et al., 2015; Armanhi et al., 2016; Browne et al., 2016). These bacterial collections have shown that it is possible to access a large proportion of the previously called 'unculturable' microbiota.

The availability of bacterial isolate collections has led to the development of synthetic community approaches in which a defined community is provided to the an environment (Faith et al., 2010; McNulty et al., 2013; Bodenhausen et al., 2014; Faith et al., 2014; Bai et al., 2015; Lebeis et al., 2015; Rolig et al., 2015; Wei et al., 2015; Kastman et al., 2016; Castrillo et al., 2017; Niu et al., 2017). The synthetic community approach has provided extemeley valuable in identifying specific bacterial strains that are well adapted to the isolation environment (Bai et al., 2015; Kastman et al., 2016; Lebeis et al., 2015; Castrillo

[1]The contents of this chapter has not been peer-reviewed. It is an eary draft, that includes work in progress, of a manuscript that will be authored by myself (Sur Herrera Paredes). Multiple people at Jeff Dangl's group contributed to this work and will be recognized with authorship. Including but not limited to undergraduate students Emily Getzen, Jose Macalino Esteban and Surojit Biswas, as well as BBSP PhD student Isai Salas González. The specific contributions are as follow: SHP and JD designed the experiments. SHP and JME performed the main combinatorial experiment. SHP and EG performed the validation experiments. SHP, SB and ISG designed and implemented the image-based phenotyping pipeline. SHP and JD analyzed data and designed figures. SHP wrote the manuscript with input from JD.

et al., 2017) respond to host (Bodenhausen et al., 2014; Lebeis et al., 2015; Castrillo et al., 2017) and environmental (McNulty et al., 2013; Castrillo et al., 2017) factors. The synthetic community approach has also been useful in identifying microbe-microbe interactions that direct the assembly of microbial communities (Kastman et al., 2016; Niu et al., 2017). Despite these successes, the majority of studies look at a single synthetic community that is thought to be representative of a particular niche. Typically, this synthetic community is designed by maximizing diversity within some experimental constraints. This approach limits the number and type of biological questions that can be asket; in particular, using a single synthetic community cannot identify microbial factors that influence host phenotypes, since only correlations between those phenotypes and microbial abundances can be obtained.

Using multiple synthetic communities allows you to associate defined and reproducible changes in microbial community compositions with host phenotypic outputs. One study manipulated the bacterial within-genus diversity in plants, and was able to identify community configurations that reduced pathogen invasion success (Wei et al., 2015). Another study generated random bacterial communities, and identified strains that modulated mouse immune and metabolic phenotypes (Faith et al., 2014). A third study, tested all possible combinations of three strains, and showed that strains modulated Zebrafish innate immune responses independently of their absolute abundance (Rolig et al., 2015). Finally, we have shown that we can combine bacterial groups to direcltly estimate the effect of those groups on plant phenotypes, and that those estimates are predictive of novel communities (Chapter 6). These studies demonstrate the power of using multiple synthetic communities to identify microbial effects on plant host.

Here we show that we can scale-up the use of unbiased combinatorial synthetic community construction to identify specific bacterial strains that alter plant phenotypes. We tested nearly 400 synthetic communities covering 54 fully sequenced bacterial isolates, and coupled the bacterial treatments with imaging-based plant phenotyping. We extend previous combinatorial synthetic community approaches by incorporating time-course phenotyping data, and show

that we are able to simultaneously identify both individual strains, and combinations of strains that alter plant size and coloration. We show that plant size is mostly influenced by bacteria in an additive manner, while plant coloration can be modulated by specific combinations.

We have used the inferences made from our combinatorial synthetic community approach to design novel communities. We are currently testing whether our predictions hold, including testing in a novel context.

7.1 The experimental design

We took an experimental design approach to identify bacterial strains that affect phenotypes of their host, either alone or in combination. Our approach consists of randomly constructing synthetic communities and associating presence/absence of individual strains with phenotypic changes in the plant (Fig. 7.1a left side). We chose 54 fully sequenced bacterial strains isolated from roots of Brassicaceae plants growing in two characterized wild soils (Lundberg et al., 2012), and constructed nearly four hundred synthetic communities of seventeen members each. In principle, multiple plant phenotypic distributions are possible, but the most relevant for our aim is the increase of variance in the phenotype in the presence of variable communities (Fig. 7.1a top right). Because the bacterial compositions are completely randomized, an increase in phenotypic variance implies that different bacteria are differentially affecting the plant phenotype of interest. Importantly, if the design is random, and the phenotypic variation high enough, the responsible strains can be identified with standard association techniques (Fig. 7.1a bottom right).

Other studies have used combinations of strains (Faith et al., 2014) or of groups of strains (Chapter 6) and demonstrated that one can draw associations with predictive power. The first study was unable to find interactions, while the second cannot pinpoint specific strains that are responsible for the observed effects. We used power analysis to determine how robust our design would be to important experimental constraints (section 7.5.1). We saw very little loss in power from doubling the number of strains (Fig. 7.1c). The number of samples per

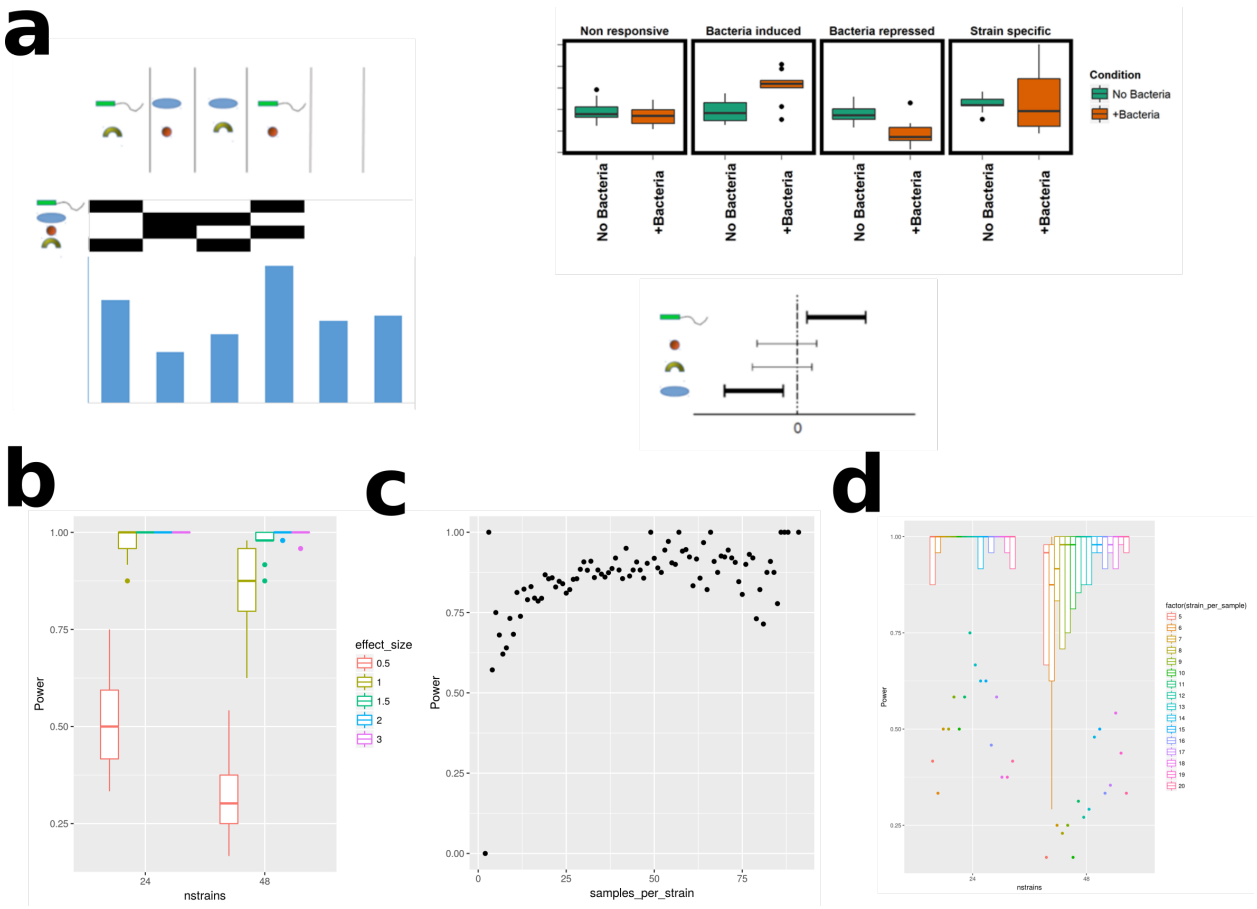


Figure 7.1: **Experimental design and power analysis**. a Experimental design approach. Left: schematic representation of four strains randomly combined into synthetic communities of size two. In this example four plants would receive one of the communities while two would remain uninoculated, and their phenotypes would be recorded. Right: potential phenotypic distributions on top; either the bacteria have a constant effect or they have a variable effect, in which case standard association methods can be used to identify the responsible strains. b Statistical power to detect associations as a function of the number of strains in the universe, and the effect size that an individual strain may have on a plant phenotype (in standard deviations). c Statistical power as a function of the number of samples in which one strain is found. d Statistical power as a function of the number of total strains and the number of strains per community. A representative of three replicates of power analysis is shown.

strain followed a typical arch pattern, which indicates that it is important to have a balanced representation of samples with and without a given strain (Fig. 7.1c. Statistical power was susceptible to the number of strains in each community, with very small communities having appreciably reduced power (Fig. 7.1d). We decided then to test 54 strains in randomized synthetic communities made of 17 strains each. This guaranteed that each strain would be present in at least one fourth of the samples of each biological replicate. We performed four independent biological replicates, drawing a different set of 96 independent communities each time.

We established an imaging-based phenotyping pipeline and we used to measure eight morphometric and nine colorimetric plant phenotypes through 49 days (section 7.5.5). Most morphometric phenotypes correlated with each other and are proxies for shoot size (Fig. 7.2 left). On the other hand, colorimetric phenotypes divide into a group that mostly correlates with green intensity, and another group that correlates with blue intensity (Fig. 7.2 right), though most of the variation in color can be explained as a function of green intensity, which is a proxy for nitrogen assimilation and photosynthesis rate (Muharam et al., 2015).

7.2 Results from combinatorial synthetic communities

A principal component analysis of all morphometric phenotypes shows that the phenotypic variation of plant shoots that have been treated with a random combination of bacteria falls in a similar, if slightly larger, range than the variation of plants not treated with a synthetic community (Fig. 7.3 left green vs magenta). Principal component analysis also showed that, as expected, time explains most of the variation related to plant size.

In order to account for the effect of time, and to leverage our time course observations, we used a Generalized Least Squares (GLS) model with a compound symmetry correlation structure (section 7.5.6). This approach retains the flexibility and simplicity of ordinary least squares methods, but relaxes the assumption of independent observations; thus allowing us to model the correlation between repeated observations on the same individual plant without over-estimating the number of independent observations. We tested the effect of each strain

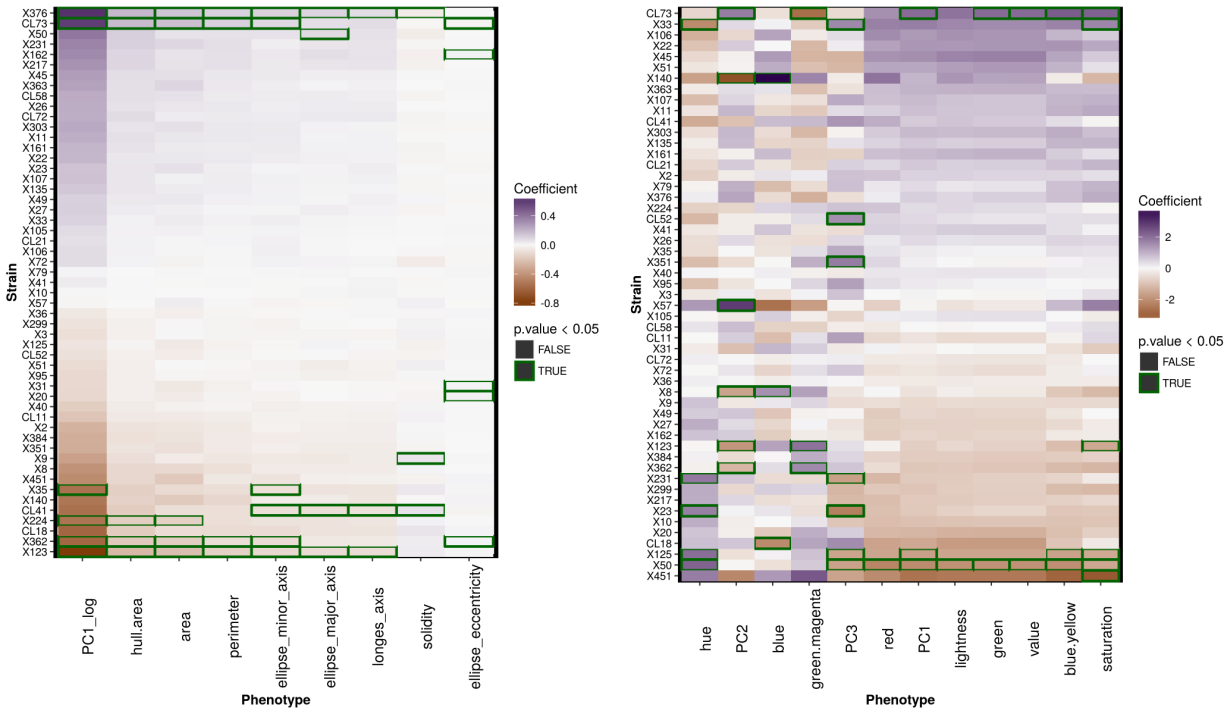


Figure 7.2: **Estimated effect (coefficient) of each strain on each phenotype**. Left shows morphometric phenotypes and right colorimetric phenotypes. Purple-Brown color scale in the heatmap indicates the whether the effect of each individual strain on a given phenotype is positive (Purple) or negative (Brown). Green rectangles indicate statistical significance (p-value < 0.05) based on the Generalized Least Squares model (section 7.5.6).

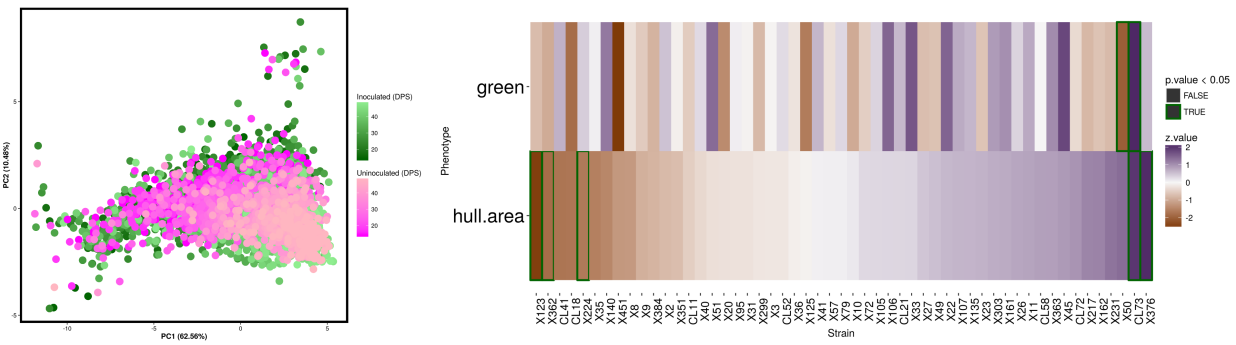


Figure 7.3: **Association of strains with plant phenotypes**. Left: principal component analysis of all morphometric features. Each dot represents a plant of a given age, which is indicated by the color. Clearer colors are older plants. DPS: days post-sowing. Right: Comparison of bacterial effects on green intensity (top) and hull area. Purple-Brown color scale in the heatmap indicates the whether the effect of each individual strain on a given phenotype is positive (Purple) or negative (Brown). Green rectangles indicates statistical significance (p-value < 0.05) based on the Generalized Least Squares model (section 7.5.6).

on each phenotype and we observed consistent results between correlated phenotypes (Fig. 7.2), but we saw that different strains influenced morphometric and colorimetric phenotype (Fig. 7.3 right). Thus, our approach allows us to simultaneously identify bacteria that alter independent plant phenotypes.

We can also ask whether the effect that bacteria have on multiple phenotypes is constant or varies with time (and plant developmental stage). We achieved this by analyzing the data according to a sliding window and, for the most part, we did not observe a strong variation of bacterial effects as a function of time (Fig. 7.4a-b). This analysis showed that bacterial effects are hard to estimate in their first couple of weeks, but that they stabilize around day 20 (Fig 7.4a-b), suggesting that events that happen around that time of plant-bacteria contact are responsible for the observed differences at later time points. Consistent with our previous results (Chapter 6), most bacteria have a slightly negative effect on plant shoot size when compared with plants that were not directly treated with a synthetic community, and positive effects were generally weaker than negative effects (Fig. 7.4a,c).

In principle, it should be possible to also associate genes in the isolates that we included in the synthetic communities, with phenotypic outcomes for the plant. We used presence/absence of KEGG orthology groups to find if the presence of specific genes were associated with either of the plant phenotypes. We found no evidence of associations of bacterial genes with plant size, but we found a significant enrichment of small p-values when we tried to associate bacterial genes with green intensity (Fig. 7.5). However, those small p-values dissipated after we corrected for multiple testing. Using the method of Storey and Tibshirani (2003) we estimated that 59% (π_0 in Storey and Tibshirani (2003)) of the bacterial genes should be associated with a change in plant shoot green intensity. This calculation does not control for the fact that many orthologue groups are highly correlated with one another, but strongly suggests that it is possible to find true associations. At this point, our study is underpowered to distinguish *bacterial gene-by-plant phenotype* relationships.

Another important question is whether specific combinations of strains produce changes in

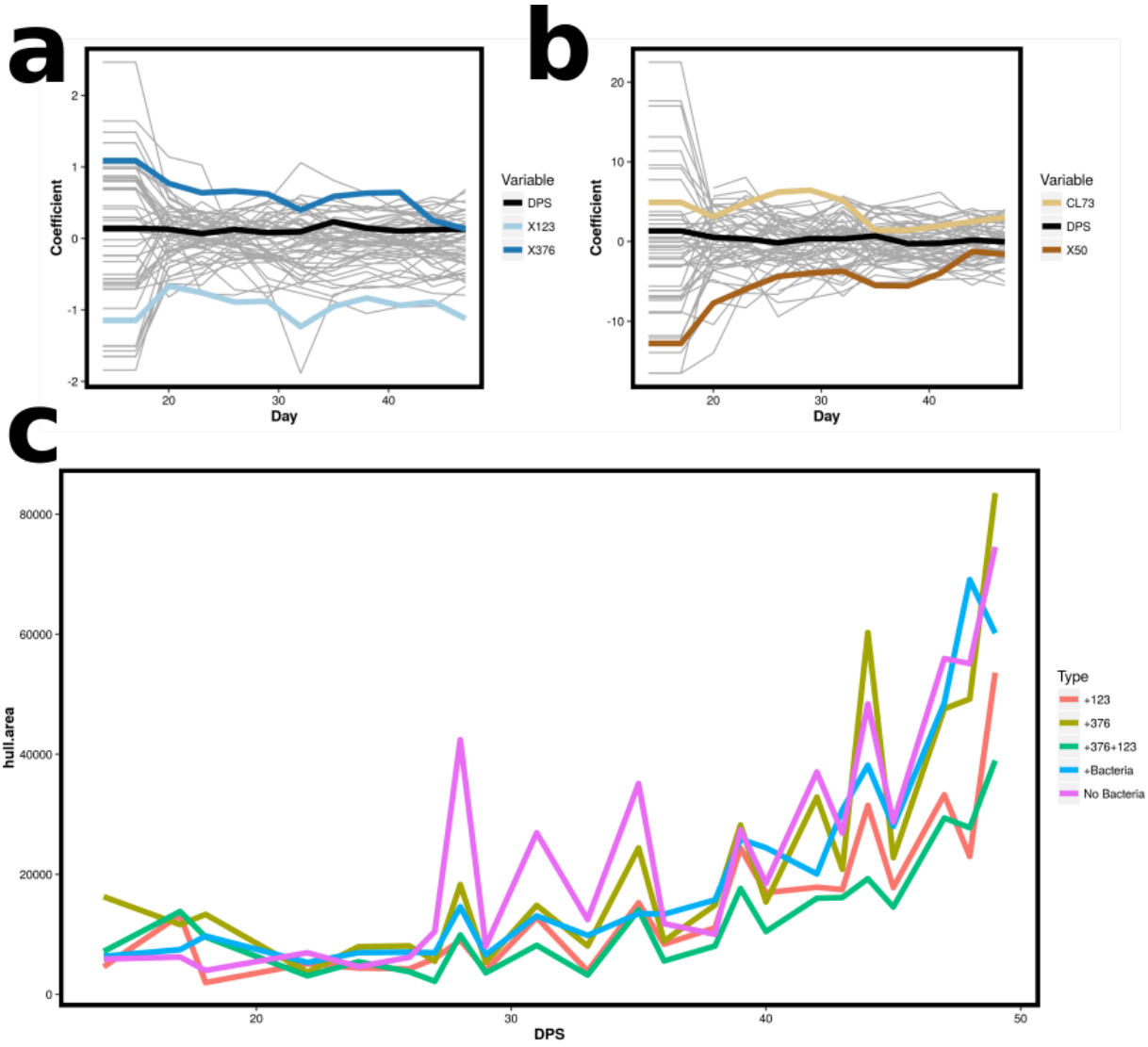


Figure 7.4: **Bacterial effect through time**. a Each line represents a strain and the y-axis value represents the average effect that the corresponding strain had on principal component one of the morphometric characters over experimental time (x-axis). Black line represents the effect of time in that particular window (i.e. the growth rate), and highlighted in blue tones are the two strains with the strongest positive and negative effects.

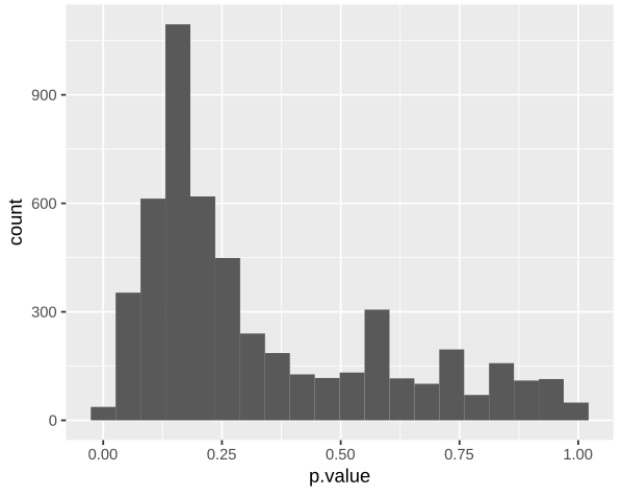


Figure 7.5: Distribution of p-values from associations between phenotypes and KEGG orthology groups.

plant phenotypes. Such interactions could represent either bacteria-bacteria interactions, or epistatic effects between plant pathways that are independently activated by different bacteria. Our previous work has shown that at the level of groups of strains, those groups act mostly additively to influence plant phosphate accumulation, shoot size, and root developmental phenotypes (Chapter 6). However, that work could not distinguish between the absence of interactions and the possibility that intra-group interactions are at least as strong as inter-group interactions. We calculated the effect of all bacteria-bacteria pairwise interactions on both shoot size (using hull area as proxy) and green intensity (section 7.5.7). We found no significant interactions for plant size (hull area), consistent with our previous results, but given the presence of correlated tests (Fig. 7.6a), our multiple testing correction is expected to be conservative. On the other hand, we found a clear signal for pairwise interactions on green intensity (7.6).

Of the 1363 pairwise interactions that happened at least once in our dataset, we found 52 involving 44 strains that were statistically significant after correction for multiple testing (q-value < 0.05; section 7.5.7). We observed that the strains that had the strongest single effects (X50 and CL73) did not interact, since plants with both of them had an intermediate level of greenness (Fig. 7.7). The number of interactions on which each strain participated

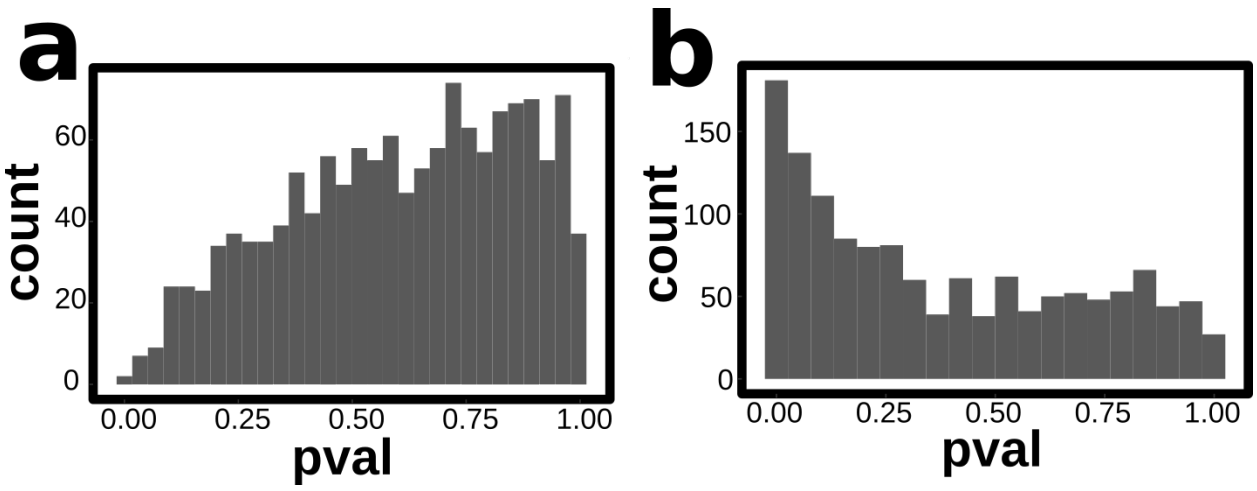


Figure 7.6: **Distributions of p-values for bacteria-bacteria interactions**. a Effect on hull area. b effect on green intensity.

was not random, with a few strains having multiple interactions and most strains having few or none (Fig. 7.7b). Overall there was a similar number of positive (25) and negative (27) pairwise interactions between bacterial strains, and no strain was more likely to have positive or negative interactions (p-value > 0.05 in all cases; hypergeometric test).

There was no obvious phylogenetic enrichment of pairwise interactions, with the top three strains by number interactions belonging to highly divergent groups, namely *Streptomyces* (CL18), *Rhizobium* (X72) and *Pseudomonas* (X50) whose phylogenetic relatedness is depicted in Fig. 7.8a. We decided to test if bacteria-bacteria interactions on plant phenotypes involved a direct interaction between both strains. We tested 86 bacterial pairs for their *in vitro* inhibition activity (section 7.5.8). Of those 86 pairs, ten showed an interaction on plant green intensity, and ten showed an *in vitro* inhibition phenotype in at least one direction. Only 2/86 had both an interactive effect on plant greenness and an *in vitro* inhibition phenotype. Those two cases involved only *Pseudomonas* strains, with both CL58 and X451 inhibiting strain X50 *in vitro*. Interestingly the presence of X50 with either CL58 or X451 was associated with a strong negative effect on plant greenness (Fig. 7.8b-c). Bacteria-bacteria inhibition mechanisms commonly involve toxin-antitoxin systems (Zhang et al., 2012; Jamet and Nassif, 2015). We hypothesize that the negative interaction on plant greenness by specific

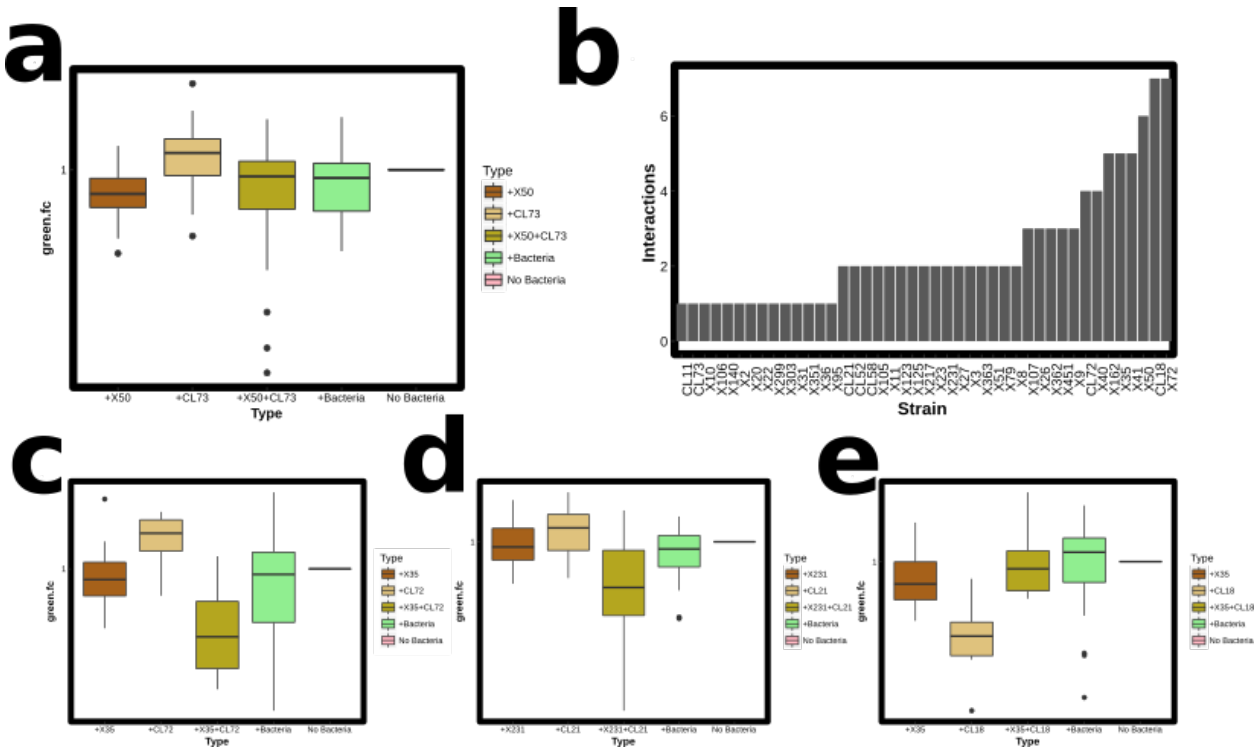


Figure 7.7: **Color is influenced by specific bacterial pairs**. a Lack of interaction between X50 and CL73, the two strains with the strongest opposite effect on plant greenness. b Number of interactions per strain. c-e examples of negative (c-d) and positive interactions between pairs of strains of plant greenness. For a, c, d and e, each box and whiskers plot represents an exclusive set of samples within each panel (i.e. each sample is only on one of the box and whiskers plot). The first set of samples includes all that have strain A but not B, second all that have strain B but not A, third samples that have both strain A and B, and fourth samples that have a synthetic community that didn't include either A nor B. Y-axis is the log fold-change in greenness with respect to the no bacteria plants, which is shown a s a flat line on the fifth position of each plot.

Pseudomonas combinations represent the activation of some of those systems that then either affect other bacteria that are needed for plant greenness or directly affect plant physiology.

Typical binary association assays are performed on agar plates, but they translate poorly into a soil environment (Glick, 2012; Bulgarelli et al., 2013). We tested the three individual strains that had the strongest effect on hull area in our syhtetic community experiments (CL73, X376 and X50). We used a calcined-clay open system in 12-well plates. Each strain was tested individually, and we also tested strain CL21 which has plant growth promoting effect, possibly mediated by increasing phosphate uptake, in agar plate assays (Chapter 6, but did not have a strong efffect on our clay pot system when other bacteria are present (Fig. 7.3 right). We measured rosette fresh weight, and number of leaves. We found that CL73 increases both total plant biomass (Fig. 7.9 left; p-value = 0.00699; ANOVA), and number of leaves (Fig. 7.9 middle; p-value = 0.00642; GLM Poisson). Strains X376, X50 and CL21, all led to larger shoots with more leaves on average, but the effect was not statisticaly silgnificant (Fig. 7.9).

We also observed that the increase in biomass produced by CL73 was not simply due to an increased number of leaves, since the ratio of the two (i.e. the weight per leaf) was also significantly higher in CL73 treated plants (Fig. 7.9 right; p-value = 0.00979; ANOVA). This indicates that CL73 speeds up Arabidopsis shoot development and increases aerial organ size.

7.3 Ongoing validation experiments

We have previously shown that binary association assays are informative regarding how bacteria will influence plant phenotypes in a community context (Chapter 6). However, the correlation was weak and combinations of strains that had the strongest individual effects did not neccessarily lead to the most effective communities. We hypothesize that bacterial effects defined directly from a community context will have better predictive accuracy of novel communities than binary associations.

To that end we designed three partially overlapping synthetic communities by: i) removing the two strains with the strongest positive effect on plant size (X376 and CL73), ii) sorting

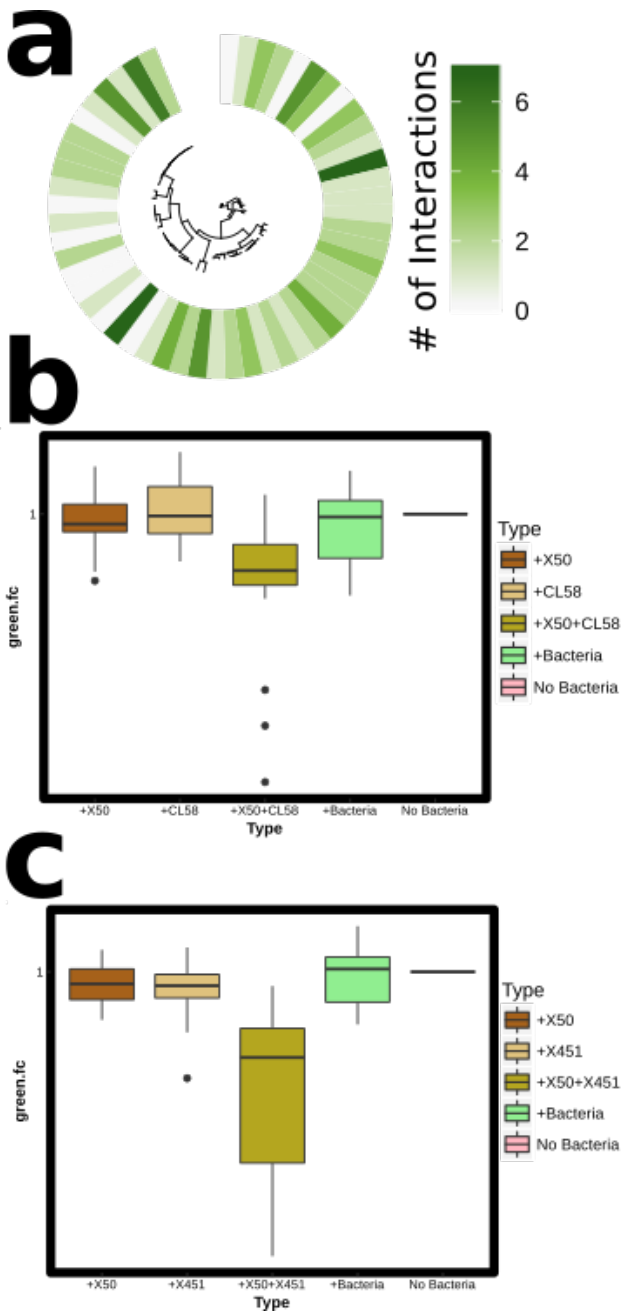


Figure 7.8: **Some interactions might be explained by *in vitro* inhibitions**. **a** Phylogenetic tree of the strains in the main experiment with the number of pairwise interactions on plant color showed by the green scale. **b** Interaction between *Pseudomonas* strains X50 and CL58. **c** Interaction between *Pseudomonas* strains X50 and X451. Box and whisker plots are defined the same way as in Fig. 7.7.

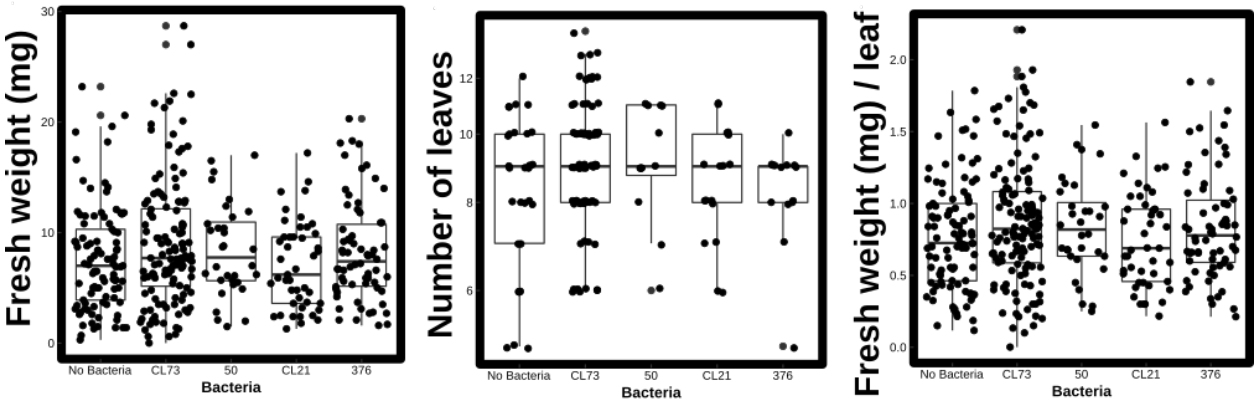


Figure 7.9: **Binary association assays in clay**. **Left**: Shoot fresh weight. **Middle**: number of leaves per seedling. **Right**: Fresh weight per leaf.

the remaining strains according to their effect, and iii) using a seventeen-strain sliding window to build communities that are expected to have positive, indifferent or negative effect on plant size. We named those communities C1, C2 and C3 respectively and we tested each of them in our 12-well plate clay system for their ability to influence plant shoot size. We tested each community alone and with the addition of each of the top two strains in terms of effect on shoot size (X376 and CL73). We have finished the imaging and harvesting, and we are in the middle image analysis, as well as DNA and RNA extraction for bacterial profiling and plant transcriptomics.

We expect that plants inoculated with C1 will be larger than plants inoculated with C3, and we expect that CL73 and X376 will be able to at least partially rescue the size differential between those to communities. We also expect that C2 inoculated plants will have an intermediate size, thus demonstrating that the effects that we estimated directly from communities can be generalized to novel communities

We are also interested in determining whether the positive effect of some bacterial strains is dependent on the abiotic environment. To that end we have taken the most positive community (C1 + X376 + CL73), and inoculated plants with it in our standard 1/4 strength MS media, as well as in a sulfur drop-down media (LowS). We have previously shown that these media has the strongest effect on bacterial abundances among several nutrient drop-

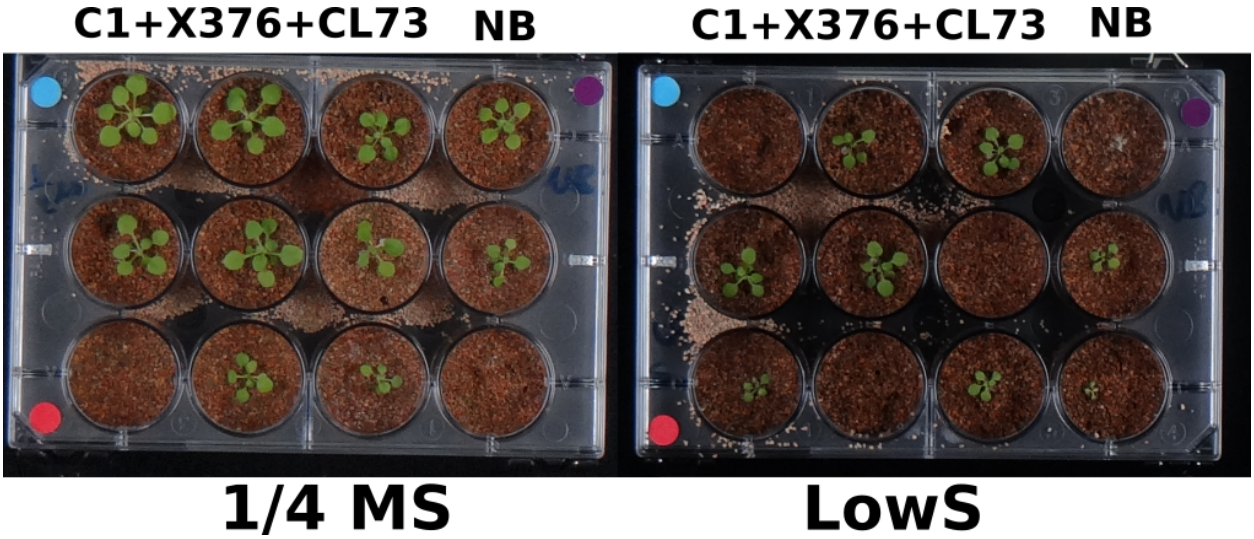


Figure 7.10: **Plants growing with and without synthetic community in two nutrient conditions**. Plants on each plate were supplemented with the nutrient solution indicated in the bottom. For each plate, the first 3 columns were inoculates with a synthetic community (C1 + X376 + CL73), and the last column did not receive any synthetic community. Pictures are taken at 31 days post germination.

downs (Chapter 3). We have previously shown that sulfur drop-down produces a decrease in size of Arabidopsis rosettes. We have finished imaging, harvesting and collecting shoot fresh weight data. We are beginning to extract DNA and RNA for bacterial profiling and plant transcriptomics.

We expect that our bacterial community will be able to at least partially rescue the size defect caused by low sulfur availability. Visual inspection seems to confirm our expectations (Fig. 7.10). We expect that there will be abundance changes in the community that colonizes the plant root and shoot in the media, and we will test if those changes correlate with changes in plant size. Given that our previous results suggest a model of strain stacking in order to modulate plant phenotypes (Chapter 6), and that the final bacterial effects were detectable relatively early (Fig. 7.4), we expect that variation in bacterial relative abundance will be poorly correlated with the effect on the plant, because

7.4 Conclusions

Identification of bacterial strains that modulate host phenotypes such as plant size is typically done via binary association experiments performed in petri-dish conditions, which

are not representative of the natural environment (Bulgarelli et al., 2013). While this has allowed for the decription of several molecular mechanisms (Glick, 2012), their relevance in field conditions remains unclear (Glick, 2012; Bulgarelli et al., 2013). At the very least, these approach igores the presence of a complex biotic background in natural environments. Microbe-microbe interactions, either direct or mediated by other microbes or the host, will limit the applicability of inferences made from binary association studies. We have shown that binary associations are informative for synthetic community function, but the correlation is weak (Chapter 6).

Our combinatorial community design allowed us to estimate the effect of individual strains on multiple plant phenotypes. We showed that different strains alter different plant phenotypes, and that plant size is mostly modulated by individual strain contributions, consistent with our previous results based on bacterial groups (Chapter 6), while plant coloration is affected both by individual strains, and specific pairs of strains. Thus, bacterial strains can act either in isolation, or in concert to influence plant phenotypes.

Our ongoing validation experiments will test whether the estimates of strain effects from community context, are better at predicting the effect of novel communities than binary association assays. We will also test wether the effect of our *designed* communities is maintained in a different context.

As the world population continues to grow, pressure to sustainably increase agricultural output increases. Microbial amendments have generated enormous interest but limited success (Bulgarelli et al., 2013). We have identified microbes that increase shoot size and greeness, a proxy for nitrogen assimilation (Muharam et al., 2015). Our experimental design explicitly favors the identification of microbes that have robust effects across biotic backgrounds. Therefore, our results will allow for fine-tuned and rational design of bacterial consortia

7.5 Methods

7.5.1 Power analysis

Power analysis was conducted in by randomly designing 96 synthetic communities in silico. The communities were of different sizes (five to 20), and strains were drawn from pools of two sizes (24 and 48). Then we assumed that only one focal strain would have a significant effect expressed in standard deviations from the mean (0.5, 1, 1.5, 2, and 3 standard deviations), and we randomly generated plant phenotypes drawn from a normal distribution with mean zero and unit variance. Samples that had the focal strain had phenotype values drawn from a normal distribution with mean equivalent to the effect size (0.5, 1, 1.5, 2, or 3), and unit variance. We tested the resulting phenotype with ANOVA using main effect terms for every strain (24 or 48 terms). We iterated this process on every strain (24 or 48) until each one had been the focal strain, and power was defined as the proportion of focal strains that were correctly captured with a p-value < 0.05. We repeated the analysis three times with three different seed numbers with similar results. Fig. 7.1b-d shows the results from a representative example.

7.5.2 Strain selection

Fifty four strains were selected from a set of Brassicaceae root derived isolates from plants growing in two previously characterized North Carolina soils (Lundberg et al., 2012). Strains were chosen because of the availability of a complete genome and because they represent the four major phyla in root associated microbiomes (Lundberg et al., 2012). For some key genera (Pseudomonas, Rhizobium, Streptomyces, Arthrobacter and Bacillus), multiple close representatives were chosen (from different soils when possible) to test consistency of bacterial effects among related strains. The full list of isolates and their taxon OID that can be used to retrieve their genomes from the Integrated Microbial Database website are provided in table 7.1.

taxon_oid	**ID**	**Genome Name**
2517572231	X2	*Rhizobium sp.* 2MFCol3.1

2513237142	X3	*Pseudomonas sp.* BZ64
2529292577	X8	*Chryseobacterium sp.* UNC8MFCol
2556921097	X9	*Arthrobacter sp.* 9MFCol3.1
2521172663	X10	*Agrobacterium sp.* 10MFCol1.1
2522125170	X11	*Microbacterium sp.* 11MF
2517572232	X20	*Pseudomonas umsongensis* 20MFCvi1.1
2519899668	X22	*Luteibacter sp.* 22Crub2.1
2563366510	X23	*Rhodococcus sp.* UNC23MFCrub1.1
2522125132	X26	*Arthrobacter nicotinovorans* 26Cvi1.1E
2522125133	X27	*Bacillus flexus* 27Col1.1E
2519899686	X31	*Arthrobacter sp.* 31Cvi3.1E
2561511224	X33	*Agrobacterium sp.* 33MFTa1.1
2521172667	X35	*Pseudomonas sp.* 35MFCvi1.1
2521172653	X36	*Pseudomonas mandelii* 36MFCvi1.1
2563366720	X40	*Flavobacterium sp.* 40S8
2563366514	X41	*Bacillus sp.* UNC41MFS5
2519899642	X45	*Pseudomonas sp.* 45MFCol3.1
2519899654	X49	*Arthrobacter sp.* 49Tsu3.1M3
2228664007	X50	*Pseudomonas sp.* KD5
2510065054	X51	*Pseudomonas brassicacearum* 51MFCVI2.1
2228664006	X57	*Rhizobium sp.* 57MFTsu3.2
2510065092	X72	*Rhizobium sp.* IBUN
2556921674	X79	*Dyella japonica* UNC79MFTsu3.2
2517572209	X95	*Bacillus sp.* 95MFCvi2.1
2517572206	X105	*Bacillus sp.* 105MF
2522125150	X106	*Bacillus sp.* 171095_106
2522125078	X107	*Bacillus sp.* 278922_107

2521172627	X123	*Bacillus sp.* 123MFChir2
2561511073	X125	*Bacillus sp.* UNC125MFCrub1.1
2517572123	X135	*Arthrobacter sp.* 135MFCol5.1
2563366508	X140	*Streptomyces sp.* 140Col2.1E
2517572124	X161	*Arthrobacter sp.* 161MFSha2.1
2517572214	X162	*Arthrobacter sp.* 162MFSha1.1
2563366516	X217	*Paenibacillus sp.* UNC217MF
2521172624	X224	*Agrobacterium sp.* 224MFTsu3.1
2523533508	X231	*Arthrobacter nicotinovorans* 231Sha2.1M6
2521172643	X299	*Streptomyces canus* 299MFChir4.1
2521172626	X303	*Streptomyces sp.* 303MFCol5.2
2522572130	X327	*Promicromonospora sukumoe* 327MFSha3.1
2521172628	X351	*Streptomyces sp.* 351MFTsu5.1
2563366511	X362	*Arthrobacter sp.* UNC362MFTsu5.1
2563366512	X363	*Rhodococcus sp.* UNC363MFTsu5.1
2521172625	X376	*Burkholderia bryophila* 376MFSha3.1
2546825545	X384	*Burkholderia* MF384
2563366509	X451	*Paenibacillus sp.* UNC451MF
2546825541	CL11	*Burkholderia* CL11
2563366515	CL18	*Streptomyces sp.* UNC401CLCol
2558309150	CL21	*Ralstonia sp.* UNC404CL21Col
2529292583	CL41	*Agrobacterium sp.* UNC420CL41Cvi
2563366513	CL52	*Paenibacillus sp.* UNCCL52
2556921015	CL58	*Pseudomonas umsongensis* UNC430CL58Col
2529292578	CL72	*Bacillus sp.* UNC437CL72CviS29
2528768222	CL73	*Bacillus sp.* UNC438CL73TsuS30

Table 7.1: Isolates used in this study.

7.5.3 Bacterial growth for synthetic communities

For each independent synthetic community experiment, bacteria were plated in LB media from glycerol stocks until the appearance of single colonies. Single colonies were used to inoculate 4-6 liquid cultures of 2xYT media that were grown for four days at 28°C. After four days, the technical replicates of the liquid culture were combined to buffer variability and the cells were washed twice with MES buffer (pH 6). Bacteria were mixed according to a randomized design using a liquid handling robot that mixed equal volumes (20μL) of each of the seventeen strains per community into a 96-well plate.

7.5.4 Plant growth for synthetic communities

Seeds were surface sterilized by washing twice with 70% ethanol and 0.1% Triton-X for 1 minute. Then they were suspended in 20% household bleach with 0.1% Triton-X for 15 minutes. Seeds were then washed five times with sterile water, re-suspended in sterile water and stratified in the dark at 4°C for three days.

Autoclaved 4in^2 calcined-clay (Diamond Pro Red Infield Conditioner) pots were then inoculated with 40mL of a 1/4 strength MS media with or without bacteria added. Seeds were then sowed on top of the calcined-clay from their water suspension averaging 6 seeds per pot. Pots were placed in flats (12 pots per flat) and flats were covered with transparent plastic lids. Plants were then transferred to a growth chamber with short day (8 hrs light, 16 hrs dark, at 21/18°C) where they were kept for the remainder of the experiment.

After two weeks, pots were thinned to one plant per pot, by keeping the largest seedling. Pots continued to be watered as needed with sterile distilled water from the top.

7.5.5 Image based phenotyping

During the experiments, each plant was imaged every 2-3 days on a professional camera stand with standard height, settings and lighting in a dark room. Each imaging session

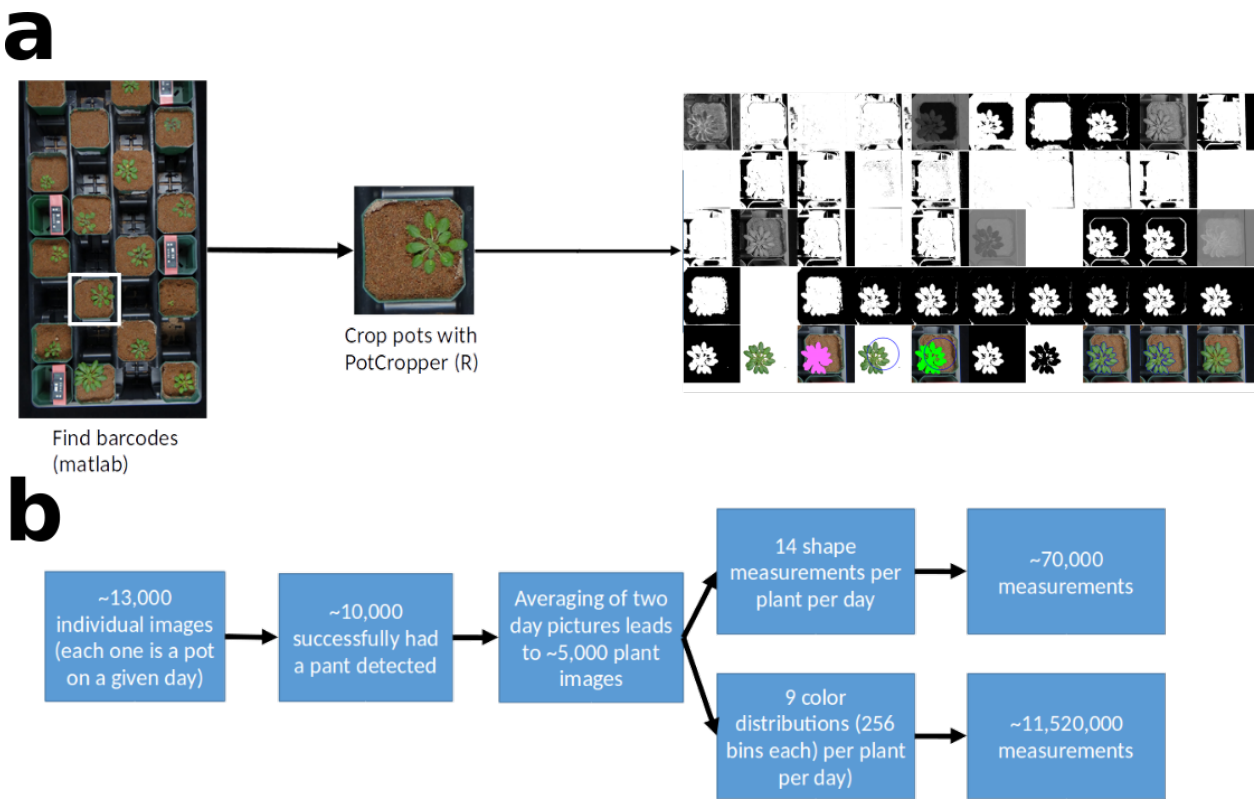


Figure 7.11: **Imaging pipeline**. a Schematic representation of the image-based phenotyping pipeline. b Number of data points obtained in this study.

individual plants were imaged 2 times each.

We implemented an image-based phenotyping platform based on PlantCV (Fahlgren et al., 2015). Briefly each image is cropped to individual pots and then PlantCV is used with custom thresholding settings to identify the region of the image that corresponds to a plant. Then all available morphometric characteristics are extracted as well as color distributions.

Phenotypes from independent images on the same session where averaged and the median of the color distributions was used as an indicator of its value.

Data is available in the combinatorix R package that will be made public upon submission for publication.

7.5.6 Estimating main effects

Standard linear models assume that observations are independent. This assumption does not hold in the case of repeated measurements on the same individual as is the case in our

time course data. Generalized least squares (GLS) models allow us to relax that assumption by specifying a correlation structure.

We tested various versions of the Autoregressive-Moving-Average model (ARMA) family, and compared them to the simpler compound correlation structure (sometimes called exchangeable correlation structure). We used the Akaike Information Criteria (Akaike, 1974) to compare the models and determined that the compound correlation structure was the best for our dataset. In this type of structure, measurements from the same plant have a correlation (ρ) that is constant for all time-points and for all individual plants. The correlation value is an extra parameter that is estimated from the data.

We modelled each phenotype separately. Morphometric phenotypes were log transformed to reduce heteroscedasticity. Color phenotypes were not transformed. We estimated the main effects of all strains simultaneously by fitting all the data from a given phenotype with a model that had one coefficient per strain, as well as terms for time and biological replicate, as indicated by the following equation:

$$y_{ij} = \beta_0 + \mathbf{D_j s} + \beta_{exp_j} + \beta_{time} t_{ij} \tag{7.1}$$

where y_{ij} is the i-th observation of the j-th plant, β_0 is the model intercept, $\mathbf{D}$ is the $n \times 54$ indicator matrix that defines which strains went into each sample, and $\mathbf{D_j}$ is the 1×54 vector indicating which strains went into sample j, $\mathbf{s}$ is the 54×1 vector of coefficients corresponding to each of the 54 strains, β_{exp_j} is the effect of experiment of sample j, β_{time} is the effect of time, and $t_i j$ is the time after sowing (in days) of the i-th observation of the j-th plant.

The correlation structure implies that the correlation between two observations depends on whether they came from the same plant, and it is defined by the following equation:

$$cor(y_{ij}, y_{kh}) = \begin{cases} \rho, & \text{if} j = h \\ 0, & \text{if} j \neq h \end{cases} \tag{7.2}$$

We fit this model and the correlation structure with the gls function from the R nlme package.

The code and data used for this analysis is available in the combinatorix R package that will be made available upon submission for publication.

7.5.7 Estimating interactions

To identify interactions, we focused on the hull area and plant green intensity. The generalized least squares approach cannot be easily utilized because there are over 1300 possible pairwise interactions and fitting a linear model with that many parameters where there are only ∼400 independent samples would result in massive loss of power and overfitting. We instead decided to test each pair of strains separately.

For every pair of strains (A and B) we classify all observations as belonging to one of the five following mutually exclusive *groups*: i) samples with strain A but not B as part of a synthetic community, ii) samples with strain B but not A as a part of a synthetic community, iii) samples with both strain A and B as part of a synthetic community, iv) samples with a synthetic community that contained neither strain A nor B, and v) samples that didn't receive a synthetic community.

Then we calculated the mean phenotypic value per combination of *group* and plant age. Because older plant are bigger, we normalized every resulting average by dividing it by the value of the fifth *group* (the no synthetic community group), and log transformed the ratio. In other words, we calculated the fold-change in phenotype with respect to no synthetic community for each *group* and plant age.

Because the resulting phenotypic values for the no synthetic community group are always 1, they are removed from the following statistical test. We perform a likelihood ratio test where we compare two ANOVA models that both contain terms for the presence of strain A

and B, but that either include or do not include an interaction term. These two models are defined with the following formulas:

$$Phenotype = strain_A + strain_B \tag{7.3}$$

$$Phenotype = strain_A + strain_B + strain_{AB} \tag{7.4}$$

Pairs of strains that have a significant p-value after correcting for multiple testing (Benjamini and Hochberg, 1995), are deemed as significant interactions.

The code and data used for this analysis is available in the combinatorix R package that will be made available upon submission for publication.

7.5.8 *In vitro* inhibitions

Bacteria bacteria inhibition assays were performed by growing each strain individually in liquid 2xYT media at 28°C. Then strains were normalized to 10^5 c.f.u./mL, assuming that optical density at 600nm of 1 is equivalent to 10^9c.f.u./mL, and resuspended in MES buffer (pH 6). A lawn of a single strain was created by spreading 200μL of one of the strain into an one tenth strength LB agar plate with sterilized glass beads. After that, 20μL of another bacteria were spotted on top of the agar lawn. Plates were sealed and incubated at 28°C, and visually inspected at 2, 5 and 7 days for a clearing on the bacterial lawn. Strains that produced a clearing on the lawn were marked as positive inhibitors for the lawn bacteria.

The data gathered, and the code to compare it to the pairwise interactions on plant phenotypes, is available in the combinatorix R package that will be made available upon submission for publication.

7.5.9 Randomization and experimental blinding

For the main synthetic community experiments, bacteria were randomized and the experimenters were blind to which samples contained which bacteria.

7.5.10 Data and software accessibility

All the the code and data presented is available in the combinatorix R package that will be made available upon submission for publication.

REFERENCES

Abramoff, M. D., Magalhaes, P. J., and Ram, S. J. (2004). Image processing with ImageJ. Biophotonics Int. 11: 36–42.

Agler, M. T., Ruhe, J., Kroll, S., Morhenn, C., Kim, S.-T., Weigel, D., and Kemen, E. M. (2016). Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLOS Biology*, 14(1):e1002352.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Ames, B. N. (1966). [10] Assay of inorganic phosphate, total phosphate and phosphatases. *Methods in Enzymology*, 8(C):115–118.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31(2):166–169.

Anderson, M. J. and Willis, T. J. (2003). Canonical Analysis of Principal coordinates: A Useful Method of Constrained Ordination For Ecology. *Ecology*, 84(2):511–525.

Angermueller, C., Pärnamaa, T., Parts, L., and Oliver, S. (2016). Deep Learning for Computational Biology. *Molecular Systems Biology*, (12):878.

Armanhi, J. S. L., de Souza, R. S. C., de Araújo, L. M., Okura, V. K., Mieczkowski, P., Imperial, J., and Arruda, P. (2016). Multiplex amplicon sequencing for microbe identification in community-based culture collections. *Scientific Reports*, 6:29543.

Arsenault, J. L., Poulcur, S., Messler, C., and Guay, R. (1996). WinRHlZO, a Root-measuring System with a Unique Overlap Correction Method. *HortScience*, 30(4):906.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., de Vos, W. M., Brunak, S., Doré, J., Antolín, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., van de Guchte, M., Guedon, E., Haimet, F., Huber, W., van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D., and Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature*, 473(7346):174–80.

Ash, C. and Allen, O. (1948). A comparison of Methods Recommended for the Surface Sterilization of Leguminous Seed. *Soil Science Society Proceedings*, pages 279–283.

Aviv, D. H., Rustérucci, C., Holt III, B. F., Dietrich, R. A., Parker, J. E., and Dangl, J. L. (2002). Runaway cell death , but not basal disease resistance , in lsd1 is SA- and NIM1 / NPR1 -dependent. *The Plant Journal*, 29(3):381–391.

Bai, Y., Müller, D. B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P. C., Spaepen, S., Remus-Emsermann, M., Hüttel, B., McHardy, A. C., Vorholt, J. A., and Schulze-Lefert, P. (2015). Functional overlap of the Arabidopsis leaf and root microbiota. *Nature*, 528(7582):364–369.

Bakker, P. a. H. M., Berendsen, R. L., Doornbos, R. F., Wintermans, P. C. a., and Pieterse, C. M. J. (2013). The rhizosphere revisited: root microbiomics. *Frontiers in Plant Science*, 4(May):1–7.

Bálint, M., Tiffin, P., Hallström, B., O'Hara, R. B., Olson, M. S., Fankhauser, J. D., Piepenbring, M., Schmitt, I., O'Hara, R. B., Olson, M. S., Fankhauser, J. D., Piepenbring, M., and Schmitt, I. (2013). Host Genotype Shapes the Foliar Fungal Microbiome of Balsam Poplar (Populus balsamifera). *PloS one*, 8(1):e53987.

Barboriak, D. P., Padua, A. O., York, G. E., and MacFall, J. R. (2005). Creation of DICOM - Aware applications using ImageJ. *Journal of Digital Imaging*, 18(2):91–99.

Barret, M., Briand, M., Bonneau, S., Préveaux, A., Valière, S., Bouchez, O., Hunault, G., Simoneau, P., and Jacques, M.-A. (2015). Emergence Shapes the Structure of the Seed Microbiota. *Applied and Environmental Microbiology*, 81(4):1257–1266.

Bates, D. M. (2010). *lme4: Mixed-effects modeling with R*.

Belkhadir, Y., Yang, L., Hetzel, J., Dangl, J. L., and Chory, J. (2014). The growth-defense pivot: Crisis management in plants mediated by LRR-RK surface receptors. *Trends in Biochemical Sciences*, 39(10):447–456.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Poweful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Benson, A. K., Kelly, S. a., Legge, R., Ma, F., Low, S. J., Kim, J., Zhang, M., Oh, P. L., Nehrenberg, D., Hua, K., Kachman, S. D., Moriyama, E. N., Walter, J., Peterson, D. a., and Pomp, D. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences*, 107(44):18933–18938.

Berendsen, R. L., van Verk, M. C., Stringlis, I. a., Zamioudis, C., Tommassen, J., Pieterse, C. M. J., and Bakker, P. a. H. M. (2015). Unearthing the genomes of plant-beneficial Pseudomonas model strains WCS358, WCS374 and WCS417. *BMC Genomics*, 16(1):539.

Bi, Y.-M., Kenton, P., Mur, L., Darby, R., and Draper, J. (1995). Hydrogen peroxide does not function downstream of salicylic acid in the induction of PR protein expression. *The Plant Journal*, 8(2):235–245.

Biswas, C., Dey, P., Satpathy, S., Sarkar, S. K., Bera, a., and Mahapatra, B. S. (2013). A simple method of DNA isolation from jute (Corchorus olitorius) seed suitable for PCR-based detection of the pathogen Macrophomina phaseolina (Tassi) Goid. *Letters in applied microbiology*, 56(2):105–10.

Blakney, A. J. C. and Patten, C. L. (2011). A plant growth-promoting pseudomonad is closely related to the Pseudomonas syringae complex of plant pathogens. *FEMS Microbiology Ecology*, 77(3):546–557.

Bodenhausen, N., Bortfeld-Miller, M., Ackermann, M., and Vorholt, J. A. (2014). A Synthetic Community Approach Reveals Plant Genotypes Affecting the Phyllosphere Microbiota. *PLoS Genetics*, 10(4):e1004283.

Bodenhausen, N., Horton, M. W., and Bergelson, J. (2013). Bacterial Communities Associated with the Leaves and the Roots of Arabidopsis thaliana. *PLoS ONE*, 8(2):e56329.

Bonardi, V., Tang, S., Stallmann, A., Roberts, M., Cherkis, K., and Dangl, J. L. (2011). Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. *Proceedings of the National Academy of Sciences*, 108(39):16463–16468.

Bonkowski, M. (2004). Protozoa and plant growth: The microbial loop in soil revisited. *New Phytologist*, 162(3):617–631.

Bouffaud, M. L., Poirier, M. A., Muller, D., and Moënne-Loccoz, Y. (2014). Root microbiome relates to plant host evolution in maize and other Poaceae. *Environmental Microbiology*, 16:2804–2814.

Bowling, S. A. (1994). A Mutation in Arabidopsis That Leads to Constitutive Expression of Systemic Acquired Resistance. *THE PLANT CELL ONLINE*, 6(12):1845–1857.

Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., Goulding, D., and Lawley, T. D. (2016). Culturing of unculturable' human microbiota reveals novel taxa and extensive sporulation. *Nature*, 533(7604):in press.

Brundrett, M. C. (2009). Mycorrhizal associations and other means of nutrition of vascular plants: understanding the global diversity of host plants by resolving conflicting information and developing reliable means of diagnosis. *Plant and Soil*, 320:37–77.

Bruto, M., Prigent-Combaret, C., Muller, D., and Moënne-Loccoz, Y. (2014). Analysis of genes contributing to plant-beneficial functions in Plant Growth-Promoting Rhizobacteria and related Proteobacteria. *Scientific reports*, 4:6261.

Bulgarelli, D., Garrido-Oter, R., Münch, P. C., Weiman, A., Dröge, J., Pan, Y., McHardy, A. C., and Schulze-Lefert, P. (2015). Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley. *Cell Host & Microbe*, 17(3):392–403.

Bulgarelli, D., Rott, M., Schlaeppi, K., van Themaat, E. V. L., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., Peplies, J., Gloeckner, F. O., Amann, R., Eickhorst, T., and Schulze-Lefert, P. (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature*, 488(7409):91–95.

Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E., Schulze-Lefert, P., van Themaat, E. V. L., and Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants. *Annual Review of Plant Biology*, 64(1):807–38.

Buscot, F. (2015). Implication of evolution and diversity in arbuscular and ectomycorrhizal symbioses. *Journal of Plant Physiology*, 172:55–61.

Bustos, R., Castrillo, G., Linhares, F., Puga, M. I., Rubio, V., Pérez-Pérez, J., Solano, R., Leyva, A., and Paz-Ares, J. (2010). A central regulatory system largely controls transcriptional activation and repression responses to phosphate starvation in arabidopsis. *PLoS Genetics*, 6(9).

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. (2009). BLAST plus: architecture and applications. *BMC Bioinformatics*, 10(421):1.

Cao, H., Glazebrook, J., Clarke, J. D., Volko, S., and Dong, X. (1997). The Arabidopsis NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankyrin repeats. *Cell*, 88(1):57–63.

Cao, W.-H., Liu, J., He, X.-J., Mu, R.-L., Zhou, H.-L., Chen, S.-Y., and Zhang, J.-S. (2007). Modulation of Ethylene Responses Affects Plant Salt-Stress Responses. *PLANT PHYSIOLOGY*, 143(2):707–719.

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. a., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. a., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. a., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7(5):335–336.

Caporaso, J. G., Lauber, C. L., Walters, W. a., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S. M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J. a., Smith, G., and Knight, R. (2012). Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME journal*, 6(8):1621–1624.

Cardinale, M., Grube, M., Erlacher, A., Quehenberger, J., and Berg, G. (2015). Bacterial networks and co-occurrence relationships in the lettuce root microbiota. *Environmental Microbiology*, 17(1):239–252.

Carvalhais, L. C., Dennis, P. G., and Schenk, P. M. (2014). Plant defence inducers rapidly influence the diversity of bacterial communities in a potting mix. *Applied Soil Ecology*, 84:1–5.

Castrillo, G., Sánchez-Bermejo, E., de Lorenzo, L., Crevillén, P., Fraile-Escanciano, A., Tc, M., Mouriz, A., Catarecha, P., Sobrino-Plata, J., Olsson, S., Leo Del Puerto, Y., Mateos, I., Rojo, E., Hernández, L. E., Jarillo, J. A., Piñeiro, M., Paz-Ares, J., and Leyva, A. (2013). WRKY6 transcription factor restricts arsenate uptake and transposon activation in Arabidopsis. *The Plant cell*, 25(8):2944–57.

Castrillo, G., Teixeira, P. J. P. L., Herrera Paredes, S., Law, T. F., de Lorenzo, L., Feltcher, M. E., Finkel, O. M., Breakfield, N. W., Mieczkowski, P., Jones, C. D., Paz-Ares, J., and Dangl, J. L. (2017). Root microbiota drive direct integration of phosphate stress and immunity. *Nature*, 543(7646):513–518.

Chi, F., Shen, S.-h., Cheng, H.-p., Jing, Y.-x., Yanni, Y. G., Dazzo, F. B., Chi, F., Shen, S.-h., Cheng, H.-p., Jing, Y.-x., Yanni, Y. G., and Dazzo, F. B. (2005). Ascending Migration of Endophytic Rhizobia , from Roots to Leaves , inside Rice Plants and Assessment of Benefits to Rice Growth Physiology. *Applied and Environmental Microbiology*, 71(11):7271–7278.

Chiu, H. C., Levy, R., and Borenstein, E. (2014). Emergent Biosynthetic Capacity in Simple Microbial Communities. *PLoS Computational Biology*, 10(7).

Clarholm, M. (1985). Interactions of bacteria, protozoa and plants leading to mineralization of soil nitrogen. *Soil Biology and Biochemistry*, 17(2):181–187.

Clarke, J. D. (2000). Roles of Salicylic Acid, Jasmonic Acid, and Ethylene in cpr-Induced Resistance in Arabidopsis. *THE PLANT CELL ONLINE*, 12(11):2175–2190.

Clarke, J. D., Liu, Y., Klessig, D. F., and Dong, X. (1998). Uncoupling PR gene expression from NPR1 and bacterial resistance: characterization of the dominant Arabidopsis cpr6-1 mutant. *Plant Cell*, 10(April):557–569.

Conrath, U., Beckers, G. J. M., Flors, V., García-Agustín, P., Jakab, G., Mauch, F., Newman, M.-A., Pieterse, C. M. J., Poinssot, B., Pozo, M. J., Pugin, A., Schaffrath, U., Ton, J., Wendehenne, D., Zimmerli, L., and Mauch-Mani, B. (2006). Priming: Getting Ready for Battle. *Molecular Plant-Microbe Interactions*, 19(10):1062–1071.

Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M., and Relman, D. a. (2012). The Application of Ecological Theory Toward an Understanding of the Human Microbiome. *Science*, 336(6086):1255–1262.

Coyte, K. Z., Schluter, J., and Foster, K. R. (2015). The ecology of the microbiome: Networks, competition, and stability. *Science*, 350(6261):663–666.

De Deyn, G. B., Cornelissen, J. H. C., and Bardgett, R. D. (2008). Plant functional traits and soil carbon sequestration in contrasting biomes. *Ecology Letters*, 11(5):516–531.

de Lorenzo, L. and Paz-Ares, J. (2017). PHOSPHATE STARVATION RESPONSE 1 acts via two cis-motifs and links plant water content with phosphate homeostasis. *in preparation*.

DeFraia, C. T., Schmelz, E. A., Mou, Z., Cleand, C., Ajami, A., Shakirova, F., Sakhabutdinova, A., Bezrukova, M., Fatkhutdinova, R., Fatkhutdinova, D., Raskin, I., Ehmann, A., Melander, W., Meeuse, B., Rhoads, D., McIntosh, L., White, R., Durrant, W., Dong, X., Raskin, I., Ryals, J., Neuenschwander, U., Willits, M., Molina, A., Steiner, H.-Y., Hunt, M., Malamy, J., Carr, J., Klessig, D., Raskin, I., Métraux, J., Signer, H., Ryals, J., Ward, E., Wyss-Benz, M., Gaudin, J., Raschdorf, K., Schmid, E., Blum, W., Inverardi, B., Enyedi, A., Raskin, I., Lee, H.-I., León, J., Raskin, I., Zhang, X., Dai, Y., Xiong, Y., Defraia, C., Li, J., Dong, X., Mou, Z., Petersen, M., Brodersen, P., Naested, H., Andreasson, E., Lindhart, U., Johansen, B., Nielsen, H., Lacy, M., Austin, M., Parker, J., Bowling, S., Guo, A., Cao, H., Gordon, A., Klessig, D., Dong, X., Clarke, J., Liu, Y., Klessig, D., Dong, X., Shah, J., Kachroo, P., Nandi, A., Klessig, D., Zhang, Y., Goritschnig, S., Dong, X., Li, X., Shirano, Y., Kachroo, P., Shah, J., Klessig, D., Heidel, A., Clarke, J., Antonovics, J., Dong, X., Nawrath, C., Métraux, J.-P., Dewdney, J., Reuber, T., Wildermuth, M., Devoto, A., Cui, J., Stutius, L., Drummond, E., Ausubel, F., Nawrath, C., Heck, S., Parinthawong, N., Métraux, J.-P., Wildermuth, M., Dewdney, J., Wu, G., Ausubel, F., Aboul-Soud, M., Cook, K., Loake, G., Malamy, J., Klessig, D., Schmelz, E., Engelberth, J., Alborn, H., O'Donnell, P., Sammons, M., Toshima, H., Tumlinson, J., Huang, W., Wang, H., Zheng, H., Huang, L., Singer, A., Thompson, I., Whiteley, A., Huang, W., Huang, L., Preston, G., Martin, N., Carr, J., Yanhong, L., Singer, A., Whiteley, A., Hui, W., Enyedi, A., Yalpani, N., Silverman, P., Raskin, I., Wiseman, A., Lee, J., Nam, J., Park, H., Na, G., Miura, K., Jin, J., Yoo, C., Baek, D., Kim, D., Jeong, J., Ishikawa, A., Kimura, Y., Yasuda, M., Nakashita, H., Yoshida, S., Nandi, A., Krothapalli, K., Buseman, C., Li, M., Welti, R., Enyedi, A., Shah, J., Zheng, Z., Mosher, S., Fan, B., Klessig, D., Chen, Z., Gupta, V., Willits, M., Glazebrook, J., Glazebrook, J., Chen, W., Estes, B., Chang, H.-S., Nawrath, C., Métraux, J.-P., Zhu, T., Katagiri, F., Eshita, S., Meuwly, P., and Metraux, J. (2008). A rapid biosensor-based method for quantification of free and glucose-conjugated salicylic acid. *Plant Methods*, 4(1):28.

Dennis, P. G., Miller, A. J., and Hirsch, P. R. (2010). Are root exudates more important than other sources of rhizodeposits in structuring rhizosphere bacterial communities? *FEMS Microbiology Ecology*, 72(3):313–327.

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006). Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, 72(7):5069–5072.

Desirò, A., Salvioli, A., Ngonkeu, E. L., Mondo, S. J., Epis, S., Faccio, A., Kaech, A., Pawlowska, T. E., and Bonfante, P. (2014). Detection of a novel intracellular microbiome hosted in arbuscular mycorrhizal fungi. *The ISME journal*, 8(2):257–70.

Dewdney, J., Lynne Reuber, T., Wildermuth, M. C., Devoto, A., Cui, J., Stutius, L. M., Drummond, E. P., and Ausubel, F. M. (2000). Three unique mutants of Arabidopsis identify eds loci required for limiting growth of a biotrophic fungal pathogen. *Plant Journal*, 24(2):205–218.

Dodds, P. N. and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plantpathogen interactions. *Nature Reviews Genetics*, 11(8):539–548.

Dombrowski, N., Schlaeppi, K., Agler, M. T., Hacquard, S., Kemen, E., Garrido-Oter, R., Wunder, J., Coupland, G., and Schulze-Lefert, P. (2017). Root microbiota dynamics of perennial Arabis alpina are dependent on soil residence time but independent of flowering time. *The ISME Journal*, 11(1):43–55.

Doornbos, R. F., Geraats, B. P. J., Kuramae, E. E., Van Loon, L. C., and Bakker, P. A. H. M. (2011). Effects of Jasmonic Acid, Ethylene, and Salicylic Acid Signaling on the Rhizosphere Bacterial Community of Arabidopsis thaliana. *Molecular Plant-Microbe Interactions*, 24(4):395–407.

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

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

Edgar, R. C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*, 10(10):996–998.

Edwards, J., Johnson, C., Santos-Medellín, C., Lurie, E., Podishetty, N. K., Bhatnagar, S., Eisen, J. A., Sundaresan, V., and Kumar, N. (2015). Structure , variation , and assembly of the root-associated microbiomes of rice. *Proceedings of the National Academy of Sciences*, 112(8):E911–E920.

Eickhorst, T. and Tippkötter, R. (2008). Improved detection of soil microorganisms using fluorescence in situ hybridization (FISH) and catalyzed reporter deposition (CARD-FISH). *Soil Biology and Biochemistry*, 40(7):1883–1891.

Elmer, W. H. (2001). Seeds as vehicles for pathogen importation. *Biological Invasions*, 3(3):263–271.

Embree, M., Liu, J. K., Al-Bassam, M. M., and Zengler, K. (2015). Networks of energetic and metabolic interactions define dynamics in microbial communities. *Proceedings of the National Academy of Sciences*, 112(50):15450–15455.

Engelbrektson, A., Kunin, V., Wrighton, K. C., Zvenigorodsky, N., Chen, F., Ochman, H., and Hugenholtz, P. (2010). Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME journal*, 4(5):642–647.

Ewald, P. (1988). Cultural vectors, virulence, and the emergence of evolutionary epidemiology. *Oxford Surveys in Evolutionary Biology*, 5:215–244.

Fahlgren, N., Feldman, M., Gehan, M. A., Wilson, M. S., Shyu, C., Bryant, D. W., Hill, S. T., McEntee, C. J., Warnasooriya, S. N., Kumar, I., Ficor, T., Turnipseed, S., Gilbert, K. B., Brutnell, T. P., Carrington, J. C., Mockler, T. C., and Baxter, I. (2015). A versatile phenotyping system and analytics platform reveals diverse temporal responses to water availability in Setaria. *Molecular Plant*, 8(10):1520–1535.

Faith, J. J., Ahern, P. P., Ridaura, V. K., Cheng, J., and Gordon, J. I. (2014). Identifying Gut Microbe-Host Phenotype Relationships Using Combinatorial Communities in Gnotobiotic Mice. *Science Translational Medicine*, 6(220):220ra11–220ra11.

Faith, J. J., Rey, F. E., O'Donnell, D., Karlsson, M., McNulty, N. P., Kallstrom, G., Goodman, A. L., and Gordon, J. I. (2010). Creating and characterizing communities of human gut microbes in gnotobiotic mice. *The ISME journal*, 4(9):1094–1098.

Feehery, G. R., Yigit, E., Oyola, S. O., Langhorst, B. W., Schmidt, V. T., Stewart, F. J., Dimalanta, E. T., Amaral-Zettler, L. a., Davis, T., Quail, M. a., and Pradhan, S. (2013). A Method for Selectively Enriching Microbial DNA from Contaminating Vertebrate Host DNA. *PLoS ONE*, 8(10).

Ferraroni, M., Matera, I., Bürger, S., Reichert, S., Steimer, L., Scozzafava, A., Stolz, A., and Briganti, F. (2013). The salicylate 1,2-dioxygenase as a model for a conventional gentisate 1,2-dioxygenase: Crystal structures of the G106A mutant and its adducts with gentisate and salicylate. *FEBS Journal*, 280(7):1643–1652.

Fiehn, O., Timothy Garvey, W., Newman, J. W., Lok, K. H., Hoppel, C. L., and Adams, S. H. (2010). Plasma metabolomic profiles reflective of glucose homeostasis in non-diabetic and type 2 diabetic obese African-American women. *PLoS ONE*, 5(12):1–10.

Fiehn, O., Wohlgemuth, G., Scholz, M., Kind, T., Lee, D. Y., Lu, Y., Moon, S., and Nikolau, B. (2008). Quality control for plant metabolomics: Reporting MSI-compliant studies. *Plant Journal*, 53(4):691–704.

Firáková, S., Šturdíková, M., and Múčková, M. (2007). Bioactive secondary metabolites produced by microorganisms associated with plants. *Biologia*, 62(3):251–257.

Flor, H. H. (1971). Current Status of the Gene-For-Gene Concept. *Annual Review of Phytopathology*, 9(1):275–296.

Franche, C., Lindström, K., and Elmerich, C. (2009). Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. *Plant and Soil*, 321:35–59.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal Of Statistical Software*, 33(1):1–22.

Fu, Z. Q. and Dong, X. (2013). Systemic Acquired Resistance: Turning Local Infection into Global Defense. *Annual Review of Plant Biology*, 64(1):839–863.

Geva-Zatorsky, N., Sefik, E., Kua, L., Pasman, L., Tan, T. G., Ortiz-Lopez, A., Yanortsang, T. B., Yang, L., Jupp, R., Mathis, D., Benoist, C., and Kasper, D. L. (2017). Mining the Human Gut Microbiota for Immunomodulatory Organisms. *Cell*, 0(0):1–16.

Gilbert, J. A., Quinn, R. A., Debelius, J., Xu, Z. Z., Morton, J., Garg, N., Jansson, J. K., Dorrestein, P. C., and Knight, R. (2016). Microbiome-wide association studies link dynamic microbial consortia to disease. *Nature*, 535(7610):94–103.

Gitaitis, R. and Walcott, R. (2007). The epidemiology and management of seedborne bacterial diseases. *Annual review of phytopathology*, 45:371–97.

Glazebrook, J., Rogers, E. E., and Ausubel, F. M. (1996). Isolation of Arabidopsis mutants with enhanced disease susceptibility by direct screening. *Genetics*, 143(2):973–982.

Glick, B. R. (2012). Plant Growth-Promoting Bacteria : Mechanisms and Applications. *Scientifica*, 2012:15.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep Sparse Rectifier Neural Networks. In *Aistats*, volume 15, page 275.

González, E., Solano, R., Rubio, V., Leyva, A., and Paz-ares, J. (2005). PHOSPHATE TRANSPORTER TRAFFIC FACILITATOR1 Is a Plant-Specific SEC12-Related Protein That Enables the Endoplasmic Reticulum Exit of a High-Affinity Phosphate Transporter in Arabidopsis. *The Plant cell*, 17(December):3500–3512.

Goodman, A. L., McNulty, N. P., Zhao, Y., Leip, D., Mitra, R. D., Lozupone, C. a., Knight, R., and Gordon, J. I. (2009). Identifying genetic determinants needed to establish a human gut symbiont in its habitat. *Cell Host & Microbe*, 6(3):279–289.

Goodrich, J., Waters, J., Poole, A., Sutter, J., Koren, O., Blekhman, R., Beaumont, M., VanTreuren, W., Knight, R., Bell, J., Spector, T., Clark, A., and Ley, R. (2014). Human Genetics Shape the Gut Microbiome. *Cell*, 159(4):789–799.

Gottel, N. R., Castro, H. F., Kerley, M., Yang, Z., Pelletier, D. A., Podar, M., Karpinets, T., Uberbacher, E. E., Tuskan, G. A., Vilgalys, R., Doktycz, M. J., and Schadt, C. W. (2011). Distinct Microbial Communities within the Endosphere and Rhizosphere of Populus deltoides Roots across Contrasting Soil Types. *Applied and Environmental Microbiology*, 77(17):5934–5944.

Greenblum, S., Turnbaugh, P. J., and Borenstein, E. (2012). Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proceedings of the National Academy of Sciences*, 109(2):594–9.

Guo, X.-Z., Zhang, G.-R., Wei, K.-J., Guo, S.-S., Gardner, J. P., and Xie, C.-X. (2013). Development of twenty-one polymorphic tetranucleotide microsatellite loci for Schizothorax o'connori and their conservation application. *Biochemical Systematics and Ecology*, 51:259–263.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methe, B., DeSantis, T. Z., Petrosino, J. F., Knight, R., and Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, 21(3):494–504.

Hacquard, S., Garrido-Oter, R., González, A., Spaepen, S., Ackermann, G., Lebeis, S., McHardy, A. C., Dangl, J. L., Knight, R., Ley, R., and Schulze-Lefert, P. (2015). Microbiota and Host Nutrition across Plant and Animal Kingdoms. *Cell Host & Microbe*, 17(5):603–616.

Hacquard, S., Kracher, B., Hiruma, K., Münch, P. C., Garrido-Oter, R., Thon, M. R., Weimann, A., Damm, U., Dallery, J.-F., Hainaut, M., Henrissat, B., Lespinet, O., Sacristán, S., Ver Loren van Themaat, E., Kemen, E., McHardy, A. C., Schulze-Lefert, P., and O'Connell, R. J. (2016). Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nature Communications*, 7(May):11362.

Hallmann, J., Quadt-Hallmann, A., Mahaffee, W. F., and Kloepper, J. W. (1997). Bacterial endophytes in agricultural crops. *Canadian Journal of Microbiology*, 43(10):895–914.

Haney, C. H., Samuel, B. S., Bush, J., and Ausubel, F. M. (2015). Associations with rhizosphere bacteria can confer an adaptive advantage to plants. *Nature Plants*, 1(6):15051.

Hardoim, P. R., van Overbeek, L. S., and van Elsas, J. D. (2008). Properties of bacterial endophytes and their proposed role in plant growth. *Trends in Microbiology*, 16(10):463–471.

Harms, K. E., Wright, S. J., Calderón, O., Hernández, a., and Herre, E. a. (2000). Pervasive density-dependent recruitment enhances seedling diversity in a tropical forest. *Nature*, 404(6777):493–495.

Harrison, M. J. (2012). Cellular programs for arbuscular mycorrhizal symbiosis. *Current Opinion in Plant Biology*, 15(6):691–698.

Hartmann, M., Lee, S., Hallam, S. J., and Mohn, W. W. (2009). Bacterial, archaeal and eukaryal community structures throughout soil horizons of harvested and naturally disturbed forest stands. *Environmental Microbiology*, 11(12):3045–3062.

Hernández, M., Dumont, M. G., Yuan, Q., and Conrad, R. (2015). Different bacterial populations associated with the roots and rhizosphere of rice incorporate plant-derived carbon. *Applied and Environmental Microbiology*, 81(6):AEM.03209–14.

Herrera Paredes, S. (2016). AMOR: Abundance Matrix Operations in R.

Herrera Paredes, S. and Lebeis, S. L. (2016). Giving back to the community: microbial mechanisms of plant-soil interactions. *Functional Ecology*, 30(7):1043–1052.

Hintner, J.-p., Lechner, C., Riegert, U., Kuhm, E., Storm, T., Reemtsma, T., Stolz, A., and Kuhm, A. E. (2001). Direct Ring Fission of Salicylate by a Salicylate 1 , 2-Dioxygenase Activity from Pseudaminobacter salicylatoxidans Direct Ring Fission of Salicylate by a Salicylate 1 , 2-Dioxygenase Activity from Pseudaminobacter salicylatoxidans. *Society*, 183(23):6936–6942.

Hiruma, K., Gerlach, N., Sacristán, S., Nakano, R. T., Hacquard, S., Kracher, B., Neumann, U., Ramírez, D., Bucher, M., O'Connell, R. J., and Schulze-Lefert, P. (2016). Root Endophyte Colletotrichum tofieldiae Confers Plant Fitness Benefits that Are Phosphate Status Dependent. *Cell*, pages 1–11.

Hooper, L. V. (2001). Commensal Host-Bacterial Relationships in the Gut. *Science*, 292(5519):1115–1118.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366.

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

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., and Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences*, 111(13):4904–4909.

Hu, Y. (2003). The Arabidopsis Auxin-Inducible Gene ARGOS Controls Lateral Organ Size. *THE PLANT CELL ONLINE*, 15(9):1951–1961.

Huang, T.-K., Han, C.-L., Lin, S.-I., Chen, Y.-J., Tsai, Y.-C., Chen, Y.-R., Chen, J.-W., Lin, W.-Y., Chen, P.-M., Liu, T.-Y., Chen, Y.-S., Sun, C.-M., and Chiou, T.-J. (2013). Identification of downstream components of ubiquitin-conjugating enzyme PHOSPHATE2 by quantitative membrane proteomics in Arabidopsis roots. *The Plant cell*, 25(10):4044–4060.

Hubert, D. a., He, Y., McNulty, B. C., Tornero, P., and Dangl, J. L. (2009). Specific Arabidopsis HSP90.2 alleles recapitulate RAR1 cochaperone function in plant NB-LRR disease resistance protein regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(24):9556–9563.

Huot, B., Yao, J., Montgomery, B. L., and He, S. Y. (2014). Growth-defense tradeoffs in plants: A balancing act to optimize fitness. *Molecular Plant*, 7(8):1267–1287.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J. H., Chinwalla, A. T., Creasy, H. H., Earl, A. M., FitzGerald, M. G., Fulton, R. S., Giglio, M. G., Hallsworth-Pepin, K., Lobos, E. A., Madupu, R., Magrini, V., Martin, J. C., Mitreva, M., Muzny, D. M., Sodergren, E. J., Versalovic, J., Wollam, A. M., Worley, K. C., Wortman, J. R., Young, S. K., Zeng, Q., Aagaard, K. M., Abolude, O. O., Allen-Vercoe, E., Alm, E. J., Alvarado, L., Andersen, G. L., Anderson, S., Appelbaum, E., Arachchi, H. M., Armitage, G., Arze, C. A., Ayyaz, T., Baker, C. C., Begg, L., Belachew, T., Bhonagiri, V., Bihan, M., Blaser, M. J., Bloom, T., Bonazzi, V., Paul Brooks, J., Buck, G. A., Buhay, C. J., Busam, D. A., Campbell, J. L., Canon, S. R., Cantarel, B. L., Chain, P. S. G., Chen, I.-M. A., Chen, L., Chhibba, S., Chu, K., Ciulla, D. M., Clemente, J. C., Clifton, S. W., Conlan, S., Crabtree, J., Cutting, M. A., Davidovics, N. J., Davis, C. C., DeSantis, T. Z., Deal, C., Delehaunty, K. D., Dewhirst, F. E., Deych, E., Ding, Y., Dooling, D. J., Dugan, S. P., Michael Dunne, W., Scott Durkin, A., Edgar, R. C., Erlich, R. L., Farmer, C. N., Farrell, R. M., Faust, K., Feldgarden, M., Felix, V. M., Fisher, S., Fodor, A. A., Forney, L. J., Foster, L., Di Francesco, V., Friedman, J., Friedrich, D. C., Fronick, C. C., Fulton, L. L., Gao, H., Garcia, N., Giannoukos, G., Giblin, C., Giovanni, M. Y., Goldberg, J. M., Goll, J., Gonzalez, A., Griggs, A., Gujja, S., Kinder Haake, S., Haas, B. J., Hamilton, H. A., Harris, E. L., Hepburn, T. A., Herter, B., Hoffmann, D. E., Holder, M. E., Howarth, C., Huang,

K. H., Huse, S. M., Izard, J., Jansson, J. K., Jiang, H., Jordan, C., Joshi, V., Katancik, J. A., Keitel, W. A., Kelley, S. T., Kells, C., King, N. B., Knights, D., Kong, H. H., Koren, O., Koren, S., Kota, K. C., Kovar, C. L., Kyrpides, N. C., La Rosa, P. S., Lee, S. L., Lemon, K. P., Lennon, N., Lewis, C. M., Lewis, L., Ley, R. E., Li, K., Liolios, K., Liu, B., Liu, Y., Lo, C.-C., Lozupone, C. A., Dwayne Lunsford, R., Madden, T., Mahurkar, A. A., Mannon, P. J., Mardis, E. R., Markowitz, V. M., Mavromatis, K., McCorrison, J. M., McDonald, D., McEwen, J., McGuire, A. L., McInnes, P., Mehta, T., Mihindukulasuriya, K. A., Miller, J. R., Minx, P. J., Newsham, I., Nusbaum, C., O'Laughlin, M., Orvis, J., Pagani, I., Palaniappan, K., Patel, S. M., Pearson, M., Peterson, J., Podar, M., Pohl, C., Pollard, K. S., Pop, M., Priest, M. E., Proctor, L. M., Qin, X., Raes, J., Ravel, J., Reid, J. G., Rho, M., Rhodes, R., Riehle, K. P., Rivera, M. C., Rodriguez-Mueller, B., Rogers, Y.-H., Ross, M. C., Russ, C., Sanka, R. K., Sankar, P., Fah Sathirapongsasuti, J., Schloss, J. A., Schloss, P. D., Schmidt, T. M., Scholz, M., Schriml, L., Schubert, A. M., Segata, N., Segre, J. A., Shannon, W. D., Sharp, R. R., Sharpton, T. J., Shenoy, N., Sheth, N. U., Simone, G. A., Singh, I., Smillie, C. S., Sobel, J. D., Sommer, D. D., Spicer, P., Sutton, G. G., Sykes, S. M., Tabbaa, D. G., Thiagarajan, M., Tomlinson, C. M., Torralba, M., Treangen, T. J., Truty, R. M., Vishnivetskaya, T. A., Walker, J., Wang, L., Wang, Z., Ward, D. V., Warren, W., Watson, M. A., Wellington, C., Wetterstrand, K. A., White, J. R., Wilczek-Boney, K., Wu, Y., Wylie, K. M., Wylie, T., Yandava, C., Ye, L., Ye, Y., Yooseph, S., Youmans, B. P., Zhang, L., Zhou, Y., Zhu, Y., Zoloth, L., Zucker, J. D., Birren, B. W., Gibbs, R. A., Highlander, S. K., Methé, B. A., Nelson, K. E., Petrosino, J. F., Weinstock, G. M., Wilson, R. K., and White, O. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

Inceolu, Ö., Al-Soud, W. A., Salles, J. F., Semenov, A. V., and van Elsas, J. D. (2011). Comparative Analysis of Bacterial Communities in a Potato Field as Determined by Pyrosequencing. *PLoS ONE*, 6(8):e23321.

Inceolu, Ö., Salles, J. F., Van Overbeek, L., and Van Elsas, J. D. (2010). Effects of plant genotype and growth stage on the betaproteobacterial communities associated with different potato cultivars in two fields. *Applied and Environmental Microbiology*, 76(11):3675–3684.

Jackman, S. (2015). pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory.

Jamet, A. and Nassif, X. (2015). New players in the toxin field: polymorphic toxin systems in bacteria. *mBio*, 6(3):e00285–15.

Jeong, M. L. (2004). Metabolic Profiling of the Sink-to-Source Transition in Developing Leaves of Quaking Aspen. *PLANT PHYSIOLOGY*, 136(2):3364–3375.

Jetiyanon, K., Wittaya-Areekul, S., and Plianbangchang, P. (2008). Film coating of seeds with Bacillus cereus RS87 spores for early plant growth enhancement. *Canadian journal of microbiology*, 54(10):861–7.

Johnston-Monje, D. and Raizada, M. N. (2011). Conservation and diversity of seed associated endophytes in Zea across boundaries of evolution, ethnography and ecology. *PloS one*, 6(6):e20396.

Jones, J. D. G. and Dangl, J. L. (2006). The plant immune system. *Nature*, 444(7117):323–9.

Joshi, N. and Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Jost, R., Pharmawati, M., Lapis-Gaza, H. R., Rossig, C., Berkowitz, O., Lambers, H., and Finnegan, P. M. (2015). Differentiating phosphate-dependent and phosphate-independent systemic phosphate-starvation response networks in Arabidopsis thaliana through the application of phosphite. *Journal of Experimental Botany*, 66(9):2501–2514.

Karthikeyan, A. S., Varadarajan, D. K., Jain, A., Held, M. A., Carpita, N. C., and Raghothama, K. G. (2007). Phosphate starvation responses are mediated by sugar signaling in Arabidopsis. *Planta*, 225(4):907–918.

Kastman, E. K., Kamelamela, N., Norville, J. W., Cosetta, C. M., Dutton, R. J., and Wolfe, B. E. (2016). Biotic Interactions Shape the Ecological Distributions of Staphylococcus Species. *mBio*, 7(5):e01157–16.

Katagiri, F. and Tsuda, K. (2010). Understanding the Plant Immune System. *Molecular Plant-Microbe Interactions*, 23(12):1531–1536.

Khan, G. A., Vogiatzaki, E., Glauser, G., and Poirier, Y. (2016). Phosphate Deficiency Induces the Jasmonate Pathway and Enhances Resistance to Insect Herbivory. *Plant Physiology*, 171(1):632–644.

Kim, Y., Tsuda, K., Igarashi, D., Hillmer, R. A., Sakakibara, H., Myers, C. L., and Katagiri, F. (2014). Mechanisms Underlying Robustness and Tunability in a Plant Immune Signaling Network. *Cell Host & Microbe*, 15(1):84–94.

Kirik, V., Bouyer, D., Schöbinger, U., Bechtold, N., Herzog, M., Bonneville, J.-M., and Hülskamp, M. (2001). CPR5 is involved in cell proliferation and cell death control and encodes a novel transmembrane protein. *Current Biology*, 11(23):1891–1895.

Kliebenstein, D. J., Figuth, A., and Mitchell-Olds, T. (2002). Genetic architecture of plastic methyl jasmonate responses in Arabidopsis thaliana. *Genetics*, 161(4):1685–1696.

Knief, C., Ramette, A., Frances, L., Alonso-Blanco, C., and Vorholt, J. a. (2010). Site and plant species are important determinants of the Methylobacterium community composition in the plant phyllosphere. *The ISME journal*, 4(6):719–28.

Koeppel, A. F. and Wu, M. (2013). Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Research*, 41(10):5175–5188.

Kover, P. X., Dolan, T. E., and Clay, K. (1997). Potential versus actual contribution of vertical transmission to pathogen fitness. *Proceedings of the Royal Society B: Biological Sciences*, 264(1383):903–909.

Kover, P. X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I. M., Purugganan, M. D., Durrant, C., and Mott, R. (2009). A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. *PLoS genetics*, 5(7):e1000551.

Kunin, V. and Hugenholtz, P. (2010). PyroTagger: A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. *The Open Journal*, page Article 1.

Kuzyakov, Y. and Domanski, G. (2000). Carbon input by plants into the soil. Review. *Zeitschrift für Pflanzenernährung und Bodenkunde*, 163(4):421–431.

Laforest-Lapointe, I., Messier, C., Kembel, S. W., Lindow, S., Brandl, M., Herre, E., Mejía, L., Kyllo, D., Rojas, E., Maynard, Z., Butler, A., Fürnkranz, M., Wanek, W., Richter, A., Abell, G., Rasche, F., Sessitsch, A., Morris, C., Kinkel, L., Lindow, S., Hecht-Poinar, E., Elliott, V., Andrews, J., Harris, R., Lambais, M., Crowley, D., Cury, J., Büll, R., Rodrigues, R., Jumpponen, A., Jones, K., Rodriguez, R., White, J., Arnold, A., Redman, R., Redford, A., Bowers, R., Knight, R., Linhart, Y., Fierer, N., Arnold, A., Mejía, L., Kyllo, D., Rojas, E., Maynard, Z., Robbins, N., Herre, E., Vorholt, J., Müller, T., Ruppel, S., Osono, T., Suda, W., Nagasaki, A., Shishido, M., Gilbert, G., Newton, A., Gravouil, C., Fountaine, J., Kim, M., Singh, D., Lai-Hoe, A., Go, R., Rahim, R., Ainuddin, A., Kembel, S., O'Connor, T., Arnold, H., Hubbell, S., Wright, S., Green, J., Kembel, S., Mueller, R., Mercier, J., Lindow, S., Atamna-Ismaeel, N., Finkel, O., Glaser, F., Mering, C., Vorholt, J., Koblížek, M., Abanda-Nkpwatt, D., Müsch, M., Tschiersch, J., Boettner, M., Schwab, W., Innerebner, G., Knief, C., Vorholt, J., Shade, A., Handelsman, J., Redford, A., Fierer, N., Burke, C., Thomas, T., Lewis, M., Steinberg, P., Kjelleberg, S., Knief, C., Ramette, A., Frances, L., Alonso-Blanco, C., Vorholt, J., Osono, T., Cordier, T., Robin, C., Capdevielle, X., Desprez-Loustau, M., Vacher, C., Finkel, O., Burch, A., Lindow, S., Post, A., Belkin, S., Rastogi, G., Tech, J., Coaker, G., Leveau, J., Delmotte, N., Knief, C., Chaffron, S., Innerebner, G., Roschitzki, B., Schlapbach, R., Schmidt, T., Rodrigues, J., Mering, C., Bulgarelli, D., Schlaeppi, K., Spaepen, S., Themaat, E., Schulze-Lefert, P., Janssen, P., Kim, M., Heo, E., Kang, H., Adams, J., Normand, P., Lindow, S., Arny, D., Upper, C., Biebl, H., Pfennig, N., Lambais, M., Lucheta, A., Crowley, D., Wright, I., Reich, P., Westoby, M., Ackerly, D., Baruch, Z., Bongers, F., Wright, S., Kitajima, K., Kraft, N., Reich, P., Wright, I., Bunker, D., Gloor, G., Hummelen, R., Macklaim, J., Dickson, R., Fernandes, A., MacPhee, R., Chelius, M., Triplett, E., Zhang, J., Kobert, K., Flouri, T., Stamatakis, A., Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F., Costello, E., Edgar, R., DeSantis, T., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen, G., Abrams, M., Kubiske, M., Farrar, J., Shipley, B., Vu, T., Niinemets, Ü., Valladares, F., Chave, J., Coomes, D., Jansen, S., Lewis, S., Swenson, N., Zanne, A., Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W., Acinas, S., Sarma-Rupavtarm, R., Klepac-Ceraj, V., Polz, M., Paradis, E., Claude, J., Strimmer, K., Wickham, H., Kembel, S., Cowan, P., Helmus, M., Cornwell, W., Morlon, H., Ackerly, D., Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M., Oksanen, M., Lozupone, C., Hamady, M., Knight, R., Anderson, M., Hochberg, Y., Bland, J., and Altman, D. (2016). Host species identity, site and time drive temperate tree phyllosphere bacterial community structure. *Microbiome*, 4(1):27.

Lambers, H., Martinoia, E., and Renton, M. (2015). Plant adaptations to severely phosphorus-impoverished soils. *Current Opinion in Plant Biology*, 25:23–31.

Lambers, H., Mougel, C., Jaillard, B., and Hinsinger, P. (2009). Plant-microbe-soil interactions in the rhizosphere: An evolutionary perspective. *Plant and Soil*, 321(1-2):83–115.

Lane, D. J. (1991). 16S/23S rRNA sequencing. In Stackebrandt, E. and Goodfellow, M., editors, *Nucleic Acid Techniqies in Bacterial Systematics*, pages 115–175. Wiley, Chichester, United Kingdom.

Lebeis, S. L., Herrera Paredes, S., Lundberg, D. S., Breakfield, N., Gehring, J., McDonald, M., Malfatti, S., Glavina del Rio, T., Jones, C. D., Tringe, S. G., and Dangl, J. L. (2015). Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science*, 349(6250):860–864.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lee, S. M., Donaldson, G. P., Mikulski, Z., Boyajian, S., Ley, K., and Mazmanian, S. K. (2013). Bacterial colonization factors control specificity and stability of the gut microbiota. *Nature*, 501(7467):426–9.

Leggett, M., Cross, J., Hnatowich, G., and Holloway, G. (2010). Challenges in commercializing a phosphate-solubilizing microorganism: Penicillium bilaiae, a case history. In Dion, P., editor, *First International Meeting on Microbial Phosphate Solubilization*, volume 102 of *Soil Biology*, pages 215–222. Springer Netherlands, Dordrecht.

Levey, S. and Wingler, A. (2005). Natural variation in the regulation of leaf senescence and relation to other traits in Arabidopsis. *Plant, Cell and Environment*, 28(2):223–231.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

Lin, W.-y., Huang, T.-k., and Chiou, T.-j. (2013). NITROGEN LIMITATION ADAPTATION , a Target of MicroRNA827 , Mediates Degradation of Arabidopsis © American Society of Plant Biologists NITROGEN LIMITATION ADAPTATION , a Target of MicroRNA827 , Mediates Degradation of Plasma Membrane Localized Phosphate. *The Plant cell*, 25(October):4061–4074.

Links, M. G., Demeke, T., Gräfenhan, T., Hill, J. E., Hemmingsen, S. M., and Dumonceaux, T. J. (2014). Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on Triticum and Brassica seeds. *The New phytologist*, 202:542–553.

Liu, Z., DeSantis, T. Z., Andersen, G. L., and Knight, R. (2008). Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Research*, 36(18):e120–e120.

Logemann, J., Schell, J., and Willmitzer, L. (1987). Improved method for the isolation of RNA from plant tissues. *Analytical Biochemistry*, 163(1):16–20.

Lopez-Velasco, G., Carder, P. a., Welbaum, G. E., and Ponder, M. a. (2013). Diversity of the spinach (Spinacia oleracea) spermosphere and phyllosphere bacterial communities. *FEMS Microbiology Letters*, 346(2):146–154.

Loy, A., Maixner, F., Wagner, M., and Horn, M. (2007). probeBase–an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Research*, 35(Database):D800–D804.

Lozupone, C. and Knight, R. (2005). UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, 71(12):8228–8235.

Lu, Y., Abraham, W. R., and Conrad, R. (2007). Spatial variation of active microbiota in the rice rhizosphere revealed by in situ stable isotope probing of phospholipid fatty acids. *Environmental Microbiology*, 9(2):474–481.

Lu, Y.-T., Li, M.-Y., Cheng, K.-T., Tan, C. M., Su, L.-W., Lin, W.-Y., Shih, H.-T., Chiou, T.-J., and Yang, J.-Y. (2014). Transgenic Plants That Express the Phytoplasma Effector SAP11 Show Altered Phosphate Starvation and Defense Responses. *Plant Physiology*, 164(3):1456–1469.

Lun, A. T., Chen, Y., and Smyth, G. K. (2016). It's DE-licious: A Recipe for Differential Expression Analyses of RNA-seq Experiments Using Quasi-Likelihood Methods in edgeR. In Mathé, E. and Davis, S., editors, *Statistical Genomics: Methods and Protocols*, volume 1418 of *Methods in Molecular Biology*, pages 391–416. Springer New York, New York, NY.

Lundberg, D. S., Lebeis, S. L., Herrera Paredes, S., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., Glavina del Rio, T., Edgar, R. C., Eickhorst, T., Ley, R. E., Hugenholtz, P., Tringe, S. G., and Dangl, J. L. (2012). Defining the core Arabidopsis thaliana root microbiome. *Nature*, 488(7409):86–90.

Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D., and Dangl, J. L. (2013). Practical innovations for high-throughput amplicon sequencing. *Nature Methods*, 10(10):999–1002.

Magoč, T. and Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963.

Maignien, L., Deforce, E. A., Chafee, M. E., Eren, A. M., and Simmons, S. L. (2014). Ecological Succession and Stochastic Variation in the Assembly of Arabidopsis thaliana Phyllosphere Communities. *mBio*, 5(1):e00682–13.

Makino, A., Mae, T., and Ohira, K. (1983). Photosynthesis and Ribulose 1,5-Bisphosphate Carboxylase in Rice Leaves. *Plant physiology*, 73(4):1002–1007.

Marasco, R., Rolli, E., Ettoumi, B., Vigani, G., Mapelli, F., Borin, S., Abou-Hadid, A. F., El-Behairy, U. a., Sorlini, C., Cherif, A., Zocchi, G., and Daffonchio, D. (2012). A drought resistance-promoting microbiome is selected by root system under desert farming. *PloS one*, 7(10):e48479.

Marschner, H. (1995). *Mineral nutrition of higher plants*, volume Academic P.

Marschner, H., Römheld, V., Horst, W. J., and Martin, P. (1986). Root-induced changes in the rhizosphere: Importance for the mineral nutrition of plants. *Zeitschrift für Pflanzenernährung und Bodenkunde*, 149(4):441–456.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):pp. 10–12.

Maruthachalam, K., Klosterman, S. J., Anchieta, A., Mou, B., and Subbarao, K. V. (2013). Colonization of spinach by Verticillium dahliae and effects of pathogen localization on the efficacy of seed treatments. *Phytopathology*, 103(3):268–80.

Masclaux, C., Valadier, M.-H., Brugière, N., Morot-Gaudry, J.-F., and Hirel, B. (2000). Characterization of the sink/source transition in tobacco (Nicotiana tabacum L.) shoots in relation to nitrogen management and leaf senescence. *Planta*, 211(4):510–518.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.

McNulty, N. P., Wu, M., Erickson, A. R., Pan, C., Erickson, B. K., Martens, E. C., Pudlo, N. a., Muegge, B. D., Henrissat, B., Hettich, R. L., and Gordon, J. I. (2013). Effects of diet on resource utilization by a model human gut microbiota containing Bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome. *PLoS biology*, 11(8):e1001637.

Mee, M. T., Collins, J. J., Church, G. M., and Wang, H. H. (2014). Syntrophic exchange in synthetic microbial communities. *Proceedings of the National Academy of Sciences*.

Mendes, R., Garbeva, P., and Raaijmakers, J. M. (2013). The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiology Reviews*, 37(5):634–663.

Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J. H. M., Piceno, Y. M., DeSantis, T. Z., Andersen, G. L., Bakker, P. A. H. M., and Raaijmakers, J. M. (2011). Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science*, 332(6033):1097–100.

Min, S., Lee, B., and Yoon, S. (2016). Deep Learning in Bioinformatics. *arXiv preprint arXiv:1603.06430*.

Misson, J., Raghothama, K. G., Jain, A., Jouhet, J., Block, M. A., Bligny, R., Ortet, P., Creff, A., Somerville, S., Rolland, N., Doumas, P., Nacry, P., Herrerra-Estrella, L., Nussaume, L., and Thibaud, M.-C. (2005). A genome-wide transcriptional analysis using Arabidopsis thaliana Affymetrix gene chips determined plant responses to phosphate deprivation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11934–11939.

Morcuende, R., Bari, R., Gibon, Y., Zheng, W., Pant, B. D., Bläsing, O., Usadel, B., Czechowski, T., Udvardi, M. K., Stitt, M., and Scheible, W. R. (2007). Genome-wide reprogramming of metabolism and regulatory networks of Arabidopsis in response to phosphorus. *Plant, Cell and Environment*, 30(1):85–112.

Motulsky, H. (2003). Prism 4 Statistics Guide: Statistical Analyses for Laboratory and Clinical Researchers.

Muharam, F., Maas, S., Bronson, K., and Delahunty, T. (2015). Estimating Cotton Nitrogen Nutrition Status Using Leaf Greenness and Ground Cover Information. *Remote Sensing*, 7(6):7007–7028.

Narang, R. a., Bruene, A., and Altmann, T. (2000). Analysis of phosphate acquisition efficiency in different Arabidopsis accessions. *Plant physiology*, 124(4):1786–99.

Naumann, M., Schüssler, A., and Bonfante, P. (2010). The obligate endobacteria of arbuscular mycorrhizal fungi are ancient heritable components related to the Mollicutes. *The ISME journal*, 4(7):862–871.

Nelson, K. E., Weinstock, G. M., Highlander, S. K., Worley, K. C., Creasy, H. H., Wortman, J. R., Rusch, D. B., Mitreva, M., Sodergren, E., Chinwalla, A. T., Feldgarden, M., Gevers, D., Haas, B. J., Madupu, R., Ward, D. V., Birren, B. W., Gibbs, R. A., Methe, B., Petrosino, J. F., Strausberg, R. L., Sutton, G. G., White, O. R., Wilson, R. K., Durkin, S., Giglio, M. G., Gujja, S., Howarth, C., Kodira, C. D., Kyrpides, N., Mehta, T., Muzny, D. M., Pearson, M., Pepin, K., Pati, A., Qin, X., Yandava, C., Zeng, Q., Zhang, L., Berlin, A. M., Chen, L., Hepburn, T. A., Johnson, J., McCorrison, J., Miller, J., Minx, P., Nusbaum, C., Russ, C., Sykes, S. M., Tomlinson, C. M., Young, S., Warren, W. C., Badger, J., Crabtree, J., Markowitz, V. M., Orvis, J., Cree, A., Ferriera, S., Fulton, L. L., Fulton, R. S., Gillis, M., Hemphill, L. D., Joshi, V., Kovar, C., Torralba, M., Wetterstrand, K. A., Abouellleil, A., Wollam, A. M., Buhay, C. J., Ding, Y., Dugan, S., FitzGerald, M. G., Holder, M., Hostetler, J., Clifton, S. W., Allen-Vercoe, E., Earl, A. M., Farmer, C. N., Liolios, K., Surette, M. G., Xu, Q., Pohl, C., Wilczek-Boney, K., and Zhu, D. (2010). A Catalog of Reference Genomes from the Human Microbiome. *Science*, 328(5981):994–999.

Nguyen, N. H., Smith, D., Peay, K., and Kennedy, P. (2015). Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist*, 205(4):1389–1393.

Niu, B., Paulson, J. N., Zheng, X., and Kolter, R. (2017). Simplified and representative bacterial community of maize roots. *Proceedings of the National Academy of Sciences*, page 201616148.

Ofek-Lalzar, M., Sela, N., Goldman-Voronov, M., Green, S. J., Hadar, Y., and Minz, D. (2014). Niche and host-associated functional signatures of the root surface microbiome. *Nature communications*, 5:4950.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., and Wagner, H. (2016). vegan: Community Ecology Package.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. (2014). *vegan: Community Ecology Package*.

Pagán, I., Montes, N., Milgroom, M. G., and García-Arenal, F. (2014). Vertical Transmission Selects for Reduced Virulence in a Plant Virus and for Increased Resistance in the Host. *PLoS Pathogens*, 10(7):23–25.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884.

Panke-Buisse, K., Poole, A. C., Goodrich, J. K., Ley, R. E., and Kao-Kniffin, J. (2015). Selection on soil microbiomes reveals reproducible impacts on plant function. *The ISME Journal*, 9(4):980–989.

Pant, B.-D., Pant, P., Erban, A., Huhman, D., Kopka, J., and Scheible, W.-R. (2015). Identification of primary and secondary metabolites with phosphorus status-dependent abundance in Arabidopsis, and of the transcription factor PHR1 as a major regulator of metabolic changes during phosphorus limitation. *Plant, cell & environment*, 38(1):172–87.

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

Pérez-Jaramillo, J. E., Mendes, R., and Raaijmakers, J. M. (2016). Impact of plant domestication on rhizosphere microbiome assembly and functions. *Plant Molecular Biology*, 90(6):635–644.

Pieterse, C. M. J., Van der Does, D., Zamioudis, C., Leon-Reyes, A., and Van Wees, S. C. M. (2012). Hormonal Modulation of Plant Immunity. *Annual Review of Cell and Developmental Biology*, 28(1):489–521.

Pozo, M. J. and Azcón-Aguilar, C. (2007). Unraveling mycorrhiza-induced resistance. *Current Opinion in Plant Biology*, 10(4):393–398.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2009). Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7):1641–1650.

Puga, M. I., Mateos, I., Charukesi, R., Wang, Z., Franco-Zorrilla, J. M., de Lorenzo, L., Irigoyen, M. L., Masiero, S., Bustos, R., Rodriguez, J., Leyva, A., Rubio, V., Sommer, H., and Paz-Ares, J. (2014). SPX1 is a phosphate-dependent inhibitor of PHOSPHATE STARVATION RESPONSE 1 in Arabidopsis. *Proceedings of the National Academy of Sciences*, 111(41):14947–14952.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raghothama, K. G. (1999). Phosphate Acquisition. *Annual Review of Plant Physiology and Plant Molecular Biology*, 50(1):665–693.

Rallapalli, G., Kemen, E., MacLean, D., Robert-Seilaniantz, A., Segonzac, C., Etherington, G., Sohn, K., and Jones, J. (2014). EXPRSS: an Illumina based high-throughput expression-profiling method to reveal transcriptional dynamics. *BMC Genomics*, 15(1):341.

Ranf, S., Eschen-Lippold, L., Pecher, P., Lee, J., and Scheel, D. (2011). Interplay between calcium signalling and early signalling elements during defence responses to microbe- or damage-associated molecular patterns. *Plant Journal*, 68(1):100–113.

Redford, A. J., Bowers, R. M., Knight, R., Linhart, Y., and Fierer, N. (2010). The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environmental Microbiology*, 12(11):2885–2893.

Redford, A. J. and Fierer, N. (2009). Bacterial succession on the leaf surface: A novel system for studying successional dynamics. *Microbial Ecology*, 58(1):189–198.

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3(2):217–223.

Richardson, A. E. and Simpson, R. J. (2011). Soil microorganisms mediating phosphorus availability update on microbial phosphorus. *Plant physiology*, 156(3):989–996.

Ristova, D., Carre, C., Pervent, M., Medici, A., Kim, G. J., Scalia, D., Ruffel, S., Birnbaum, K. D., Lacombe, B., Busch, W., Coruzzi, G. M., and Krouk, G. (2016). Combinatorial interaction network of transcriptomic and phenotypic responses to nitrogen and hormones in the Arabidopsis thaliana root. *Science Signaling*, 9(451):rs13–rs13.

Roberts, D. W. (2016). labdsv: Ordination and Multivariate Analysis for Ecology.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Rodríguez, H. and Fraga, R. (1999). Phosphate solubilizing bacteria and their role in plant growth promotion. *Biotechnology advances*, 17(4-5):319–339.

Rodriguez, R. J., Henson, J., Van Volkenburgh, E., Hoy, M., Wright, L., Beckwith, F., Kim, Y.-O., and Redman, R. S. (2008). Stress tolerance in plants via habitat-adapted symbiosis. *The ISME Journal*, 2(4):404–416.

Rolig, A. S., Parthasarathy, R., Burns, A. R., Bohannan, B. J., and Guillemin, K. (2015). Individual Members of the Microbiota Disproportionately Modulate Host Innate Immune Responses. *Cell Host & Microbe*, 18(5):613–620.

Rousk, J. and Bengtson, P. (2014). Microbial regulation of global biogeochemical cycles. *Frontiers in Microbiology*, 5(7441):305–307.

Santhanam, R., Luu, V. T., Weinhold, A., Goldberg, J., Oh, Y., and Baldwin, I. T. (2015). Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping. *Proceedings of the National Academy of Sciences*, 112(36):E5013–E5020.

Schachtman, D. P., Reid, R. J., Ayling, S. M., S, D. B. D. P., and A, S. S. S. M. (1998). Phosphorus Uptake by Plants : From Soil to Cell. *Plant Physiology*, 116:447–453.

Schlaeppi, K., Dombrowski, N., Oter, R. G., Ver Loren van Themaat, E., and Schulze-Lefert, P. (2014). Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. *Proceedings of the National Academy of Sciences*, 111(2):585–592.

Schlötterer, C., Kofler, R., Versace, E., Tobler, R., and Franssen, S. U. (2015). Combining experimental evolution with next-generation sequencing : a powerful tool to study adaptation from standing genetic variation. *Heredity*, 114:431–440.

Schmitz, A. M. and Harrison, M. J. (2014). Signaling events during initiation of arbuscular mycorrhizal symbiosis. *Journal of Integrative Plant Biology*, 56(3):250–261.

Schnitzer, S. A., Klironomos, J. N., HilleRisLambers, J., Kinkel, L. L., Reich, P. B., Xiao, K., Rillig, M. C., Sikes, B. A., Callaway, R. M., Mangan, S. A., van Nes, E. H., and Scheffer, M. (2011). Soil microbes drive the classic plant diversityproductivity pattern. *Ecology*, 92(2):296–303.

Schulz, B. J. E., Boyle, C. J. C., and Sieber, T. N. (2006). *Microbial Root Endophytes*, volume 9 of *Soil Biology*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Schwachtje, J., Karojet, S., Kunz, S., Brouwer, S., and van Dongen, J. T. (2012). Plant-growth promoting effect of newly isolated rhizobacteria varies between two Arabidopsis ecotypes. *Plant Signaling & Behavior*, 7(6):623–627.

Schweizer, F., Fernández-Calvo, P., Zander, M., Diez-Diaz, M., Fonseca, S., Glauser, G., Lewsey, M. G., Ecker, J. R., Solano, R., and Reymond, P. (2013). Arabidopsis Basic Helix-Loop-Helix Transcription Factors MYC2, MYC3, and MYC4 Regulate Glucosinolate Biosynthesis, Insect Performance, and Feeding Behavior. *The Plant cell*, 25(8):3117–32.

Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., Mitter, B., Hauberg-Lotte, L., Friedrich, F., Rahalkar, M., Hurek, T., Sarkar, A., Bodrossy, L., van Overbeek, L., Brar, D., van Elsas, J. D., and Reinhold-Hurek, B. (2012). Functional Characteristics of an Endophyte Community Colonizing Rice Roots as Revealed by Metagenomic Analysis. *Molecular Plant-Microbe Interactions*, 25(1):28–36.

Shade, A., Mcmanus, P. S., and Handelsman, J. (2013). Unexpected Diversity during Community Succession in the Apple Flower Microbiome. *mBio*, 4(2):e00602–12.

Sharma, S. B., Sayyed, R. Z., Trivedi, M. H., and Gobi, T. a. (2013). Phosphate solubilizing microbes: sustainable approach for managing phosphorus deficiency in agricultural soils. *SpringerPlus*, 2(1):587.

Shin, H., Shin, H. S., Dewbre, G. R., and Harrison, M. J. (2004). Phosphate transport in Arabidopsis: Pht1;1 and Pht1;4 play a major role in phosphate acquisition from both low- and high-phosphate environments. *Plant Journal*, 39(4):629–642.

Silva, T. R., Valdman, E., Valdman, B., and Leite, S. G. (2007). Salicylic acid degradation from aqueous solutions using Pseudomonas fluorescens HK44: parameters studies and application tools. *Brazilian Journal of Microbiology*, 38(1):39–44.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smith, M. I., Yatsunenko, T., Manary, M. J., Trehan, I., Mkakosya, R., Cheng, J., Kau, a. L., Rich, S. S., Concannon, P., Mychaleckyj, J. C., Liu, J., Houpt, E., Li, J. V., Holmes, E., Nicholson, J., Knights, D., Ursell, L. K., Knight, R., and Gordon, J. I. (2013). Gut Microbiomes of Malawian Twin Pairs Discordant for Kwashiorkor. *Science*, 339(6119):548–554.

Smith, P. A. (2014). Why Tiny Microbes Mean Big Things for Farming. *National Geographic*.

Smith, S. E. and Smith, F. a. (2012). Fresh perspectives on the roles of arbuscular mycorrhizal fungi in plant nutrition and growth. *Mycologia*, 104(1):1–13.

Spoel, S. H. and Dong, X. (2008). Making Sense of Hormone Crosstalk during Plant Immune Responses. *Cell Host & Microbe*, 3(6):348–351.

Spor, A., Koren, O., and Ley, R. (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*, 9(4):279–290.

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., Sander, C., and Xavier, J. B. (2013). Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *PLoS Computational Biology*, 9(12):e1003388.

Stewart, A. D., Logsdon, J. M., and Kelley, S. E. (2005). An empirical study of the evolution of virulence under both horizontal and vertical transmission. *Evolution; international journal of organic evolution*, 59(4):730–739.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–5.

Sul, W. J., Cole, J. R., Jesus, E. D. C., Wang, Q., Farris, R. J., Fish, J. a., and Tiedje, J. M. (2011). Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. *Proceedings of the National Academy of Sciences*, 108(35):14637–14642.

Swenson, W., Wilson, D. S., and Elias, R. (2000). Artificial ecosystem selection. *Proceedings of the National Academy of Sciences*, 97(16):9110–4.

Testen, A. L., Jiménez-Gasco, M. d. M., Ochoa, J. B., and Backman, P. A. (2013). Molecular detection of Peronospora variabilis in quinoa seeds and phylogeny of the quinoa downy mildew pathogen in South America and the United States. *Phytopathology*, pages 1–30.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop. *COURSERA: Neural networks for machine learning*.

Timm, C. M., Campbell, A. G., Utturkar, S. M., Jun, S.-R., Parales, R. E., Tan, W. A., Robeson, M. S., Lu, T.-Y. S., Jawdy, S., Brown, S. D., Ussery, D. W., Schadt, C. W., Tuskan, G. A., Doktycz, M. J., Weston, D. J., and Pelletier, D. A. (2015). Metabolic functions of Pseudomonas fluorescens strains from Populus deltoides depend on rhizosphere or endosphere isolation compartment. *Frontiers in Microbiology*, 6(OCT):1–13.

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111.

Tsuda, K., Sato, M., Stoddard, T., Glazebrook, J., and Katagiri, F. (2009). Network Properties of Robust Immunity in Plants. *PLoS Genetics*, 5(12):e1000772.

Turnbaugh, P. J., Hamady, M., Yatsunenko, T., Cantarel, B. L., Duncan, A., Ley, R. E., Sogin, M. L., Jones, W. J., Roe, B. a., Affourtit, J. P., Egholm, M., Henrissat, B., Heath, A. C., Knight, R., and Gordon, J. I. (2009a). A core gut microbiome in obese and lean twins. *Nature*, 457(7228):480–4.

Turnbaugh, P. J., Ridaura, V. K., Faith, J. J., Rey, F. E., Knight, R., and Gordon, J. I. (2009b). The Effect of Diet on the Human Gut Microbiome: A Metagenomic Analysis in Humanized Gnotobiotic Mice. *Science Translational Medicine*, 1(6):6ra14–6ra14.

van der Lelie, D., Taghavi, S., Monchy, S., Schwender, J., Miller, L., Ferrieri, R., Rogers, A., Wu, X., Zhu, W., Weyens, N., Vangronsveld, J., and Newman, L. (2009). Poplar and its Bacterial Endophytes: Coexistence and Harmony. *Critical Reviews in Plant Sciences*, 28(5):346–358.

van Elsas, J. D., Trevors, J. T., and Starodub, M. E. (1988). Bacterial conjugation between pseudomonads in the rhizosphere of wheat. *FEMS Microbiology Letters*, 53(5):299–306.

Van Nuland, M. E., Wooliver, R. C., Pfennigwerth, A. A., Read, Q. D., Ware, I. M., Mueller, L., Fordyce, J. A., Schweitzer, J. A., and Bailey, J. K. (2016). Plant-soil feedbacks: connecting ecosystem ecology and evolution. *Functional Ecology*, 30(7):1032–1042.

Vellend, M. (2010). Conceptual synthesis in community ecology. *The Quarterly review of biology*, 85(2):183–206.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Virgin, H. W. and Todd, J. A. (2011). Metagenomics and Personalized Medicine. *Cell*, 147(1):44–56.

Vitousek, P. M., Porder, S., Houlton, B. Z., and Chadwick, O. a. (2010). Terrestrial phosphorus limitation: mechanisms, implications, and nitrogenphosphorus interactions. *Ecological Applications*, 20(1):5–15.

Vorholt, J. a. (2012). Microbial life in the phyllosphere. *Nature Reviews Microbiology*, 10(12):828–840.

Wagg, C., Bender, S. F., Widmer, F., and van der Heijden, M. G. A. (2014). Soil biodiversity and soil community composition determine ecosystem multifunctionality. *Proceedings of the National Academy of Sciences*, 111(14):5266–5270.

Wagner, M. R., Lundberg, D. S., del Rio, T. G., Tringe, S. G., Dangl, J. L., and Mitchell-Olds, T. (2016). Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications*, 7:12151.

Wang, D., Yang, S., Tang, F., and Zhu, H. (2012). Symbiosis specificity in the legume - rhizobial mutualism. *Cellular Microbiology*, 14(3):334–342.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. (2007). Naive Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267.

Wang, S., Mohamed, A.-r., Caruana, R., Bilmes, J., Plilipose, M., Richardson, M., Geras, K., Urban, G., and Aslan, O. (2016). Analysis of Deep Neural Networks with the Extended Data Jacobian Matrix. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 718–726.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., Schwartz, M., and Venables, B. (2016). gplots: Various R Programming Tools for Plotting Data.

Wei, Z., Yang, T., Friman, V.-P., Xu, Y., Shen, Q., and Jousset, A. (2015). Trophic network architecture of root-associated bacterial communities determines pathogen invasion and plant health. *Nature Communications*, 6:8413.

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

Wright, D. A., Swaminathan, J., Blaser, M., and Jackson, T. A. (2005). Carrot seed coating with bacteria for seedling protection from grass grub damage. *New Zealand Plant Protection*, 58:229–233.

Xie, Y. (2016). knitr: A general-purpose package for dynamic report generation.

Yamada, K., Saijo, Y., Nakagami, H., and Takano, Y. (2016). Regulation of sugar transporter activity for antibacterial defense in Arabidopsis. *Science*, 354(6318):1427–1430.

Yang, L., Teixeira, P. J. P. L., Biswas, S., Finkel, O. M., He, Y., Salas-Gonzalez, I., English, M. E., Epple, P., Mieczkowski, P., and Dangl, J. L. (2017). Pseudomonas syringae Type III Effector HopBB1 Promotes Host Transcriptional Repressor Degradation to Regulate Phytohormone Responses and Virulence. *Cell Host & Microbe*, pages 1–13.

Yeoh, Y. K., Paungfoo-Lonhienne, C., Dennis, P. G., Robinson, N., Ragan, M. a., Schmidt, S., and Hugenholtz, P. (2016). The core root microbiome of sugarcanes cultivated under varying nitrogen fertilizer application. *Environmental Microbiology*, 18(5):1338–1351.

Yi, X., Du, Z., and Su, Z. (2013). PlantGSEA: a gene set enrichment analysis toolkit for plant community. *Nucleic acids research*, 41(Web Server issue):98–103.

Yourstone, S. M., Lundberg, D. S., Dangl, J. L., and Jones, C. D. (2014). MT-Toolbox: improved amplicon sequencing using molecule tags. *BMC Bioinformatics*, 15(1):284.

Yu, G., Smith, D. K., Zhu, H., Guan, Y., and Lam, T. T. Y. (2016). ggtree: An r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, pages 28–36.

Yuan, Z., Druzhinina, I. S., Labbé, J., Redman, R., Qin, Y., Rodriguez, R., Zhang, C., Tuskan, G. A., and Lin, F. (2016). Specialized Microbiome of a Halophyte and its Role in Helping Non-Host Plants to Withstand Salinity. *Scientific Reports*, 6(May):32467.

Zamioudis, C., Mastranesti, P., Dhonukshe, P., Blilou, I., and Pieterse, C. M. J. (2013). Unraveling Root Developmental Programs Initiated by Beneficial Pseudomonas spp. Bacteria. *PLANT PHYSIOLOGY*, 162(1):304–318.

Zamioudis, C. and Pieterse, C. M. J. (2012). Modulation of Host Immunity by Beneficial Microbes. *Molecular Plant-Microbe Interactions*, 25(2):139–150.

Zarraonaindia, I., Owens, S. M., Weisenhorn, P., West, K., Hampton-marcell, J., Lax, S., Bokulich, N. a., Mills, D. a., Martin, G., Taghavi, S., Lelie, D. V. D., and Gilbert, A. (2015). The Soil Microbiome Influences Grapevine-Associated Microbiota. *mBio*, 6(2):1–10.

Zeileis, A., Kleiber, C., and Jackman, S. (2008). Regression Models for Count Data in R. *Journal of Statistical Software*, 27(8).

Zelezniak, A., Andrejev, S., Ponomarova, O., Mende, D. R., Bork, P., and Patil, K. R. (2015). Metabolic dependencies drive species co-occurrence in diverse microbial communities. *Proceedings of the National Academy of Sciences*, 112(20):201421834.

Zeng, Q., Sukumaran, J., Wu, S., and Rodrigo, A. (2015). Neutral Models of Microbiome Evolution. *PLOS Computational Biology*, 11(7):e1004365.

Zgadzaj, R., James, E. K., Kelly, S., Kawaharada, Y., de Jonge, N., Jensen, D. B., Madsen, L. H., and Radutoiu, S. (2015). A Legume Genetic Framework Controls Infection of Nodules by Symbiotic and Endophytic Bacteria. *PLOS Genetics*, 11(6):e1005280.

Zhang, D., de Souza, R. F., Anantharaman, V., Iyer, L. M., and Aravind, L. (2012). Polymorphic toxin systems: Comprehensive characterization of trafficking modes, processing, mechanisms of action, immunity and ecology using comparative genomics. *Biology direct*, 7(1):18.

Zhang, Y., Zhang, Y., Goritschnig, S., Goritschnig, S., Dong, X., Dong, X., Li, X., and Li, X. (2003). A Gain-of-Function Mutation in a Plant Disease Resistance Gene Leads to Constitutive Activation of Downstream Signal Transduction Pathways in. *Society*, 15(11):2636–46.

Zhao, H., Sun, R., Albrecht, U., Padmanabhan, C., Wang, A., Coffey, M. D., Girke, T., Wang, Z., Close, T. J., Roose, M., Yokomi, R. K., Folimonova, S., Vidalakis, G., Rouse, R., Bowman, K. D., and Jin, H. (2013). Small RNA profiling reveals phosphorus deficiency as a contributing factor in symptom expression for citrus huanglongbing disease. *Molecular Plant*, 6(2):301–310.

Zhu, Q., Riley, W. J., Tang, J., and Koven, C. D. (2016). Multiple soil nutrient competition between plants, microbes, and mineral surfaces: Model development, parameterization, and example applications in several tropical forests. *Biogeosciences*, 13(1):341–363.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

Zuppinger-Dingley, D., Schmid, B., Petermann, J. S., Yadav, V., De Deyn, G. B., and Flynn, D. F. B. (2014). Selection for niche differentiation in plant communities increases biodiversity effects. *Nature*, 515(7525):108–111.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (2009). Zero-Truncated and Zero-Inflated Models for Count Data. In *Mixed Effects Models and Extensions in Ecology with R*, Statistics for Biology and Health, pages 261–293. Springer New York, New York, NY, 1 edition.