THE UNIVERSITY OF CHICAGO

DIVERSITY AND HERITABILITY IN *ARABIDOPSIS THALIANA* LEAF MICROBIOMES

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

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY
HANNAH WHITEHURST HACKLEY

CHICAGO, ILLINOIS
DECEMBER 2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

My deep gratitude goes to the community of people who helped me complete my PhD. They gave me support in a variety of ways, from statistics tutorials to warm meals. While I am unable to name everyone in these pages, I am fortunate to share this milestone with those mentioned and unmentioned here.

First, I would like to thank Joy Bergelson for her invaluable guidance during my graduate studies. Joy possesses a remarkable ability to simplify complex concepts, making them engaging and comprehensible. Her systematic thinking, adept statistical skill set, and forward-thinking approach in ecology have profoundly influenced my academic growth. I hope to emulate her creativity and fastidiousness in my future work.

The members of the Bergelson lab taught me a variety of statistical methods, microbiological techniques, and ecological theory. I am profoundly thankful for the friendship and collaboration of Caroline Oldstone-Jackson, Rebecca Satterwhite, Hanna Maerkle, Keven Dooley, Megan Kennedy, and Andrew Gloss. The ever mass-spectacular Tim Morton kept my spirits high and lab materials stocked. I already miss our lunches crowded around the office table in Erman.

The research in the following chapters is built largely off the extensive work done by Ben Brachi. I appreciate our conversations brainstorming ideas for the heritability estimates. Riley Leff and Feng Huang were essential in the isolate collection. Their enthusiasm was contagious - a particularly impressive feat given the tediousness of streaking thousands of bacteria.

My support at the University of Chicago extended beyond the lab. My fellow graduate students at the University of Chicago helped me navigate the academic landscape while staying satiated with Medici milkshakes. Thank you especially to Kristina Fialko and Chloe Nash. The work of Jeff Wisniewski, Bonnie Brown, and Mary Johnson kept the Ecology and Evolution department running smoothly and provided research logistical support. Audrey

Aronowsky shared personal and professional advice that helped me complete my work, for which I am deeply grateful.

Cheryl Swift introduced me to ecology in during my undergraduate education, and I'll never forget our early morning trips to the field. Her continued mentorship helped me enter the PhD program. I also had the privilege to work with Pieter Faber in the University of Chicago sequencing core. His patience and expertise provided me the molecular genetics skill set to start my graduate research, without which none of my dissertation would be possible. I also benefited greatly from the practical skill set afforded to me through a rotation, and continued collaboration, with Jocelyn Malamy. I am grateful for the feedback of my committee members, which undoubtedly improved my work: Stefano Allesina, Maureen Coleman, Greg Dwyer, and Mercedes Pascual.

Patricia Sents, Steve Whitehurst, Zach Whitehurst, Ashley Whitehurst, my nephews, aunts, uncles, cousins, and the Hackley family all supported and loved me unconditionally as I was focused on graduate school these past years. I am indebted to them. My dear friends Erin Brewer, Jessica Little, Megan Hamm, Tanja Florin, Aaron Brewer, Timothy Little, Mark Hamm, Travis Hoyne, and Rob Miles were my pillars in Chicago. Kristina Hamm and Megan Hovick were my inestimable companions in New York. Brie Ross and Melissa Davidson always encouraged me to embrace my passion for science, even as teenagers. Thank you all.

Lastly, Peter Hackley was my cornerstone and best friend through graduate school. Words fail to convey the appreciation and joy I find in his companionship. The endless late-night conversations, walks to the lab on the weekends, and laughs we shared all refueled me when I was running on fumes. Any success I may have is unequivocally tied to him. On to the next chapter, with love.

ABSTRACT

Microbes are associated with all complex organisms, influencing host fitness, local ecology and evolutionary trajectories. Thus, there is burgeoning interest in engineering microbiomes for practical applications, such as sustainable agriculture and precision medicine. However, the emergent phenotype from confounding host, microbe, and environmental interactions within diverse microbiomes proves challenging to characterize, let alone engineer. One method of microbiome characterization is to quantify differential abundances of distinct bacteria taxonomic groups among hosts. The development of high throughput sequencing facilitates this type of characterization using variable regions of marker genes to taxonomically group the microbes. However, the canonical 16S v5-v7 gene region used to assess plant microbiomes is relatively constrained, effectively grouping distinct bacteria into higher levels and potentially masking host-microbe interactions.

Here, I propose decomposing taxonomic groups to lower levels, facilitating finer groupings of bacteria to more accurately describe the respective microbe-microbe interactions and subsequently investigate host-genotype effects on the microbiome abundance phenotype. I first describe the collection and classification of natural bacteria isolates collected from *Arabidopsis thaliana* plants from the field. In the next chapter, I describe the development of a new marker gene database using gyrase subunit-β (*gyrB*). Using the isolates from the previous chapter in combination with published data sets of *A. thaliana* microbiomes, I show that *gyrB* provides both finer taxonomic resolution and stronger correlations between genetic and genomic distances compared to the canonical 16S v5-v7 marker gene. Lastly, I apply *gyrB* sequencing to leaf microbiome community abundances from replicated *A. thaliana* field experiments. I show that using *gyrB*, compared to 16S v5-v7, and including microbe-microbe interactions improves model fits for broad-sense heritability estimates. I use these data to perform Genome Wide Association Studies (GWAS), and identify host gene-candidates potentially shaping the microbiome.

CHAPTER 1

INTRODUCTION

Microbes influence host fitness in plants and animals, including humans. By altering host phenotypes and variation among hosts within host populations, microbes also likely affect evolutionary potential (Henry et al. 2021). Given the broad potential of the microbiome to improve host health and fitness, engineering microbiomes is an exciting avenue for practical applications including sustainable agriculture (Ke, Wang, and Yoshikuni 2021) and precision medicine (Schmidt, Raes, and Bork 2018). However, the ecological machinery and component bacteria required for a beneficial microbiome remain elusive, in part due to their complexity (Albright et al. 2022). Hosts harbor hundreds to thousands of distinct bacterial strains with variable functions. Moreover, microbiome phenotypes are determined by a confluence of dynamic variables including microbe, host and abiotic factors, as well as the interactions among them (Figure 1.1). Here, I broadly review the factors affecting host phenotype, focusing on microbe-microbe interactions in the context of the host. I then show how my research in the following three chapters helps to more precisely measure host-microbiome interactions in order to assess the strength and limitations of host-microbiome interactions in the context of host health.

1.1 Environmental effects on host phenotype

The environment encapsulates a wide swathe of abiotic characteristics, including climate, water and nutrient availability. Variation in the environment necessarily defines the host and microbes that are able to survive and thus determines the pool of microbes that could colonize the host. For this reason, it is not surprising that the environment plays a predominant role in determining microbiome diversity. Up to 20% of human gut microbiome diversity can be attributed to environmental factors, compared to $<2\%$ attributable to host

ancestry or polymorphisms (Rothschild et al. 2018). Still, the variance attributable to the host provides insights to the tunable elements of the microbiome beyond what is available in the environment; while the environment plays a significant role in microbiome and host phenotype, I primarily focus on microbe-microbe and host-microbe interactions.

1.2 Microbe-microbe interactions

In addition to environmental effects, microbe-microbe interactions influence co-occurring microbial abundance through positive, negative, and asymmetric interactions (see Coyte and Rakoff-Nahoum 2019 for a review of microbiome interactions through the lens of human gut microbiomes). For example, sloth skin microbiomes were shown to produce anti-microbial peptides that directly interfere with the propagation of several known pathogens (Rojas-Gätjens et al. 2022). Indirect competition over shared resources also likely leads to the exclusion of microbes, but the nature of interactions is hard to predict: neither phylogenetic distance, pairwise testing nor putative metabolic functions are reliable predictors of indirect competitive exclusion in complex communities (Li et al. 2019; Sundarraman et al. 2020; Garza et al. 2018).

One way to infer microbe-microbe interactions is through the development of networks using co-occurrence data, with nodes representing taxonomically distinct microbes and edges representing the putative interactions (Figure 1.1). To generate microbiome networks, taxonomic abundances are gathered by sequencing the microbial community in host samples that can be considered microbiome "snapshots" collected across time or space. Data are compiled into a matrix of samples with respective relative abundances of microbial taxonomies sequenced. This matrix enables us to infer interactions through models that assess if the changes in abundance of any given microbe is correlated with the changes in abundance of another microbe (e.g. Kurtz et al. 2015; Friedman and Alm 2012). This method is particularly useful because it requires only DNA analysis of metacommunities, rather than the

collection of isolates and subsequent competition experiments, both of which can be costly, time-intensive, and biased by collection methods.

Microbiome samples collected from host tissue often contain a large amount of host DNA intermingled with the microbial DNA, sometimes comprising up to 90% of a metagenomic sample (Marquet et al. 2022). Marker gene amplification enriches for microbiome DNA through the use of primers designed to bind to conserved regions of ubiquitous microbial genes flanking a genetically variable region but not bind to host DNA. The amplicon sequences are crossed-referenced against taxonomic databases for the respective marker gene, such as the SILVA database for 16S ribosomal subunit (Quast et al. 2013). However, the 16S genetic distance of highly conserved marker genes does not have high concordance with genomic distance (Hassler et al. 2022), likely under-representing taxonomic diversity and reducing accuracy for inferred interactions among microbes.

The amplification pipeline also results in read counts per microbial taxa that do not necessarily correlate to input molecules, and so must be transformed to relative abundances. Relative abundance, i.e. compositional data sets, are prone to many statistical pitfalls (Morton et al. 2019; Friedman and Alm 2012). Compositional data are not independent; increasing the value of one taxa requires the decrease of another. This is particularly problematic given the ecology of pathogens, which increase the overall load of microbes in a microbiome without affecting the true counts of other microbes. For example, Karasov et al. (2019) noted that a pathogen invasion increased the microbial count in a plant by adding to an otherwise consistent count of non-pathogens in the community. Thus, the compositional data can create the false impression that non-pathogen microbes decrease in abundance when there is a pathogen invasion but in actuality there is no change in true abundance. This is one example illustrating how compositional data lead to artifactual correlations and can even lead to reverse correlation estimates.

1.3 Host genotype

Physiology of the host organism, ranging from tissue specification to immune responses, can affect colonization and persistence of microbes. Thus, host genotype (and, by extension, phenotype) remain an important component in understanding the impact of microbiomes on the host. For example, infection from *C. difficile* pathogen in humans is a primary cause of Irritable Bowel Diseases (IBD) and is correlated to impaired function of host immune responses, which can be identified through complex trait polymorphisms (Knights et al. 2014).

While there are clear examples of host-genotypes having tight correlations with specific microbes, the influence of host genotype on the structure of the microbiome is less understood. Hosts could, in principle, exert a general influence over the hundreds of microbes within the microbiome, for example through secondary metabolites or general immune responses. Alternatively, hosts could exert control over a few ecologically significant microbes, which subsequently shape the remaining microbiome through direct or indirect interactions (Foster et al. 2017).

One metric used to assess host genotype effects is heritability, which quantifies the proportion of variance in a given microbe's abundance attributable to host genotype. Further analysis has the potential to identify candidate host genes driving the observed variance - understanding the host genetic components shaping microbiome diversity can illuminate important biological pathways driving microbiome diversity and persistence. Heritability is calculated as the variance attribution to the genotype relative to the total variance, which also includes residual variance (considered environmental variance, see Chapter 4 Box 1 for equations). Due to its reliance on residual variance, heritability is susceptible to the well-documented inaccuracies due to ill-fitting models and non-normally distributed parameters. Thus, careful consideration is required for the accurate assessment of heritability estimates.

A large fraction of host-associated microbes have been reported as heritable in plants and

animals, although subsequent analysis to identify host-genes driving this heritability result in few significant candidate genes (Brachi et al. 2022). This is sometimes referred to as the "missing heritability" problem. I hypothesize that missing heritability can be attributed to poor taxonomic classification of microbes through the 16S marker gene, which results in inaccurate genotype effects. Second, I hypothesize that microbe-microbe interactions play a significant role in microbial variance which, when unaccounted for, leads to skewed estimates of host-genotype effects. Thus, improved taxonomic identification and incorporation of microbe-microbe interactions would lead to better identification of host-gene candidates driving heritability. In my research, I first built a gyrase subunit β (*gyrB*) marker gene database for improved taxonomic identification and then tested my model using data from complex microbial communities in *Arabidopsis thaliana*.

1.4 Overview of research

1.4.1 Study system

Arabidopsis thaliana was the first fully sequenced plant genome (The Arabidopsis Genome Initiative 2000). The small flowering annual is a diploid and is highly-selfing. Thus, plants of the same genotype can typically be considered highly homozygous and clonal. In a large international collaboration, researchers generated seed stocks and sequencing data for over 1000 distinct genotypes, providing excellent computational and experimental tools for comparative genotype studies (The 1001 Genomes Consortium 2016; Korte and Ashley 2013).

1.4.2 Chapter 2: Living library of isolates

In Chapter 2, I describe the collection of approximately 4,000 leaf endophytes from natural *A. thaliana* plants. We collected two leaves from 10 plants in each of five populations across Sweden and cultured bacteria on eight distinct media types to improve the probability of

collecting taxonomically diverse isolates. I show that we collected the majority of anticipated taxa, as inferred from culture-independent amplicon sequencing of similar samples from prior years. I use these isolates to experimentally validate the *gyrB* marker gene in Chapter 3. I use the *gyrB* database from Chapter 3 to compare taxonomic diversity of the isolate collection using 16S v5-v7 compared to *gyrB*. As hypothesized, I show there are more distinct strains identified using *gyrB*. I highlight putative ecologically-significant isolates and pathogens identified in the collection. The living isolate library also provides a valuable collection of bacteria isolates for future studies relying on natural isolates.

1.4.3 Chapter 3: gyrase subunit β marker-gene database development

In Chapter 3, I assess the effectiveness of using *gyrB* as a marker gene in complex *A. thaliana* leaf microbial communities compared to the commonly used 16S rRNA v5-v7. The 16S marker gene method was developed to amplify a variable region of the ribosomal subunit found in all bacteria while excluding off-target amplification of host plant DNA, however, the tightly constrained gene only allows differentiation of bacteria down to the family level. An alternative marker gene, gyrase subunit B (*gyrB*), has higher variability and is commonly used in phylogenetic studies to characterize bacteria down to the genus or species taxonomic level. Barret et al. (2015) demonstrated its effectiveness for characterizing soil communities, although other study systems report varying levels of success. I built out a *gyrB* taxonomic database to improve taxonomic identification, showing improved taxonomic identification of isolates using *gyrB* compared to previously published databases. Moreover, I demonstrate that *gyrB* is more tightly correlated to genomic diversity compared to the 16S rRNA (v5-v7) marker gene.

1.4.4 Chapter 4: Heritability in A. thaliana *leaf microbiomes*

In Chapter 4, I assess heritability of the leaf microbiome in A. thaliana. I compare 16S v5-v7 and *gyrB* taxonomic profiles generated for the same samples to compare taxonomic resolution on heritability estimates. Surprisingly, I show that 16S rRNA v5-v7 and *gyrB* analysis of isolates yield similar heritability estimates across families, indicating that heritability of strains does not always correlate to more power when compared to broader taxonomic groupings as seen with 16S. However, I argue that *gyrB* methods still result in more "true" correlations with host genotype given that I am able to identify more statistically significant host gene candidates using *gyrB* compared to the 16S. I also show that including microbe-microbe interactions improves our ability to explain variance observed in microbial abundances as opposed to only the host genotype.

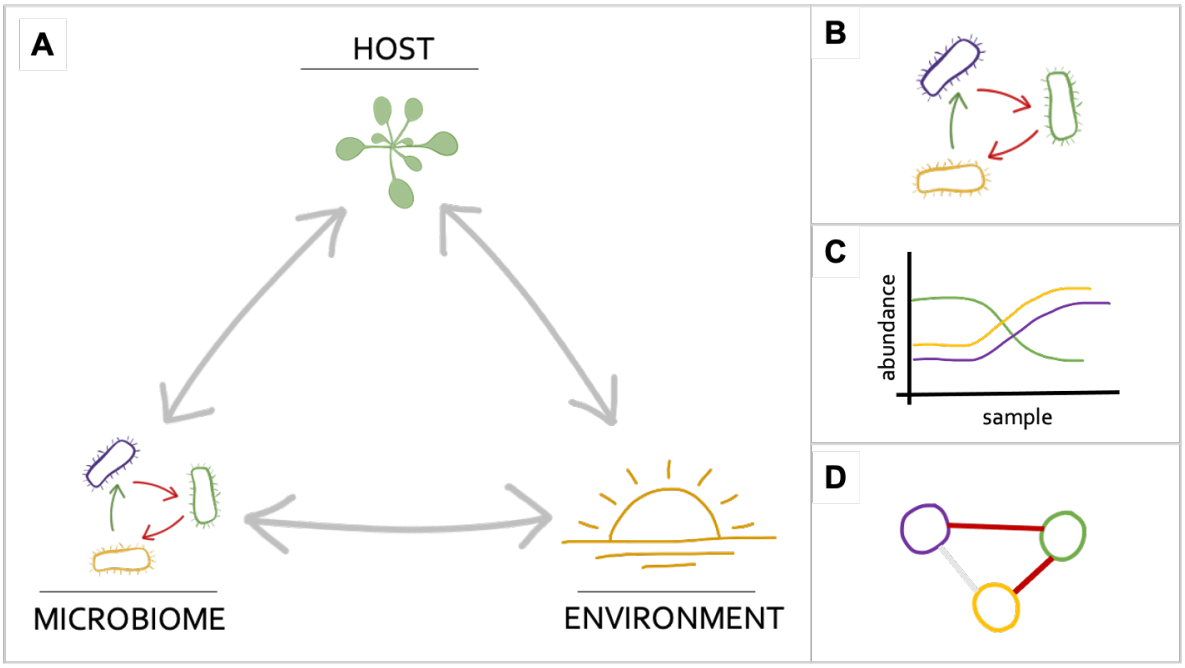

Figure 1.1: Host, microbes, and environment interact dynamically to shape the host phenotype (A). Originally conceptualized by Stevens (1960) to represent disease emergence in plants through pathogens, the model has been expanded to represent host phenotypes more broadly, include feedback loops within the variables, and consider the microbiome beyond pathogens (Bernardo-Cravo et al. 2020). Here we explore the microbe-microbe interactions in more depth. Microbes interact with each other through indirect competition and direct interaction, such as anti-microbial peptide production (B) which can be captured through abundance data collected from host samples (C). Abundance data is used to generate networks to analyze for community-level characterization, representing bacteria as nodes and inferred interactions as edges (D). Some true biological interactions may not be accurately captured in networks because the microbes do not report enough variance in the abundance data to infer correlation (represented by the gray edge in (D)). False positives may also emerge due to sampling error and the nature of compositional data. Conceptualization of network generation recapitulated from (Kurtz et al. 2015).

CHAPTER 2

DIVERSITY OF THE *ARABIDOSIS THALIANA* LEAF MICROBIOME

2.1 Introduction

Plants host a variety of microbes, and the confluence of microbial community members, environment, host genotype and phenotype substantially influence a plant's health and fitness (Brader et al. 2017). Computational tools in combination with culture-independent analysis of microbiomes provide convenient high-throughput inference of host-microbe interactions, such as through the heritability estimates and microbial networks explored in Chapter 4. However, verifying estimated correlations requires empirical testing of naturally occurring microbial isolates. Here I describe the extensive collection, and associated database, of more than 4,000 natural isolates from *Arabidopsis thaliana* plants for future empirical testing. I provide identification of 600 isolates as putative pathogens, identified by taxonomic assignment of marker genes. I further demonstrate the value of the isolate library by using an isolate, *Brevundimonas sp* (B38), to verify direct, positive microbe effects on host plant *A. thaliana* as hypothesized through prior culture-independent methods.

2.2 Results

2.2.1 Successful collection of 4,128 bacteria strains

We plated each of the glycerol leaf stocks across eight distinct media types, which were selected to have different nutrient availability to collect bacteria with various, distinct metabolic requirements (Appendix A, Table 6.1). Colony forming units (CFUs) were allowed to grow for up to three weeks at 28°C, and as new colonies formed, we re-streaked them onto a new plate. Once re-streaked, and CFUs formed, we picked one for growth in liquid Nutrient

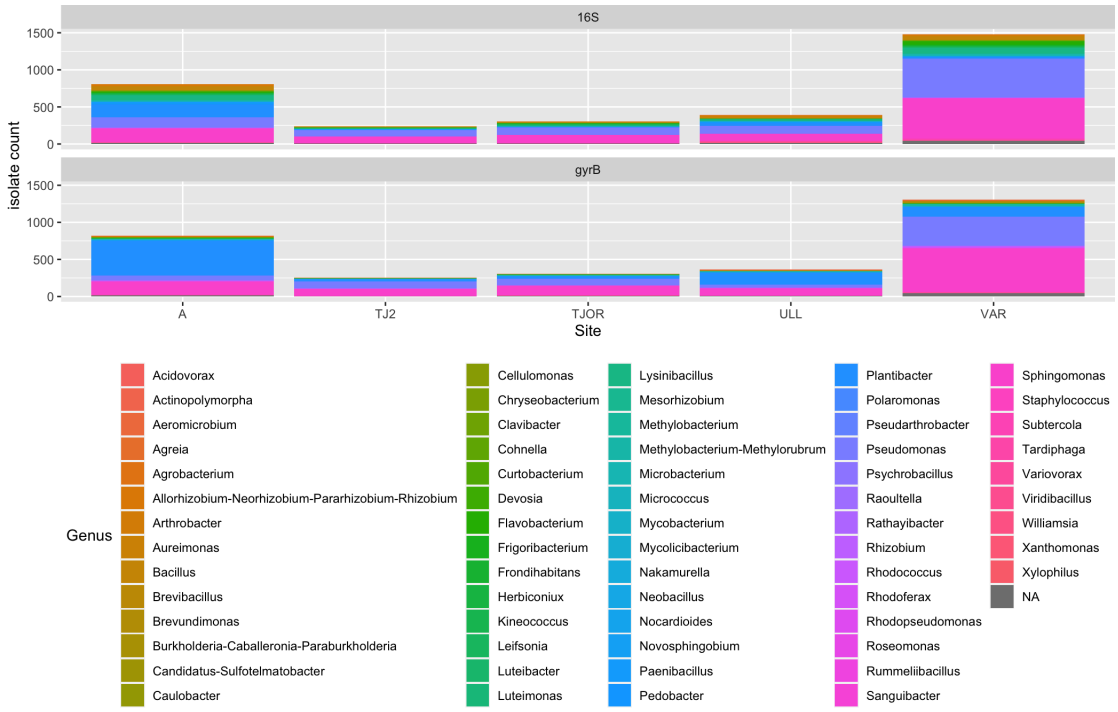

Figure 2.1: Comparison of *gyrB* and 16S taxonomic composition of isolates collected from 100 *A. thaliana* rosette leaves of 50 plants taken from a total of 5 locations across Sweden.

Broth (NB) in matrix tubes. Liquid cultures grew at 28°C until turbid. We then mixed the liquid cultures in glycerol for a final percentage of 15-25% and stored in -80°C.

We collected 4,128 isolates. We generated *gyrB* and 16S v5-v7 (hereafter 16S) sequences for approximately 3,000 of the isolates. The taxonomic compositions of the 16S and *gyrB* genera are similar, with the exception that the *gyrB* database classified notably more *Plantibacter* compared to 16S. We identified 568 *gyrB* and 486 16S distinct amplicon variants. All data, including site, plant, leaf number, sequence, and media used, is available for each isolate in an online database developed in-house through the R Shiny package at https://hannahwhitehurst.shinyapps.io/isolate-tracker/.

2.2.2 Sampling recovered a substantial portion of the anticipated diversity, but not all

To assess the success of our culture methods in capturing the complete taxonomic diversity, we planned to compare the diversity of our isolates to culture-independent taxonomic profiles of the leaf samples using amplicon sequencing of the glycerol and leaf mixture. However, we had limited material and were unable to sequence the diversity of the macerated leaf material directly. Instead, we turned to four data collections on 3,515 leaf-endophyte communities taken from similar locations in prior years: Ullstorp (2011, 2012) and Adal (2011, 2012). We combined the read counts of the 16S operational taxonomic units (OTUs) identified in the Ullstorp and Adal data sets across both years, filtering out any OTUs representing less than 0.01% of the read counts. The Ullstorp and Adal sets, when combined, had 144 OTUs with greater than 0.01% read counts; we found 82 of the OTUs (>97% identity). The rate of successfully culturing isolates varied across the taxa (Table 1.1). Our sampling methods were excellent at recovering Pseudomonadacea, Sphingomonadacea, Rhizobiaceae, and Beijerinckiaceae, with over 80% of the expected OTUs found. Our sampling methods were less successful in collecting isolates in the Burkholderiaceae, Kineosporiaceae, and Nocardioidaceae families, with less than 50% of the anticipated OTUs found.

2.2.3 Acquisition of plant pathogens for future studies

We used the BacDrive database (Reimer et al. 2022) to generate a list of plant pathogen species (n=179). Our isolate collection contains 14 of the 179 putative pathogen species (608/4,128 isolates).

We performed whole genome sequencing on a selection of isolates through Illumina shotgun sequencing and assembled genomes with SPAdes (Bankevich et al. 2012). We analyzed 104 high quality whole genome assemblies (>100,000 bp n50, <10% redundancy, >90% complete as quantified by QUAST (Gurevich et al. 2013) and anvi'o (Eren et al. 2015), filtering

	Found	Unfound	% Found
Acetobacteraceae	0	2	0.0
Beijerinckiaceae	8	1	88.9
Burkholderiaceae	15	32	31.9
Caulobacteraceae	4	0	100.0
Deinococcaceae	0	2	0.0
Devosiaceae	3	1	75.0
Flavobacteriaceae	5	0	100.0
Geodermatophilaceae	0	1	0.0
Kineosporiaceae	1	3	25.0
Methylophilaceae	0	2	0.0
Microbacteriaceae	4	1	80.0
Nocardiaceae	0	1	0.0
Nocardioidaceae	2	7	22.2
Polyangiaceae	0	1	0.0
Pseudomonadaceae	7	0	100.0
Rhizobiaceae	9	2	81.8
Solirubrobacteraceae	0	1	0.0
Sphingomonadaceae	20	2	90.9
uncultured	0	1	0.0
Weeksellaceae	0	1	0.0
Xanthobacteraceae	5	0	100.0
Total	**83**	**61**	**57.6**

Figure 2.2: Top OTUs (>0.01% relative abundance, n= 142) identified through culture-independent taxonomic classification of 3515 *A. thaliana* plants collected across two years in Adal and Ullstorp, Sweden, and the number of isolates that were found through culturing (>97% percent identity)

Putative Pathogen	n isolates
Clavibacter insidiosus	1
Clavibacter michiganensis	26
Curtobacterium flaccumfaciens	12
Dickeya solani	1
Pseudomonas amygdali	1
Pseudomonas savastanoi	2
Pseudomonas syringae	271
Pseudomonas tolaasii	5
Pseudomonas viridiflava	268
Rathayibacter tritici	3
Rhodococcus fascians	11
Sphingomonas melonis	3
Spiroplasma citri	3
Xanthomonas albilineans	1

Table 2.1: Putative pathogens in Swedish isolate collection

all contigs <3000 bp).

2.2.4 Validating the ecological significance of B38 on A. thaliana

In addition to using bacterial isolates for testing plant-pathogen interactions, we can begin to better understand beneficial and neutral interactions. Previous members of our lab identified potential beneficial microbes in *A. thaliana* leaf endophyte communities using culture-independent methods (Brachi et al. 2022). They used 16S amplicon sequencing to generate co-occurrence data of microbes among 1100 plants at each of four locations in Sweden, replicated over two years. They then identified ecologically significant bacteria through network analysis in combination with host fitness estimates. While several bacteria were identified as potentially beneficial to the host, the correlations needed to be empirically tested. We found several of the putative bacteria hubs in our isolate collection (Appendix A, Table 7.1) and selected one of these hub isolates, B38, for single-inoculation testing to determine host effects.

We first performed shotgun sequencing and genome assembly for B38. We identified B38

as *Brevundimonas sp.* using amplicon sequences and core gene similarities using anvi'o (Eren et al. 2015). The final genome had approximately 130x coverage, with an estimated size of 3,753,256 across 8 contigs (N50=771,495). We did not find evidence of effectors, virulence factors, or antibiotic resistance genes (percent identity matches >97% using ABRicate, NCBI VFDB, and publication databases (Seemann 2023; Chen et al. 2016; Dillon et al. 2019). See Methods for details.

We selected an *A. thaliana* host genotype (#6136) that harbored B38 in the field experiments. We sterilized seeds and planted them on 1/2 MS media in 12-well plates, then we transferred the plates to a growth chamber for two weeks to germinate and grow (Methods). We randomly assigned each of 600 plants to either a B38 drip inoculation group or a control drip inoculation group. We measured the plant area immediately prior to treatment, 7 days post inoculation, and again at 14 days post inoculation using a custom python script (Methods). After accounting for within and between plate variation, plants inoculated with B38 showed an increase of 5.375 (SE = 1.973) mm^2 new plant growth compared to control plants when measuring growth between days 7-14 (F= 7.3981, df = 1, P value = $6.7e^{-3}$); corresponding to a 10.22% increase in growth.

2.3 Discussion

We used a variety of media comprised of a spectrum of nutrients in an attempt to increase the diversity of microbes we collected. Still, we did not capture the entire breadth of microbial taxonomic diversity that is indicated from previous amplicon sequencing of microbial communities from similar plants. This is likely a limitation of the selected media, humidity, temperature, and sampling amounts. We increased sampling depth as reported in previous collection efforts, but yielded similar percentages of recovery of the most abundant microbe taxa (Bai et al. 2015). Future isolation efforts may consider adding additional media types to improve isolation.

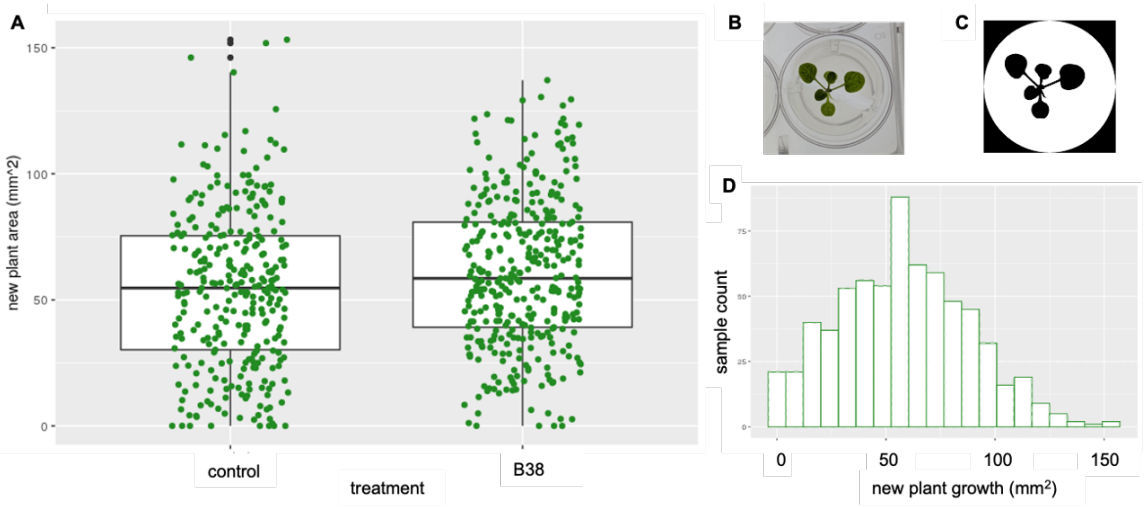

Figure 2.3: B38 effects on host growth. New plant growth of plants treated with either the B38 or control inoculum was measured between 7-14 days post treatment (mm^2) separated by treatment (A) and shown together (D), showing significant increase in growth for plants treated with B38 after controlling for within and between plant variation (F= 7.3981, df = 1, P value = $6.7e^{-3}$). Measurements were made using a custom python script, which took input images taken of wells (B) and converted to cropped, binary images (C).

Some isolates did not successfully generate an amplicon sequence. This could be in part due to genomic DNA loss through the high-throughput methods. Additionally, Karasov et al. (2019) showed preliminary findings using shotgun sequencing in metagenomics samples that Swedish populations have up to 1.5X oomycetes to host plant DNA; we anticipate that many of the unidentified isolates are oomycetes and not compatible with the primers used in this study.

We successfully collected a variety of diverse and ecologically significant microbes identified in previous studies. We empirically validated the host effects of one such hub, *Brevundimonas sp.*, B38, in the lab, which we could not have done without the isolate collection. We demonstrate the value of culture collections in validating hypotheses developed through culture-dependent ecological studies; the resulting publication is provided at the end of this chapter. In addition to identification of beneficial microbe candidates, we also identify 14 distinct putative pathogens, which totaled 607 isolates.

The collection provides a broad resource for microbial community studies in the lab. For example, we have already used the collection in identifying primer bias against a common *Plantibacter* isolate, shown in Chapter 3. The *Plantibacter* microbe was not identified as abundant using standard 16S primers, but through genomic sequencing and multi-locus amplicon sequencing, we identified the base-pair mismatch and have switched primers for future microbiome community analysis. Thus, we illustrate the potential of this collection to facilitate ecological and technical studies in the lab.

The online application allows researchers to easily identify isolates of interest using meta data parameters or BLAST identity matches. For metadata, users can filter isolates found by location, and identify isolates that came from the same plant or leaf. For example, if a user wanted to study genomic variation within co-occurring microbes of the same species, they could upload a BLAST sequence, find matches, then filter for isolates found in the same location, plant, and leaf. We have shared the application script online for free usage of the inventory tracking tool in lieu of expensive laboratory inventory management systems and to encourage accessibility of lab isolate resources.

2.4 Methods

All data was analyzed using R version 4.1.1 (R Core Team 2021) unless otherwise noted.

2.4.1 Sample Collection

Plant populations were identified in five locations in Sweden: Ullstorp, Varhallen, Tjor, Tjor-2 and Adal. For each of 10 plants in each population, we collected 2 leaves, surface sterilized with sterile ddH20 followed by two brief ethanol washes, followed by one additional rinse in sterile water. Leaves were macerated and stored in 20% glycerol on ice in the field, then -80°C until processing.

2.4.2 Media preparation

Media components (Appendix A, Table 6.1) were mixed in 1L batches and autoclaved at 121°C for 40 minutes. Once cool enough to handle, we poured 25 mL into each petri dish and dried. In a sterile laminar flow, we plated 45 ul of the leaf glycerol mixture using wide-mouth pipette tips onto each respective media, then spread with a sterile glass spreader. We wrapped Parafilm around plates before placing them in an incubator at 28°C. We made controls of each media, and if any colonies formed on the controls, the sample was discarded and redone.

2.4.3 Isolate propagation and storage

We checked media plates daily for three weeks. Each new colony was restreaked onto new plates of the respective media. Once new CFUs formed, a colony was stabbed and propagated into 250 ul liquid Nutrient Broth in matrix tubes, incubated at 28°C in an orbital shaker at 280 RPM. Once the sample achieved turbidity, glycerol was added to a final concentration of 15-25%, gently mixed, and stored at -80°C.

2.4.4 DNA extraction of isolates and amplicon sequencing

We pelleted turbid, 250ul cultures of each isolate by centrifuging 6600g for 15 minutes. We then added 255ul of the lysozyme mixture (250 ul TES, 350 units NEB Lysozyme Ready Lyse #R1810M, 3.5 ul RNAse A), mixed, and incubated at room temperature for 30 minutes. We then added a mixture of 250 ul of TES + 2% SDS (10 mM Tris-HCl pH 8, 1 mM EDTA, 100 mM NaCl + 2% SDS (w/v)) and 1 mg/mL Proteinase K (NEB #P8107S), and incubated at 55°C for 4 hours. We then added 300ul 5M NaCl to precipitate protein/SDS complexes, mixed, then centrifuged at 7000G for 5 minutes. We withdrew the clear supernatant into a fresh plate, then added 300ul Solid Phase Reversible Immobilization Beads and incubated for 10 minutes at room temperature. The samples were flash spun and placed on a magnetic

stand for 1 minute. We withdrew the clear supernatant and discarded. We rinsed the beads three times with 80% ethanol, let dry for 5 minutes, added 150 uL molecular grade water, mixed, and placed back on the magnetic stand for 5 minutes. The clear supernatant was transferred to a sterile plate for library preparation. We performed two-step PCR as described in Illumina, using custom primers and internal barcodes as described in Chapter 3 and Appendix B.

2.4.5 ASV tallies

We, trimmed reads, estimated error rates (learnErrors), generated dereplicated sequences, applied the DADA2 algorithm, merged reads (maxMismatch =0, minOverlap = 10), then removed chimeras using DADA2 (Callahan et al. 2016). The DADA2 parameters for optimal maximum error rates and trim positions for 16S and *gyrB* amplicons for each flowcell were determined through FIGARO (Weinstein et al. 2019); however, *gyrB* amplicon sizes were below the software threshold and additional adjustments were required. See scripts for details. We grouped reads if there was read length variation or alignment shifts but were otherwise identical using DADA2 “collapseNoMismatch". We then merged ASV tables for each amplicon across sequencing runs and classified sequences with the DADA2 naïve baysian classifier using the SILVA 16S database, nr99_v138.1 (Quast et al. 2013) or an in house *gyrB* database for 16S and *gyrB* identification, respectively.

2.4.6 Isolate genome sequencing and assembly and pathogen analysis

We generated Illumina Nextera sequencing libraries using standard protocols (Illumina FC-121-1030). Libraries were sequenced on the Illumina Miseq using PE 300 v3 chemistry (Illumina MS-102-3003). We trimmed reads for adapters using the standard recommended BBduk parameters (ktrim=r k=23 mink=11 hdist=1) and assembled genomes using SPAdes (kmer=21, min=500). We verified genome completeness using anvi’o (anvi-estimate-genome-

completeness) and contig assembly metrics with QUAST. Assembled genomes were blasted against the effector database generated by (Dillon et al. 2019) using NCBI BLAST command line tool and retaining all hits with percent matching identity >97% .Genomes were also assessed for potential virulence factors using ABRigate with a threshold set at >97% identity.

2.4.7 B38 inoculation on A. thaliana

. We prepped 1/2 MS media (Murashige & Skoog medium including Nitsch vitamins, bioWORLD containing 500 mg/L MES (2-Morpholinoethanesulfonic acid hydrate), pH 5.7 to 1.8.) and poured 1.5 mL into each well of 24-well plates. We surface sterilized seeds from *A. thaliana* genotype 6136 using chlorine gas incubation, and placed 1-2 seeds in each well. We let the seeds vernalize for 4 days at 4°C, then placed 24-well plates in a growth chamber with 16 hours of light at 16°C. The plants were treated with either B38 or control inoculum between days 13 and 15 in the chamber.

B38 cultures were grown in R2A liquid media for approximately 3 days until reaching turbidity (OD600=0.2). We centrifuged cultures at 1,800 relative centrifugal force (RCF) at 18°C for 7 minutes, decanted, and suspended the pellet in 0.1 M MgSO4. We randomly selected an equal number of plants for B38 inoculation (B38 + 0.1 M MgSO4) or control (0.1 M MgSO4) treatment. We pipette dripped 180ul of the respective treatment inoculation in a sterile laminar flow hood. Plates were sealed with Micropore tape before returning to the growth chambers. We took photo measurements of each plant prior to inoculation, then days 7 and 14 post inoculation. The plates showed signs of high humidity; we scored the plants blindly for waterlogged characteristics (in regard to B38 or control treatment) by categorizing leaf morphology. Curled, white/translucent leaves, and stunted plants were removed from analysis, for a loss of 422 plants. We wrote a custom python script which quantified plant surface area in each well by scaling based on the wells' diameter. We then converted images into a binary format and measured nonwhite pixels within each well. We manually inspected

images, and for any images that were not accurately processed, we used ImageJ (Schneider, Rasband, and Eliceiri 2012) to perform the described pipeline.

2.5 Supplementary

The following paper (Brachi et al. 2022) was associated with work described in this chapter. Reproduced with permission by the *Proceedings of the National Academy of Sciences* (PNAS) default license.

Brachi, Benjamin et al. (2022). "Plant genetic effects on microbial hubs impact host fitness in repeated field trials". In: *Proceedings of the National Academy of Sciences of the United States of America* 119.30. issn: 10916490. doi: 10.1073/pnas.2201285119.

PNAS INAUGURAL ARTICLE | GENETICS OPEN ACCESS

Plant genetic effects on microbial hubs impact host fitness in repeated field trials

Benjamin Brachi[a,b], Daniele Filiault[c,1], Hannah Whitehurst[a,1], Paul Darme[a], Pierre Le Gars[a], Marine Le Mentec[a], Timothy C. Morton[a], Envel Kerdaffrec[c], Fernando Rabanal[c], Alison Anastasio[a], Mathew S. Box[d], Susan Duncan[d], Feng Huang[a,e], Riley Leff[a], Polina Novikova[c], Matthew Perisin[a], Takashi Tsuchimatsu[c], Roderick Woolley[a], Caroline Dean[d], Magnus Nordborg[c], Svante Holm[f], and Joy Bergelson[a,g,2]

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2018.

Contributed by Joy Bergelson; received January 31, 2022; accepted June 3, 2022; reviewed by Peter Balint-Kurti, Thomas Mitchell-Olds, and Venkatesan Sundaresan

Although complex interactions between hosts and microbial associates are increasingly well documented, we still know little about how and why hosts shape microbial communities in nature. In addition, host genetic effects on microbial communities vary widely depending on the environment, obscuring conclusions about which microbes are impacted and which plant functions are important. We characterized the leaf microbiota of 200 *Arabidopsis thaliana* genotypes in eight field experiments and detected consistent host effects on specific, broadly distributed microbial species (operational taxonomic unit [OTUs]). Host genetic effects disproportionately influenced central ecological hubs, with heritability of particular OTUs declining with their distance from the nearest hub within the microbial network. These host effects could reflect either OTUs preferentially associating with specific genotypes or differential microbial success within them. Host genetics associated with microbial hubs explained over 10% of the variation in lifetime seed production among host genotypes across sites and years. We successfully cultured one of these microbial hubs and demonstrated its growth-promoting effects on plants in sterile conditions. Finally, genome-wide association mapping identified many putatively causal genes with small effects on the relative abundance of microbial hubs across sites and years, and these genes were enriched for those involved in the synthesis of specialized metabolites, auxins, and the immune system. Using untargeted metabolomics, we corroborate the consistent association between variation in specialized metabolites and microbial hubs across field sites. Together, our results reveal that host genetic variation impacts the microbial communities in consistent ways across environments and that these effects contribute to fitness variation among host genotypes.

Arabidopsis thaliana | genome-wide association study | microbiome | fitness | microbial hubs

Hosts harbor complex microbial communities that are thought to impact health and development (1). Human microbiota has been implicated in a variety of diseases, including obesity and cancer (2). Efforts are thus underway to determine the host factors shaping these communities (3, 4), and to use next-generation probiotics to inhibit colonization by pathogens (5). Similarly, in agriculture, there is great hope that selection on plant traits shaping the composition of the microbiota will help mitigate disease and increase crop yield in a sustainable fashion. Indeed, the Food and Agriculture Organization of the United Nations has made the use of biological control and growth-promoting microbial associations a clear priority for improving food production (6).

Plant-associated microbes can be beneficial in many ways, including improving access to nutrients, activating or priming the immune system, and competing with pathogens. For example, seeds inoculated with a combination of naturally occurring microbes were found to be protected from a sudden-wilt disease that emerged after continuous cropping (7). Thus, it would be advantageous to breed crops that promote the growth of beneficial microbes under a variety of field conditions, a prospect that is made more likely by the demonstration of host genotypic effects on their microbiota (8–11). However, microbial communities are complex entities that are influenced by the combined impact of host factors, the abiotic environment, and microbe–microbe interactions (12). Indeed, several studies have found a strong influence of the environment on estimates of host genotype effects (8, 13, 14). Although most, if not all, studies exploring the influence that host genotype exerts on microbial communities suggest that such plant control could be beneficial to plant performance, almost nothing is known about the relationship between host genotype effects on microbial communities and on plant performance or fitness. Consequently, the extent to which host plants can control microbial communities to their advantage, especially in a consistent manner across multiple environments, remains unclear.

Significance

Recent demonstrations of a genetic basis for variation among hosts in the microbiome leave unresolved the question of how commonly host genetic effects influence individual microbes, and whether these effects impact host fitness. We used replicated field studies in the north and south of Sweden to map host genetic effects in microbial community networks using genome-wide association mapping. By focusing on consistent effects across sites, we found effects of genetic variation on important microbial hubs that contributed to plant fitness in a manner robust to the environment. Our results suggest that ongoing efforts to harness host genotype effects on the microbiome for agricultural purposes can be successful and highlight the value of explicitly considering abiotic variation in those efforts.

Reviewers: P.J.B., Agricultural Research Service, US Department of Agriculture; T.M.-O., Duke University; and V.S., University of California, Davis.

The authors declare no competing interest.

[1]D.F. and H.W. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: jb7684@nyu.edu.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2201285119/-/DCSupplemental.

Published July 22, 2022.

Here, we combine large-scale field experiments in natural environments, extensive microbial community analysis, and genome-wide association mapping to 1) determine how host genotype affects different microbial community members, and thus shapes the overall microbiome; 2) estimate host genotype effects on microbial communities across eight environments and investigate the contribution of those effects to the performance of plant genotypes; and 3) use genome-wide association mapping to identify key pathways that shape the leaf microbial communities across multiple environmental conditions.

Snapshot of Microbial Community Variation

We performed a set of field experiments that included natural inbred lines of *Arabidopsis thaliana* (hereafter "accessions") originally collected throughout Sweden, mainly in two climatically contrasted regions of the country (Dataset S1); *A. thaliana* in the north of Sweden experiences long, snowy winters, and, as a consequence, plants are typically found on south-facing slopes of rocky cliffs. *Arabidopsis* populations in the south of Sweden, on the other hand, tend to be associated with agricultural or disturbed fields that experience highly variable snow cover over the winter months. We used replicate experiments in four representative *Arabidopsis* sites, two each in the north (sites NM and NA) and south (sites SU and SR) of Sweden. Experiments were repeated across 2 years, for a total of eight experiments.

Each experiment was organized in a complete randomized block design including 24 replicates of 200 sequenced accessions (15), established as seedlings in a mixture of 10% native and 90% potting soil and timed to coincide with local germination flushes in late summer. Many of the microbiome members from our experiments were also found within the leaves of *A. thaliana* plants that we collected in the field in southern Sweden in 2017, suggesting that this percentage of native soil was sufficient to seed a representative microbiome (Dataset S2). Immediately upon snowmelt in early spring, we sampled and freeze dried five or six whole rosettes per accession. DNA was extracted from the freeze-dried rosettes, and both the ITS1 portion of the *Internal Transcribed Spacer* (*ITS*) and the V5 to V7 regions of the *16S* RNA gene were sequenced to characterize the fungal and bacterial communities, respectively (9, 12, 16). The sequences obtained were clustered into operational taxonomic units (OTUs) using Swarm to generate community matrices (17) (see *Count Table Filtering*). The frequency distributions of OTUs were highly skewed, with the top 10 most common OTUs contributing, on average, 59% of the reads in each experiment (ranging from 45 to 78%). Throughout this study, we chose to focus on the microbes represented by at least 0.01% of the sequencing reads per experiment. While rare microbes may impact host performance and have important ecological roles (18), we would not have had the power to estimate heritability or map host control of these species. Taxonomic assignments indicate that the fungal communities were dominated by Leotimycetes and Dothideomycetes, while the bacterial communities included high proportions of Alphaproteobacteria and Actinobacteria (*SI Appendix*, Fig. S1).

In a principal coordinate (PC) analysis, differences between northern and southern sites explained 10% and 5% of the overall diversity in the fungal and bacterial communities, respectively, while differences between the two consecutive years explained 5% and 3%. This level of differentiation among experiments likely underestimates that present in the native soil, as it has been shown that hosts filter the microbial community to reduce site-to-site differences (19, 20) (Fig. 1). In addition, there may have been a homogenizing effect of using a combination of local and potting soil. Irrespective of how well our treatments mimicked natural microbial communities, our analysis of eight common garden experiments permits assessment of the consistency across time and space of plant genetic effects on their associated microbial communities.

Host Genetic Effects on the Microbiota

Our experiments provided a unique opportunity to investigate associations between host genetic variation and their resident microbiomes, within the context of environmental variation across time and space. We focused on PC from simple unconstrained PC analysis (PCoA) within each experiment in order to summarize the variation among communities including hundreds or thousands of species with a few dimensions, and then calculated the proportion of variance explained by the host genotype (hereafter heritability or H^2). Within each experiment, we found significant heritability of PC of the microbial communities (*SI Appendix*, Table S1), suggesting that genetic variation in the host significantly impacts at least a fraction of the microbiota, in line with results of previous studies (8–10, 12, 21, 22).

Significant heritability of the resident microbiome could arise from host genotypes exerting weak control over many community members, or by targeting a few microbes that then influence the relative abundance of others through microbe–microbe interactions. In order to investigate these hypotheses, we modeled the log-ratio transformed counts of individual OTUs with random-effect linear models and revealed significant genotypic effects (with the 95% CI of heritability not overlapping zero) for between 10.13% and 21.93% of all OTUs, depending on the site and year (Fig. 2 *A–D* and *SI Appendix*, Fig. S2 *A–D*). The latter explanation thus seems more likely, given that the influence of the host appears focused on relatively few OTUs, although it remains to be investigated whether heritable microbial hubs influence other members of the microbiome (see below). We found no evidence that either fungal or bacterial communities are systematically more impacted by host effects than the other (Fig. 2 *A–D* and *SI Appendix*, Fig. S2 *A–D*), nor that mean relative abundance was strongly correlated with OTU heritability (*SI Appendix*, Fig. S3).

Host Genetics Correlate Most Strongly with Ecologically Central Microbes

Having found that host effects are concentrated on a small proportion of OTUs, we investigated the possibility that these heritable OTUs trigger a broader community-level change in the microbiota. First, we computed networks of microbe cooccurrence for each experiment. We explored the ecological importance of heritable OTUs by computing networks of microbe cooccurrence for each experiment using the SPIEC-EASI (SParse InversE Covariance Estimation for Ecological Association Inference) pipeline (23). Although our networks included both fungal and bacterial OTUs, most significant cooccurrences involved OTUs within each domain, with an average of only 7.76% (min = 6.64%, max = 9.91%) of edges connecting fungal and bacterial OTUs. We quantified the ecological importance of OTUs using two common characteristics of nodes in a network ("degree" and "betweenness centrality") (12), defining ecologically important "hubs" in each network as OTUs in the 95% tail of both of these statistics (*SI Appendix*, Fig. S4). We identified, on average, 16.5 microbial hubs per experiment (ranging from 11 to 24), representing 78 unique OTUs across all eight experiments (43 bacterial OTUs and 35 fungal OTUs). These hubs were

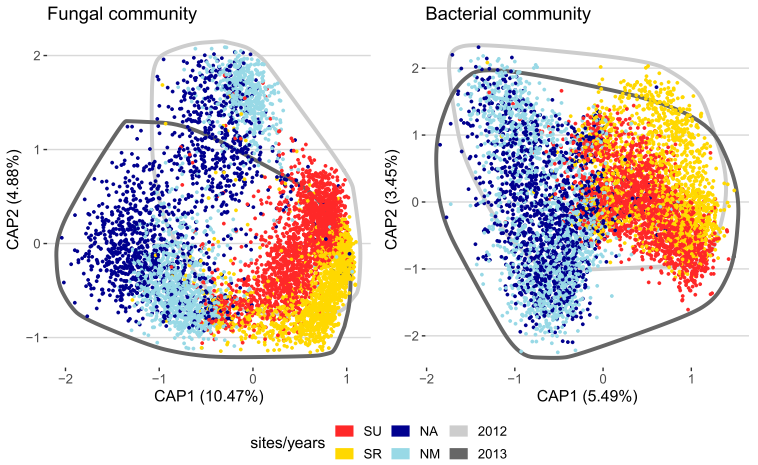

Fig. 1. Plants grown in different environments have different microbial communities. The plots represent the projection of each sample on the plane defined by the first two constrained components of the fungal and bacterial communities, describing variation among sites and years. The percentages in parentheses are the proportion of the total inertia (square root of the Bray–Curtis dissimilarity) explained by each component. The colors of the points indicate the site from which samples were collected. Experiments from the south are represented in red (SU) and yellow (SR), and experiments from the north are represented in blue (NR) and dark blue (NA). All points from 2012 and 2013 are encircled by a darker and lighter gray line, respectively.

connected to an average of 19.62% (min = 14.50%, max = 25.23%) of the edges in the networks, indicating that they are likely important in structuring the microbial community. In addition, hubs were involved in proportionally more interactions between fungi and bacteria than the rest of the community (*SI Appendix*, Table S3).

Next, we asked whether heritable OTUs are more likely to be ecologically important hubs, because this could open the door to community-level impacts of host genetic variation. Across all eight experiments, we detected 23 OTUs that were both heritable and hubs at least once (*SI Appendix*, Table S2 and Dataset S2). This represents a significant enrichment of hub OTUs among heritable OTUs (Wilcoxon rank sum test: $n = 8$, W = 57, P value = 0.007), suggesting that host effects on the microbiota preferentially influence the relative abundance of ecologically important microbes. In fact, hub OTUs were often among the OTUs with the highest heritability within each experiment; these hub OTUs stand out in that we find no general relationships between heritability and either betweenness or degree (*SI Appendix*, Fig. S5). To further explore how heritability is distributed among members of microbial communities, we mapped broad-sense heritability onto the ecological network. In six out of eight experiments, we observed a significant negative relationship between heritability and the distance (number of network edges) to the nearest heritable hub (combined P value = $3.96e^{-25}$, using Fisher's method for combining P values) (24) (Fig. 2 *E–H* and *SI Appendix*, Fig. S2 *E–H*). This pattern reveals that host genetic variation impacts the structure of microbial communities, although whether this occurs due to shared host effects on many microbes or host effects on hubs that then percolate in the microbial community through microbe–microbe interactions is unclear.

To discern the contribution of microbe–microbe interactions in the propagation of host genetic effects across the microbial networks, we took advantage of our replicates of each host genotype to permute counts for each OTU. We reasoned that, if microbe–microbe interactions were largely responsible for the patterns of cooccurrence that we observed, then microbial cooccurrences would be diminished by our permutations of replicates within host genotypes. The same diminution would be evident if patterns of microbial cooccurrence were due to microenvironmental variation within experiments, independent of host genotype, although strong microenvironmental effects would have interfered with our ability to detect heritable OTUs. On the other hand, if OTUs tended to cooccur due to shared host genotype effects, then our permutations would have little impact. In the networks computed from the permuted datasets, on average, 91.39% (ranging from 87.67 to 95.2% across our eight experiments) of all OTUs that previously cooccurred with at least one other OTU (with degree > 0) had fewer associations with other microbes. Overall, networks computed from the permuted data had, on average, 75% fewer edges (ranging from 62 to 87%). This indicates that most microbe–microbe associations were not due to shared host genotype effects. Thus, although host genetic variation drove the cooccurrences for a fraction of OTUs, we interpret our empirical networks as consistent with a shared role of host genetics and microbe–microbe interactions, with host genotypes most strongly impacting microbial hubs that then influence other members of the microbial communities.

Not only did the heritable hubs seem to have an impact that percolated through the microbial community, they were widely distributed among accessions, sites, and years. We were able to identify 127 fungal and bacterial OTUs that were found in at least 50% of samples in all experiments. Interestingly, OTUs that were heritable hubs at least once were overrepresented in this core microbiota ($\chi^2 = 51.98$, degree of freedom [df] = 1, P value = 5.58e-13). This was not an artifact of their being widespread; significant heritability estimates were detected across the entire range of prevalence. Indeed, prevalence of OTUs explained less than 2.6% of variation in OTU heritability across all experiments (F statistic = 110.66, df = 4176, P value < 2.2e-16; *SI Appendix*, Fig. S6). Thus, ecologically important OTUs with greatest associations to host genotypes were unusual in being widespread among plants in multiple experiments. Host effects on the fungal OTU #8 (hereafter F8) are especially important; this OTU showed significant heritability ($H^2 > 0$) in five

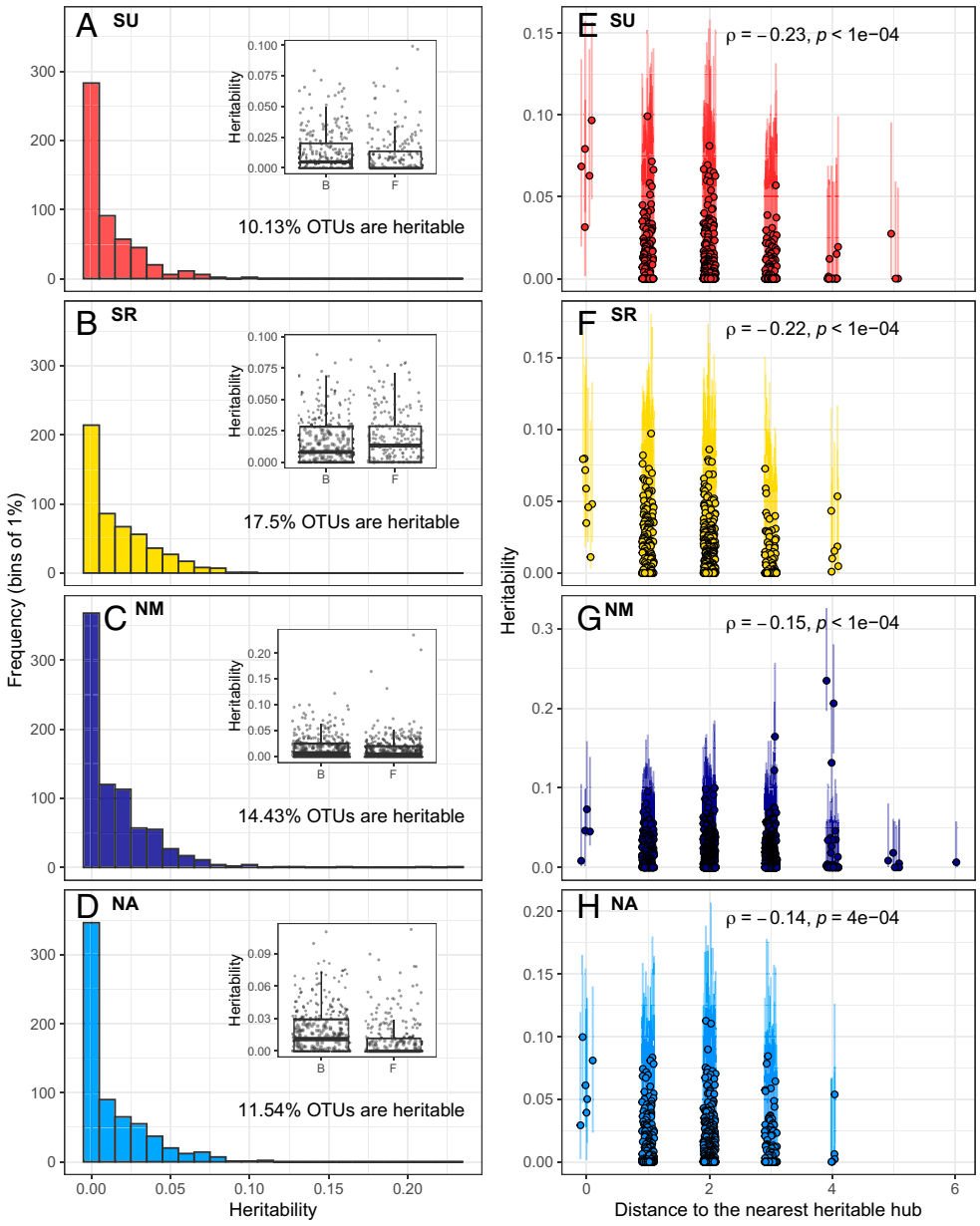

Fig. 2. The effect of host genetic variation on the microbial community targets relatively few OTUs and percolates across the network. This figure corresponds to observations in the set of four experiments performed in 2013. The same figure is available for the 2012 experiments in *SI Appendix*, Fig. S3. (*A–D*) Each frame presents the distribution of heritability estimates for individual OTUs in one site. In each frame, *Inset* graph is a box and whiskers plot contrasting the heritability (*y* axis) of bacterial (*B*) and fungal (*F*) OTUs. (*E–H*) The heritable hubs are represented at a distance of zero (hub). The other points are OTUs connected to heritable hubs, directly (distance = 1) or indirectly (distance > 1). The *x* axis represents the number of edges in the network separating an OTU and its nearest heritable hub. The correlation coefficients presented are Spearman rank correlations between heritability and distances to the heritable hub(s) (including zero).

out of the seven experiments in which it was a hub (*SI Appendix*, Table S2), suggesting that natural variation in *A. thaliana* influences its microbiota with some consistency across environments. The widespread prevalence of these heritable hubs suggests that variation at particular host genes associates with particular hubs across time and space, potentially providing a means to impact the microbiota in a robust fashion.

Variation in Performance of Host Genotypes Explained by Their Influence on Microbial Hubs

The extent to which natural variation among host genotypes in their associated microbes translates into fitness differences has yet to be determined. Our experiments included additional replicates of all genotypes that were left to flower and mature in the field. We harvested mature stems in early summer and used high-throughput image analysis to measure the size of reproductive stems, an estimate of lifetime investment in reproduction in this annual species. This measure encompasses variation in both the number of siliques and their size (which can increase as a function of seed number and seed size) but correlated well with seed production in an independent experiment (*SI Appendix*, Fig. S7) (24). We thus call our estimate "seed-set" in what follows. We observed that plant seed-set estimates were positively correlated across experiments (*SI Appendix*, Fig. S8), suggesting fitness variation among accessions was relatively consistent across sites. We therefore asked whether host effects on microbial hubs contributed to some genotypes producing more seeds across all environments investigated. Specifically, we used random intercept models to estimate genotype effects on both heritable microbial hubs and seed-set in a series of analyses that jointly considered all eight experiments and investigated the relationship between these two effects (see *Heritable Hubs and Seed-Set across Environments*).

We found that the host genotype explained, on average across experiments, 6.88% (with a 95% CI [5.52, 8.34]) of seed-set. Host genotype effects on the relative abundances of 19 of our 23 heritable microbial hubs, quantified as random intercept deviation, were similarly modest, explaining up to 4% of the variation (Fig. 3*A*; four heritable hubs were not detected in more than two experiments and were removed for this analysis). We used multiple regression to estimate genetic correlations between host genotype effects on seed-set and on microbial hubs. We detected positive correlations between accession effects on seed-set and accession effects on three heritable hubs, F8, B38, and B13, as well as a negative correlation between accession effects on seed-set and accession effects on F5 (Fig. 3*B*). The variation explained by host genotype on the relative abundances of microbial hubs explained 12.4% of the host genotype effects on seed-set.

These results reveal that a sizable percentage of genetic variance in seed-set is shared with genetic variation associated with the relative abundance of a few broadly distributed microbial hubs, consistent with a causal relationship between genotype and seed-set mediated by heritable microbial hubs. Of course, the proportion of shared genetic variation between seed-set and heritable microbial hubs is unlikely to be equally important across time and space. In fact, in analyses performed on an experiment-by-experiment basis, we found that relationships between host effects on hubs and on seed-set were stronger in southern Sweden, where we detected significant relationships in both sites and both years (*SI Appendix*, Table S4).

Overall, our results highlight the importance for plants of controlling their leaf microbial community and suggest that breeding plants for their effects on specific members of microbial communities has the potential to significantly increase plant productivity.

Effect of Hubs on Growth in Controlled Condition

In an effort to confirm that the genetic correlations observed between heritable hubs and plant seed-set were due to an interaction between host and microbial species, we returned to the field to collect wild *A. thaliana* leaves, cultured ~3,900 bacterial isolates from within these leaves (25), and sequenced both the 16S

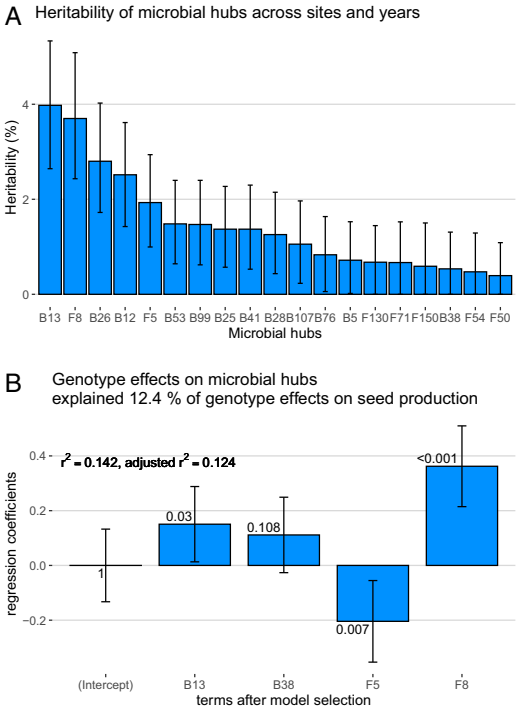

Fig. 3. Relationship between host genotype seed-set and influence on microbial hubs across sites and years. (*A*) Proportion of heritable hub relative counts explained by host effects across all sites and years. (*B*) Coefficients for the linear regression explaining lifetime seed production variation among accessions with accession effects on microbial hubs across experiments (after model selection). The numbers near each bar are the *P* values associated with each term.

RNA gene and gyrase-B. These sequences included 100% matches for 10 of the 43 bacterial hubs, among which 4 were heritable hubs (*SI Appendix*, Table S2). Among successfully cultured heritable hubs was B38 which appeared to contribute positively to the seed-set of accessions in our field experiments (Fig. 3). This isolate derived from Vårhallarna, in southern Sweden (*SI Appendix*, Table S5). We subsequently performed shotgun whole genome sequencing of B38 which we identified as *Brevundimonas sp.* The assembled and annotated genome did not identify putative pathogenic or virulence genetic factors present in the genome.

If there is an interaction between B38 and the host, the growth-promoting effect of B38 could be either direct or indirect, mediated through other members of the community. To test the direct effect of B38 on host growth, we grew *Arabidopsis* plants of an accession (#6136) from the south of Sweden chosen to have intermediate relative abundance of B38 in the field. Plants were grown under sterile conditions in 1/2 MS media under long-day conditions in the growth chamber, with and without B38 inoculation. Approximately 2 wk after germination, over 600 plants were randomly selected for either drip inoculation with the control or B38 inoculum, and measured for surface area growth over the following 2 wk. Accounting for variation in plant growth among trials and plates within trials, we found that plants treated with B38 grew 5.375 (SE = 1.973) mm^2 larger than control plants ($F = 7.3981$, df = 1, P value = $6.7e^{-3}$) between days 7 and 14, corresponding to a 10.22% growth increase.

The microbial hubs could, in principle, influence host fitness directly, for example, by contributing to growth, or indirectly through their influence on other beneficial members of the microbial community (26). Here we show that B38 directly improves host growth over early life stages in isolation from the rest of the microbial community. This result is consistent with our field observations, where we found a positive correlation between genetic variation associated with B38 and with seed-set, suggesting that, in this instance, the correlation is causative. The possibility of additional indirect interactions in the field cannot, of course, be excluded.

Mapping Host Genetic Associations with the Relative Abundances of Microbial Hubs across Experiments

Our observation that host control of the relative abundance of four microbial hubs explains ~12% of variation in seed-set among *Arabidopsis* genotypes grown in eight field trials suggests the potential to reveal host genes that can enhance plant performance in the presence of microbes, particularly across environments. Toward this end, we performed genome-wide association mapping for host genotype effects on microbial hubs ($n = 19$) and seed-set across all experiments. Despite significant differences among accessions, genome-wide association (GWA) analysis yielded few peaks with P values below accepted significance thresholds after correction for multiple testing. Specifically, we found only two significant associations, both for microbial hub B41. The first is located on chromosome 1 at position 29909876 in AT1G79510 annotated as a pseudogene. The second is on chromosome 4 on positions 15704377, 15704472, and 15704478. These consecutive single-nucleotide polymorphisms (SNPs) are located between *YUC-1* (AT4G32540), involved in auxin biosynthesis, and *LEUNIG* (AT4G32551), involved in the development of the leaf blade and floral organs.

A potentially more powerful strategy to detect minor quantitative trait loci (QTL) involves computing local association scores along the genome. The assumption underlying this method is that neighboring markers in linkage disequilibrium with causal mutations will also carry association signals; thus, aggregating P values increases power (27). This method identified 344 nonoverlapping loci (hereafter QTLs), with sizes ranging from 93 bp to 150,926 bp, including a total of 25,529 SNPs. Out of the 344 QTLs, only 27 included SNPs associated with multiple traits (Dataset S3).

To investigate functions underlying these associations, we tested pathway and Gene Ontology (GO) term enrichment (biological processes only) (28, 29). Each annotated gene was assigned the highest absolute SNP effect within 5 kb, and we used a combination of methods based on effect sizes accounting for multiple testing, overlapping gene lists, and the potential aggregation of functions and associations along the genome (30–33); we identified 29 enriched GO terms related to biological processes across 16 traits (Datasets S4 and S5), including genes involved in the response to virus (GO:0009615) and nematodes (GO:0009624), hypersensitive response (GO:0009626), and response to chitin (GO:0010200), all of which are related to interactions with other organisms. Three enriched GO terms directly concern auxins and their transport (GO:0009926, GO:0010540, and GO:0009734); auxins have previously been documented to contribute to shaping plant interactions with beneficial bacteria (34, 35). Specialized metabolites also appear to be involved in shaping the relative abundance of microbial hubs. Indeed, hub B107 is associated with genes in the geranylgeranyl diphosphate metabolism (GO:0033385), the universal precursor of diterpenes, which include carotenoids, gibberellins, and hormones such as abscisic acid. In addition, loci associated with B76 are enriched in genes related to specialized metabolite biosynthesis (GO:0044550) and genes involved in the synthesis of sinapoyl glucose and sinapoyl malate (PWY-3301), a side branch in the synthesis of phenylpropanoids. Genes involved in the synthesis of glucosinolates from phenylalanine (like glucotropaeolin in ref. 36, PWY-2821) and hexahomomethionine [specifically, 8-(methylsulfinyl)octyl-glucosinolate (36), PWYQT-4475] are also enriched in loci associated with B5 and F71, respectively.

The functions highlighted by our analysis are in line with other studies suggesting the involvement of specialized metabolites, auxins, and the immune system in influencing the leaf microbial communities (37, 38). Our analysis also highlights less obvious functions, like fatty acid and brassinosteroids biosynthesis (Dataset S5). This is especially true for beneficial members of the community. For example, loci associated with the relative abundance of the beneficial microbial hub B38 are enriched for transition metal ion transport (GO:0000041), response to carbohydrates (GO:0009743), and fatty acid biosynthesis (PWY-4381).

Plant Specialized Metabolites Correlated with Microbial Hub Abundance

Our biological processes and pathway enrichment analysis suggest that specialized metabolites are involved in shaping microbial hubs. To support this result, we quantified 20 compounds using untargeted metabolomics in a subset of the field samples in which we characterized the rosette microbiome. These compounds were chosen to be abundant, allowing annotation, while limiting the number of tests required to explore their association with microbial hubs. We found that the relative abundance of 14 out of 19 hubs was significantly correlated with at least one of 11 specialized metabolites (after correction for multiple testing), 6 of which displayed significant heritability across field sites ranging from 1 to 38% (*SI Appendix*, Fig. S9 *A* and *B*).

The molecule 8-(methylsulfinyl)octyl-glucosinolate (36) (260_GSL_8MSO in *SI Appendix*, Fig. S9 and Table S6) displayed the strongest relationship with multiple microbial hubs in the field (*SI Appendix*, Fig. S9*A* and Table S6), as well as significant heritability under field conditions (*SI Appendix*, Fig. S9*B*). The variation among accessions of this abundant glucosinolate was less evident in the greenhouse and in sterile conditions (*SI Appendix*, Fig. S9*B*), however, leaving open the possibility that the correlation is induced by one or more of the microbial hubs. In contrast, other molecules significantly related to the abundance of microbial hubs in the field across experiments (354_C_Cy-GRGF_785 and 358_F_R-K-R_577; *SI Appendix*, Table S6) are heritable in all conditions, and variation among accessions in the field is positively correlated with the variation among accessions in the greenhouse. This suggests that these flavonoids are constitutively and consistently produced by accessions and influence microbial hubs in a manner that is robust to heterogeneity among field experiments.

Conclusion

In this study, we show that, not only does host genetic variation influence the microbiome, it does so consistently. Host genotype effects are centered on ecologically important hub species, and appear to percolate through the microbial community, at least in part as a result of microbe–microbe interactions. Our replicate field experiments were instrumental in allowing us to reveal consistent host effects on the leaf microbiome via common and widespread hub species.

Furthermore, we found that the influence of host genetics on a handful of prevalent microbial hubs has a far-reaching impact on the community, and is associated with a substantial fraction of the variation in our fitness estimates among accessions. Although these relationships are correlational, a causal relationship is plausible (39), and, indeed, we were able to culture one of the identified hubs and confirm a direct positive effect on host fitness experimentally.

Understanding how host performance or fitness components are influenced by their ability to shape microbial communities could provide a basis for breeding crops favoring microbes that are beneficial to both growth and resistance to pathogens. We successfully mapped variation in host microbe interactions using genome-wide association, and our results suggest that natural and artificial selection can act on plant traits such as leaf specialized metabolites, auxins, and the immune system to improve plant performance through effects on microbial communities (40, 41). In addition, we found that at least some plant metabolites are expressed in a consistent manner that is robust to variation among our experiments and correlates with the relative abundance of microbial hubs. Our results therefore suggest that ongoing efforts to harness host genotype effects on the microbiome for agricultural purposes can be successful, and highlight the value of explicitly considering abiotic variation in those efforts.

Materials and Methods

Field Experiments. This study uses a set of 200 diverse accessions (inbred lines; *SI Appendix*, Table S1) that were previously resequenced (15). The seeds were produced simultaneously in the greenhouse of the University of Chicago under long-day conditions, except for a 12-wk vernalization period at 4 °C, required to induce flowering. The seeds for the common garden experiments were cold stratified in water at 4 °C for 3 d before being planted in trays of 66 open-bottom wells, each measuring 4 cm in diameter. For each experiment, trays were filled with a mix of 90% standard greenhouse soil and 10% local soil. The local soil was collected at the site where each experiment was established, within 2 d of seeds being planted in each year. The standard greenhouse soil was bought in a single order for the four experiments each year. The sites chosen for the experiments were as follows:

SU: Ullstorp (agricultural field, lat: 56.067, long: 13.945)
SR: Ratchkegården (agricultural field, lat: 55.906, long: 14.260)
NM: Ramsta (agricultural field, lat: 62.85, long: 18.193)
NA: Ådal (south-facing slope, lat: 62.862, long 18.331)

The sites were chosen to be *Arabidopsis* habitats and located near known natural populations. Each experiment included three complete randomized blocks, including eight replicates per accession. Experiments were sown in pairs (two in the north and two in the south) over 6 d, corresponding to the sowing of one block a day, alternating between the two experiments (between 7 and 12 August in the north, and between 31 August and 5 September in the south, in both years). The trays were placed in a common garden the morning after sowing under row tunnels to avoid disturbance by precipitation and to favor germination (on the campus of Mid Sweden University in the north and Lund University in the south). Trays were watered as needed, and missing seedlings were transplanted between cells within blocks and then thinned to one per cell after 9 d. Seventeen days after sowing, trays were laid in the field in their final location over tilled soil. For each experiment, the blocks were laid across the most obvious environmental gradient (exposition, shading, slope, soil humidity, …). The pierced bottom of the cells allowed the roots to grow through and reach the soil, as was verified upon harvest. The same protocol was followed in 2011 and 2012.

Sample Collection and DNA Extractions. The rosettes used to characterize the microbial community were harvested in the spring of 2012 and 2013 only a few days after the plants were exposed, following snowmelt. We harvested two randomly selected replicates per accession in each experimental block. Upon harvest, rosettes were placed in sealed paper envelopes, placed on dry ice, and then kept at −80 °C until lyophilized (*S1 Appendix, Supplementary Methods*). DNA extractions were performed on powdered lyophilized rosette tissue. The protocol used included two enzymatic digestions to maximize yield from gram-negative bacteria (42) but otherwise followed (43). Further details about sample processing and DNA extractions are given in *SI Appendix, DNA Extraction*.

PCR and Sequencing. To describe the microbial communities, we amplified and sequenced fragments of the taxonomically informative genes *16S* and *ITS* for bacteria and fungi, respectively. For bacteria, we amplified the hypervariable regions V5, V6, and V7 of the *16S* gene using the primers 799F (5′-AACMGGATTAGATACCCKG-3′) and 1193R (5′-ACGTCATCCCCACCTTCC-3′) (9, 44). For fungi, we amplified the ITS-1 region using the primers ITS1F (5′-CTTGGTCATTTAGAGGAAGTAA-3′) (16) and ITS2 (5′-GCTGCGTTCTTCATCGATGC-3′) (45). The sequencing was performed using 11 MiSeq 500 cycle V2 kits following ref. 46. Primer design (47), PCR conditions (48), and sequencing methods (49, 50) are presented in more detail in *SI Appendix, PCR and Sequencing*.

Sequence Processing and Clustering. The demultiplexed fastq files generated by MiSeq reporter for the first read of each run were quality filtered and truncated to remove potential primer sequences and low-quality base calls using the program cutadapt (51). The reads were then further filtered and converted to fasta files using the FASTX-Toolkit (-q 30 -p 90 -Q33). The fasta files for each run were then dereplicated using AWK code provided in the swarm git repository (https://github.com/torognes/swarm) (17). The resulting dereplicated fasta files were filtered for PCR chimeras using the vsearch uchime_denovo command (https://github.com/torognes/vsearch). The dereplicated fasta files for each run were then combined and further dereplicated at the study level. The fasta files were then used as input for OTU clustering using swarm (-t 4 -c 20000). The clustering identified 150,412 and 251,065 OTUs for the fungal and bacterial communities, respectively. The output files were combined into two separate community matrices using a custom python script (available at GitLab, https://forgemia.inra.fr/bbrachi/microbiota_paper) (52). The taxonomy of each OTU was determined using the quiime2 2019.1 v8 feature classifier trained on the UNITE V8 and SILVA 1.32 database for bacteria and fungi, respectively (53, 54).

Count Table Filtering. The count tables obtained for both the bacterial and fungal communities were filtered in successive steps by removing the following:

1) samples corresponding to empty wells and additional plant genotypes present in the experiments sampled by mistake (leaving 7,476 and 7,240 samples for the fungal and bacterial count tables, respectively)
2) samples with less than 1,000 reads (leaving 6,678 and 6,819 samples for the fungal and bacterial count tables, respectively)
3) OTUs not represented in at least 10 reads in at least five samples (leaving 1,381 and 993 OTUs for the fungal and bacterial count tables, respectively)
4) for the bacterial community, OTUs assigned to plant mitochondria (leaving 993 OTUs in the bacterial count table, no OTUs assigned to plant mitochondria)
5) for a second time, samples with less than 1,000 reads (leaving 6,656 and 6,783 samples for the fungal and bacterial count tables, respectively).

The final count tables used in the study included 993 OTUs and 6,793 samples for the bacterial communities and 1,381 OTUs and 6,656 samples for the fungal community.

The counts for the bacterial community included between 570 and 1,051 samples per experiment. The counts for the fungal community included between 530 and 996 samples per experiment.

Differentiation of the Microbial Communities among Sites and Years. This analysis was performed for the fungal and bacterial communities independently, including all samples and only OTUs with read counts above 0.01% of total read counts (after the filtering described above) across sites and years. To investigate how the microbial communities differed among sites and years, we performed a constrained ordination on log-transformed read counts using the capscale function in the R-package Vegan (55) and following ref. 56. The log transformation offers the advantage of removing large differences in scale among variables. The capscale function performs canonical analysis of PC, an

analysis similar to redundancy analysis (rda), but based on the decomposition of a Bray–Curtis dissimilarity matrix among samples (instead of Euclidean distance in the case of rda). This allows identification of the dimension that maximized the variance explained by components, while discriminating groups of samples, here sites and years, with the formula "$Y \approx$ site + year + site * year" where Y is the count matrix normalized to 1,000 reads and transformed with $\log(x + 1)$ (23, 56).

Core Microbiota. In order to define a core microbiota, we counted, for each OTU, the number of site/year combinations in which it was prevalent. We defined "prevalent" as being present in at least 50% of the samples in a given site/year. We performed this analysis using count tables for each experiment with the filtering described in the previous paragraph. Therefore, for an OTU to be designated as a member of the core microbiota, it needed to have nonzero counts in more than 50% of the samples within each site/year combination and, due to previously described filtering, needed to be represented by at least 10 reads in five of those samples across all site/year combinations (see *Count Table Filtering*).

Heritability of the Microbiota. In this analysis, count tables were split per site and year before filtering for OTUs represented by more than 0.01% of the reads (after the filtering described in *Count Table Filtering*) for each of the bacterial and fungal communities. The resulting 16 counts tables were normalized to 1,000 reads per sample and used to calculate 16 Bray–Curtis pairwise dissimilarity matrices among samples. Count tables were not rarefied. Relative abundances were multiplied by the minimum depth of 1,000 reads. These matrices were then decomposed into 10 PC. For each component, we estimated broad sense heritability (hereafter H^2), that is, the proportion of variance explained by a random intercept effect capturing the identity of the accessions present in the experiment (plate effects had limited impact on H^2 estimates but were included in the models) in models following

$$Y_{ik} \sim \beta_j.\text{Plate}_{ij} + a_k + \varepsilon_{ik}, \quad [1]$$

where Y_{ik} was one of the 10 PC, β is the effect of the plate, Plate is the design matrix capturing the assignation the ith sample to the jth plate, and $a_k \sim \mathcal{N}(0, \sigma_a^2)$ is the random intercept term capturing the effect of the kth accession and $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma_e^2)$ captures the residual variance. Heritability (H^2) was estimated as the percentage of variance explained by the random accession intercept,

$$H^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}. \quad [2]$$

Mixed models were fitted using the function lmer in the lme4 R package (57). We computed 95% confidence intervals (CIs) using 1,000 bootstraps, and components were considered to have significant H^2 when their CIs did not overlap zero (lower bound of the CI $\geq$ 0.01).

Heritability of Individual OTUs. This analysis was also performed per site, year, and community, as in the microbiota H^2 estimation analysis. In this analysis, counts were transformed to centered log-ratios (CLR; after adding one to all counts to handle zeros) using a dedicated function in the R package mixOmics (58, 59). Individual transformed OTU counts were modeled with a model following Eq. **3**,

$$Y_{ik} \sim a_k + \varepsilon_{ik}, \quad [3]$$

where Y_{ik} is the vector of transformed counts for one OTU, and $a_k \approx \mathcal{N}(0, \sigma_a^2)$ is the random intercept term capturing the effect of the kth accession. $\varepsilon_k \sim \mathcal{N}(0, \sigma_e^2)$ captures the residual variance. H^2 estimates and CIs were computed as the proportion variance explained by the accession term $a_k \sim \mathcal{N}(0, \sigma_a^2)$ for each OTU (Eq. **2**). We computed 95% CIs using 1,000 bootstraps, and OTUs were considered to have significant heritability when their CIs did not overlap zero (lower bound of the confidence interval $\geq$ 0.01). H^2 estimates for our estimate of seed-set (see below) were estimated the same way using a Box–Cox transformation.

Microbe–Microbe Cooccurrence Networks. Microbe–microbe cooccurrence networks were computed for the fungal and bacterial communities together, using the count tables per site/year and filtering OTUs represented by less than 0.01% of the reads within each community. The count tables were then combined into the same table and analyzed using the SPIEC-EASI (v1.1) pipeline (23). This method computes sparse microbial ecological networks in a fashion robust to compositional bias and uses conditional independence to identify true ecological interactions, meaning that a connection between two OTUs will be significant when one provides information about the other, given the state of all other OTUs in the network. This means that covariance among OTUs induced by microenvironmental and host genetic variation is controlled. SPIEC-EASI was run using the neighborhood selection framework, and model selection was regularized with parameters set to a minimum lambda ratio of $1e^{-2}$ and a sequence of 50 lambda values (see documentation for SPIEC-EASI and the huge R package, which provides regularization functions) (60).

Network Statistics. The inferences of microbe–microbe ecological interactions inferred using SPIEC-EASI were passed to the igraph package (61), which was used for enforcing simplicity of graphs (no edges that connect vertices to themselves or duplicated edges), computing degree and betweenness centrality of vertices, computing distances between vertices, and plotting. With each of the eight networks thus computed, hubs were defined as OTUs with degree and betweenness centrality both in the 5% tail of their respective distributions. We then checked the overlap between heritable OTUs and hubs, and the overrepresentation of heritable OTUs among hubs was tested using a simple χ^2 test across all site/year combinations. The relationship between distances to heritable hubs (OTUs that are both hubs and have significant H^2) and heritability was investigated using Spearman's rank correlation coefficient. Distances were calculated as the number of edges between OTUs and the closest heritable hub in the network. OTUs not connected to heritable hubs were assigned a distance equal to one more than the maximum distance observed for OTUs connected to heritable hubs.

In order to investigate whether the microbe–microbe associations detected in the networks were mostly due to host genetic effects shared among microbes, we performed permutations of the count tables for each site and year as follow:

1) Compute read counts per sample.
2) Perform a log-ratio transformation of the count table (count + 1).
3) Compute heritability estimates for each OTU (H^2; see *Heritability of Individual OTUs*).
4) For each OTU, and for each *Arabidopsis* genotype, resample the log-ratio transformed counts without replacement across samples. This permutation scheme maintains shared host effects on OTUs but breaks up correlations among OTUs that are independent of the host genotype.
5) Compute new heritability estimates on the permuted data for each OTU (H^2P), which is equal to H^2.
6) Transform the nonpermuted and the permuted log-ratio transform count tables back to proportions using the softmax function (https://rpubs.com/FJRubio/softmax) and then back to counts using the counts per sample computed in step 1 above.
7) Infer interaction networks from both these new count tables using SPIEC-EASI (see *Microbe–Microbe Cooccurrence Networks*).

Estimation of Seed-Set. The experiments each included eight replicates per block per accession (24 replicates per experiment). While we harvested two replicates per block (six replicates per experiment) for microbiota analysis, the remaining plants were left to grow, flower, and produce seeds in the field. We harvested the mature stems of all remaining plants at the end of the spring, when all plants had finished flowering and siliques were mature, and stored them flat in individual paper envelopes. We estimated lifetime seed production (seed-set) by the size of the mature stems. After removing remaining traces of roots and rosettes, each mature plant was photographed on a black background, using a digital single-lens reflex camera (Nikon 60D) mounted on a copy-stand and equipped with a 60-mm macro lens (Nikon 60mm). The photographs were segmented [using custom scripts in R based on the EBimage package (62)] to isolate plants from the image background and estimate the total surface of the image they occupied.

We validated this method with mature plants harvested from a previous experiment that was planted in NM in fall 2010, and that included the 200 accessions used in this study. We counted siliques and estimated the average silique size for 1,607 mature stems that were also photographed. The total silique length produced per plant (number * average size) was highly correlated

with our size estimates based on image analysis (Spearman's rho = 0.84) and displayed a clear linear relationship.

Relationship between Host Effects on Microbial Hubs and Seed-Set. To investigate the relationship between host genotype effects on heritable hubs and seed-set in each experiment, we computed estimates of accession effects (best unbiased linear predictors [BLUPs]) for both log-ratio transformed heritable hubs and Box–Cox transformed seed-set estimates. We then fitted multiple regressions for each site/year combination aiming to explain seed-set variation among accessions with their influence over microbial hubs and following Eq. **4**.

$$f_i \sim \sum_{j=1}^{n} \left[\left(\beta_j . h_{ij} \right) + \left(\gamma_j . h_{ij}^2 \right) \right] + \varepsilon_i, \quad [4]$$

where f_i is the seed-set estimate of the ith accession (BLUP), and h_{ij} is the effect of the ith accession on the jth hub. β_j is the regression coefficient for the jth hub, and γ_j is the regression coefficient for the jth hub squared. $\varepsilon_k \sim \mathcal{N}(0, \sigma_e^2)$ captures residual variance per accession. We then performed forward/backward model selection to obtain the final models presented in *SI Appendix*, Table S4.

Heritable Hubs and Seed-Set across Environments. We next investigated host effects on heritable hubs and seed-set across all eight experiments. Similarly to previous analyses, count tables were split per site and year before filtering for OTUs represented by more than 0.01% of the reads (after the filtering described in *Count Table Filtering*) for each of the bacterial and fungal communities. The resulting 16 count tables were then transformed (CLR) and combined into one before fitting a mixed model following Eq. **5**,

$$Y_{ik} \sim \beta_j . exp_{ij} + a_k + \varepsilon_{ik}, \quad [5]$$

where Y_{ik} is the vector of transformed counts for one OTU, β_j is the effect of the experiment j, exp_{ij} is the design matrix capturing the assignation the ith sample to the jth experiment, $n = 8$), and $a_k \sim \mathcal{N}(0, \sigma_a^2)$ is the random intercept term capturing the effect of the kth accession. $\varepsilon_{ik} \sim \mathcal{N}(0, \sigma_e^2)$.

Seed-set data were analyzed the same way, except we performed Box–Cox transformation of the data. The lambda parameter for the Box–Cox transformation was estimated using the same model, but without the random accession term. Heritability was calculated according to Eq. **2**.

For both heritable microbial hubs and seed-set, we retrieved random intercept accession effects (BLUPs) and fitted a multiple linear regression following Eq. **6**,

$$F_i \sim \sum_{j=1}^{n} \left[\left(\beta_j . H_{ij} \right) + \left(\gamma_j . H_{ij}^2 \right) \right] + \varepsilon_i, \quad [6]$$

where F_i is the effect of the ith accession ($n = 200$) on seed-set (across all experiments), H_{ij} is the effect of accession i on hub j across all experiments, and H_{ij}^2 is the squared effect of accession i on hub j. β_j and γ_j are the corresponding regression coefficient for hub j and $\varepsilon_i \approx \mathcal{N}(0, \sigma_e^2)$ captures the residual variance per accession. The final model was obtained after backward/forward model selection based on AIC.

Isolation, Culture, and Identification of Microbial Hubs.

Bacteria sampling from wild* A. thaliana *plants. We collected two leaves from 10 plants at five locations in Sweden (*SI Appendix*, Table S5) which we stored in 20% glycerol at −20 °C. Wild *A. thaliana* microbial isolates were collected using modified methods that were previously described (25), using six distinct media selected to capture a diverse set of bacterial isolates (63). After isolating and cultivating colonies, we performed DNA extraction and identified over 3,900 isolates using 16S and gyraseB sequencing [*SI Appendix, Bacteria Sampling from Wild A. thaliana Plants* (64)]. Matches to our experimental OTUs are indicated in Dataset S2. Of the isolates identified, we focused on the heritable hub, B38, which appears to contribute to seed-set in the field.

B38 Whole Genome Assembly. We used a low-input method for Illumina library preparation (65). Briefly, ∼2 ng of extracted DNA was used in a reduced volume (5 μL) tagmentation reaction with TDE1 (incubate 55°C for 10 mins, room temperature for 5 mins). The tagmentation reaction was added to a 15-μL PCR, adding the Illumina adapters (Kapa HiFi Hotstart PCR kit KK502, standard Illumina adapters and cycling). The library was cleaned with 0.8× volume SPRI (solid-phase reversible immobilization) beads, quantified on the Bioanalyzer, and run on the MIseq2500 using paired end 300 chemistry. Reads were trimmed for adapters (BBDuk, ktrim = r, k = 23, mink = 11, hdist = 1 tbo) and quality across a sliding window (k = 4, trimq = 20) (66). Reads were assembled using SPAdes (using the settings –isolate -k 21,33,55,77) and annotated with the software Prokka designed for rapid prokaryotic genome annotation (67, 68).

Plant Growth Assays with B38.

Plant growth. *A. thaliana* accession 6136 from Southern Sweden was used in the growth assays. In our field experiments, it displayed average relative counts for B38 (rank 102 of 199). The plant assay used slightly modified methods as previously described (69). The seeds were exposed to chlorine gas for sterilization: In a bell jar with dessicant, an open 1.5-mL tube with seeds was placed next to a 50-mL beaker with 40 mL of Chlorox bleach and 1 mL of hydrochloric acid, sealed with parafilm, and incubated for 4 h. Sterilized seeds were subsequently sown on 24-well tissue plates containing 1.5 mL of 1/2 MS media (Murashige & Skoog medium including Nitsch vitamins, bioWORLD) containing 500 mg/L MES (2-Morpholinoethanesulfonic acid hydrate), pH 5.7 to 5.8. Plates were wrapped in parafilm and vernalized in the dark at 4 °C for 4 d. The plates were individually wrapped with micropore tape to prevent environmental contamination and transferred to a growth chamber with 16 h of light at 16 °C. The plants were treated with either B38 or control inoculum between days 13 and 15 postvernalization. The plates were returned to the chamber to grow for another 14 d.

B38 inoculation. The B38 isolate grew in R2A liquid media in an orbital shaker for approximately 3 days, until the optical density at a wave length of 600 (OD_{600}) reached 0.2. To ensure no environmental contamination, a portion of the inoculum was saved for DNA extraction and subsequent 16S Sanger sequencing verification. The liquid cultures were pelleted by centrifuging at 1,800 relative centrifugal force (RCF) at 18°C for 7 min, decanted, and resuspended in 0.1 M $MgSO_4$. The plants in each 24-well plate were randomly selected to receive the infection (B38 + 0.1 M $MgSO_4$) or control (0.1 M $MgSO_4$) treatment. Each plant was drip inoculated using pipettes with 180 μL of the selected treatment. The plates were rewrapped in micropore tape and returned to the growth chamber.

Measuring plant growth. We performed three trials of 11, 28, and 23 plates, totaling 62 twenty-four-well plates. Plants were not treated and were removed from the experiment if they had less than three true leaves, cracked agar, or failed to germinate, resulting in a total of 1,094 plants. The plants were individually photographed immediately before inoculation, then again at 7 and 14 d postinoculation. The images were processed using a custom script employing cv2 in Python (70), which quantified plant surface area in each well by scaling based on the wells' size, converting images into binary images, and measuring nonwhite pixels within each well (i.e., plant surface area). The output images were manually inspected, and any images which failed to be accurately processed were manually measured using the same pipeline described above, but using Image J.

Due to the high humidity of the plates and the drip inoculation, 422 plants showed signs of waterlog stress. Plants were scored for symptoms of stress induced by waterlogging (blindly with regard to B38 inoculation) as categorized by translucent/white leaves or stunted growth, and were removed from the experiment.

We used a linear mixed model (Eq. **4**) accounting for variation in plant growth among trials and plates within trials to estimate the effect of B38 inoculation.

$$G_{ij} \sim \beta . T_i + p_j + \varepsilon_{ij}, \quad [7]$$

In Eq. **4**, G_{ij} is the growth of ith plant in the jth plate/assay combination. β is the estimate of the treatment effect compared to the controls (intercept), and T_i is the treatment (inoculation with a B38 or control solution); $p_j \sim \mathcal{N}(0, \sigma_p^2)$ is the random intercept effect capturing variation among plates in assays ($n = 62$ plates across three trials). $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ captures the residual variances.

Genome-Wide Association Mapping.

Single polymorphism calling and filtering. SNPs used in this study were generated in the context of the 1001 Genome Project (71) and published in Long et al. (15). As pipelines evolved, we reran SNP calling to ensure optimal quality (*SI Appendix, Single Polymorphism Calling and Filtering*).

Phenotype preparation and association analysis. Association mapping analyses were performed for the 11 heritable microbial hubs for which we estimated host genotype effects across experiment and accession seed-set estimates.

Association analyses were performed using a classical one-trait mixed model accounting for genetic relatedness among accessions (kinship) (71).

In order to take advantage of linkage disequilibrium and gain power by grouping association statistics in contiguous markers, we computed local association scores (27). We followed the instructions provided by the authors and defined the parameter X_i as the 0.999 quantile of the distribution of $-log(p-value)-1$ rounded to the closest integer for each trait investigated (19 microbial hubs and seed-set). The approach highlights regions, which we call QTLs.

The null association model (without fixed SNP effect) from Gemma allows us to estimate SNP-based heritability or pseudoheritability (72), which is the proportion of variance explained by the random accession effect, accounting for the genetic similarity among accessions. To investigate whether the regions highlighted by the local score approach included true positives, we computed SNP-based heritability for each trait, each time using three sets of SNPs to compute the kinship matrix: 1) all the SNPs in the genome over 10% frequency, 2) all the SNPs within QTLs identified by the local score approach, and 3) all SNPs not included in the QTLs identified by the local score approach.

Pathway enrichment analysis. To investigate biological functions associated with seed-set of accessions or their influence over microbial hubs, we searched for enrichment in annotated pathways (in the BIOCYC database) and GO categories (biological processes only) in *A. thaliana*. Gene-set enrichment methods are designed for assays that directly assign *P* values or effects to individual genes (i.e., RNA sequencing experiments). Here, for each trait, each gene was attributed the largest absolute SNP effect within a distance of 5 kb on each side and followed the setRank procedure that accounts for overlapping categories and multiple testing. We set the parameter "setPCutoff" to 0.01 and set the "fdrCutoff" to 0.05 (30). To account for specificities of gene-set enrichment in the context of association mapping, we also tested the enrichment of the gene groups identified by setRank using a weighted Kolmogorov–Smimov score (31) and a permutation scheme accounting for the nonindependence of marker effects due to linkage disequilibrium along the genome, as well as the potential clustering of genes with similar function (32, 33). Briefly, enrichment was calculated using a weighted Kolmogorov sum using gene effect rank (and not a gene effect significance threshold) (31). Enrichments were then tested against an empirical distribution generated from $1e^5$ permutations. For each permutation, chromosomes are randomly reordered and reoriented, and the whole genome is shifted (or "rotated") by a random number, before reassigning SNP effects to genes and calculating enrichment for the groups of genes of interest. We considered only categories with empirical *P* values below 0.05.

Untargeted Metabolomics.

Plant material and sample preparation. This analysis uses three sets of samples. The first are samples collected from the experiments in Sweden and correspond to a subset of those used for the microbial community. In particular, we chose samples from the four experiments established in 2012 and focused on a subset of 50 accessions selected to span the genetic variation among hosts in our mapping population. The second set of samples correspond to six replicates of the same 50 genotypes grown in the University of Chicago greenhouse during the summer 2014 under long-day conditions (16-h light period), in standard culture soil. After 28 d, plants were vernalized for 3 wk at 4 °C, and leaf samples were collected after vernalization, immediately flash frozen in liquid nitrogen, freeze-dried, and stored at room temperature. The third set corresponds to three replicates of the same 50 genotypes, grown on sterile agar medium (Murashige & Skoog with Nitsch vitamins) in individual well plates in a growth chamber with a 16-h light period (long-day condition). Seeds were sterilized by a 70% ethanol bath for 10 min, and manipulated under a sterile hood. Samples were collected after 28 d of growth, flash frozen, freeze-dried, and stored at room temperature.

Dried samples from the three sets were coarsely ground, and distributed in 18 ninety-six-well plates with two ceramic grinding beads per well (10 mg per well ± 2 mg). Samples were randomized across all plates to limit confounding of biological effects. In addition, each plate included 16 random samples (1/6) from each experimental unit (greenhouse, sterile, and the four field experiments).

Specialized Metabolite Extraction and Liquid Chromatography–MS Analysis. The extraction protocol was designed to extract polar compounds such as glucosinolates and flavonoids. Samples in plates were ground using a Geno/Grinder (SPEX SamplePrep 2010) at 1,750 rpm for 2 min. The extraction buffer (70% methanol, 30% water, internal standard: quercetin, 0.0708 mM) was added using a Tecan pipetting robot (100 μL per milligram of dry material). Samples were shaken at room temperature for 2 h and filtered on 96-well filter plates (0.45 μm) on a vacuum manifold. The flow-through was collected in 96-well plates and stored at 4 °C.

Samples were autoinjected through a Zorbax SB-C18 2.1 × 150 mm, 3.5-μm column on an Agilent Q-TOF liquid chromatography–MS with dual electrospray ionisation (ESI, Agilent 6520) with the following parameters: 325 °C gas temperature, 6 $L{\cdot}min^{-1}$ drying gas, 35-eV fixed collision energy, 35 psig nebulizer, 68-V skimmer voltage, 750-V OCT 1 RF Vpp, 170-V fragmentor, and 3,500-V capillary voltage. Mass accuracy was within 2 ppm to 5 ppm. Samples were eluted with 0.1% formic acid in water (A) and 100% acetonitrile (B) using the following separation gradient: 95% A injection followed by a gradient to 90% A at 1 min, 45% A at 6 min, and 100% B at 6.5 min with 4-min hold and 3-min equilibration. An external standard (sinigrin, 1 mM) was run four times before each plate and one time every 20 samples to monitor and maintain run quality. Compounds were characterized using retention times and fragmentation patterns of chromatograms with automatic agile integration in Agilent Mass Hunter Software (Qualitative Analysis B6 2012), and fragments were compared to online databases, massbank (massbank.jp) and plantCyc (plantcyc.org). The XCMS package for peak detection in R (cran.r-project.org) was used to align chromatograms, adjust retention times, and group the peaks. For every molecule, a "barcode" peak was chosen to have a unique retention time and mass to charge ratio (*m/z*) combination. The size of these peaks relative to the internal standard, Quercetin, was used to quantify each molecule in every sample.

Statistical analysis. The peaks' intensities relative to the internal standard were used to capture molecule concentration variation. Standardized intensities were square root transformed before analysis. Heritability of individual compounds in the three conditions was performed using random intercept models identical to those used to estimate OTU heritability. A fixed "site" effect was added for the field samples. In the greenhouse and sterile conditions, a simple random accession term was used to quantify heritability and estimate accession effects (BLUPs). Those accession effects were used to estimate genetic correlation between specialized metabolites in the field and the greenhouse. We used Pearson's correlation coefficient and corrected the corresponding *P* values for false discovery rate (FDR; $n = 20$).

For the field samples, we modeled the relationships between the relative abundances of 19 microbial hubs and the relative intensity of 20 compounds (*SI Appendix*, Table S6) using a linear model following Eq. **8**,

$$H_i \sim \beta 1_s.S_{si} + \beta 2.M_i + \beta 3_s.S_{si}.M_i + \varepsilon_i, \quad \textbf{[8]}$$

where H_i are the log-ratio transformed counts of one of the 19 microbial hubs used for mapping, $\beta 1_s$ are the four site effects, S_{si} is the design matrix assigning sample *i* to site *s*, β2 is the effect of one of the 20 molecules identified in our untargeted screen, and M_i is the relative intensity of the molecules measured in sample *i*. $\beta 3_s$ are site-specific regression coefficients (interactions between the site and molecule effects). $\varepsilon_i \sim \mathcal{N}(0, \sigma_e^2)$ captures the residual variances. We fitted 380 models (19 hubs and 20 molecules) and used *F* tests to estimate term significance. All *P* values corresponding to the molecule effect β2 were corrected for FDR ($n = 380$).

Repeatability of Analysis and Data Availability. All scripts used to perform the analyses presented in this paper, as well as nonessential but complementary figures, are available in the GitLab repository https://forgemia.inra.fr/bbrachi/microbiota_paper (52).

Data tables for OTU counts, seed-set estimates, and plant growth data for the B38 experiment are also available in a Zenodo repository (73).

Metabarcoding Illumina sequences (*ITS* and *16S* amplicons) and the B38 sequence data have been deposited in National Center for Biotechnology Information under BioProject PRJNA707473 (74).

ACKNOWLEDGMENTS. Thanks go to Mia Holm for her hospitality and wonderful dinners after hard work in the field as well as help during harvesting; to Einar Holm for helping with field work and taking photos of harvested plants; to Torbjörn Säll for assistance with sampling and providing greenhouse space in Lund; and, finally, to the Kleen family, the Öhman family, Nils Jönsson, and the Rathckegården farm for allowing us

to install our experiments on their land. Thanks go to Timothée Flutre and Talia Karasov for helpful discussions on previous versions of the manuscript. Thanks go to Man Yu from the C.D. lab, who helped generate stem images used for seed-set estimates and manual seed-set estimates. This work was funded by a grant from the National Health Institute (Grant R01 GM 083068) to J.B., M.N., and C.D.; by a Dropkin Foundation Fellowship to B.B.; and with support from the University of Chicago and New York University (to J.B.). B.B. has received the support of the European Union in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills/AgreenSkills+ fellowship (under Grant Agreement 267196). P.D., M.L.M., and P.L.G. are students in the Magistère de Génétique Graduate Program at Université de Paris. Computing resources and storage were provided by the Center for Research Informatics, funded by the Biological Sciences Division at the University of Chicago with additional funding provided by the Institute for Translational Medicine; CTSA Grant UL1 TR000430 from the NIH; the genotoul bioinformatics platform Toulouse Occitanie, France (Bioinfo Genotoul, http://bioinfo.genotoul.fr); and Bordeaux Bioinformatics Center at the University of Bordeaux, France.

Author affiliations: [a]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; [b]University of Bordeaux, INRAE, BIOGECO, F-33610 Cestas, France; [c]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna BioCenter, 1030 Vienna, Austria; [d]Gene in the Environment, John Innes Center, Norwich, NR47UH, United Kingdom; [e]South China Botanical Garden, Chinese Academy of Sciences, Guangzhou, 510520, China; [f]Department of Natural Sciences, Mid-Sweden University, HLV SE-851 Sundsvall, Sweden; and [g]Center for Genomics and System Biology, Department of Biology, New York University, New York, NY, 10003

Author contributions: B.B., D.F., C.D., M.N., S.H., and J.B. designed research; B.B., D.F., H.W., P.D., P.L.G., M.L.M., T.C.M., E.K., F.R., A.A., M.S.B., S.D., F.H., P.N., T.T., R.W., R.L., M.N., S.H., and J.B. performed research; M.P. contributed methods, feedback on experimental design and new reagents/analytic tools; B.B., H.W., and F.R. analyzed data; M.N. contributed comments on the design of the experiments and the manuscript; and B.B. and J.B. interpreted analyses and wrote the paper.

1. E. J. van Opstal, S. R. Bordenstein, Rethinking heritability of the microbiome. *Science* **349**, 1172–1173 (2015).
2. M. Vétizou *et al.*, Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science* **350**, 1079–1084 (2015).
3. M. A. Abdul-Aziz, A. Cooper, L. S. Weyrich, Exploring relationships between host genome and microbiome: New insights from genome-wide association studies. *Front. Microbiol.* **7**, 1611 (2016).
4. J. K. Goodrich *et al.*, Human genetics shape the gut microbiome. *Cell* **159**, 789–799 (2014).
5. E. G. Pamer, Resurrecting the intestinal microbiota to combat antibiotic-resistant pathogens. *Science* **352**, 535–538 (2016).
6. United Nations Food and Agriculture Organization, "Sustainable agriculture for biodiversity-biodiversity for sustainable agriculture" (Food and Agriculture Organization of the United Nations report I9577EN/1/05.2018, 2018, https://www.fao.org/3/I6602E/i6602e.pdf).
7. R. Santhanam *et al.*, Native root-associated bacteria rescue a plant from a sudden-wilt disease that emerged during continuous cropping. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5013–E5020 (2015).
8. M. R. Wagner *et al.*, Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nat. Commun.* **7**, 12151 (2016).
9. M. W. Horton *et al.*, Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. *Nat. Commun.* **5**, 5320 (2014).
10. J. A. Peiffer *et al.*, Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6548–6553 (2013).
11. D. S. Lundberg *et al.*, Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**, 86–90 (2012).
12. M. T. Agler *et al.*, Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol.* **14**, e1002352 (2016).
13. A. Rochefort *et al.*, Influence of environment and host plant genotype on the structure and diversity of the *Brassica napus* seed microbiota. *Phytobiomes J.* **3**, 326–336 (2019).
14. A. M. Veach *et al.*, Rhizosphere microbiomes diverge among *Populus trichocarpa* plant-host genotypes and chemotypes, but it depends on soil origin. *Microbiome* **7**, 76 (2019).
15. Q. Long *et al.*, Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. *Nat. Genet.* **45**, 884–890 (2013).
16. M. Gardes, T. D. Bruns, *ITS* primers with enhanced specificity for basidiomycetes–Application to the identification of mycorrhizae and rusts. *Mol. Ecol.* **2**, 113–118 (1993).
17. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014).
18. A. Jousset *et al.*, Where less may be more: How the rare biosphere pulls ecosystems strings. *ISME J.* **11**, 853–862 (2017).
19. K. Beilsmith, M. Perisin, J. Bergelson, Natural bacterial assemblages in *Arabidopsis thaliana* tissues become more distinguishable and diverse during host development. *MBio.***12**, 2020.03.04.958165 (2021).
20. D. Bulgarelli *et al.*, Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* **488**, 91–95 (2012).
21. J. Bergelson, J. Mittelstrass, M. W. Horton, Characterizing both bacteria and fungi improves understanding of the *Arabidopsis* root microbiome. *Sci. Rep.* **9**, 24 (2019).
22. S. Deng *et al.*, Genome wide association study reveals plant loci controlling heritability of the rhizosphere microbiome. *ISME J.***15**, 3181–3194 (2021).
23. Z. D. Kurtz *et al.*, Sparse and compositionally robust inference of microbial ecological networks. *PLOS Comput. Biol.* **11**, e1004226 (2015).
24. F. Roux, J. Gasquez, X. Reboud, The dominance of the herbicide resistance cost in several *Arabidopsis thaliana* mutant lines. *Genetics* **166**, 449–460 (2004).
25. Y. Bai *et al.*, Functional overlap of the *Arabidopsis* leaf and root microbiota. *Nature* **528**, 364–369 (2015).
26. K. Farrar, D. Bryant, N. Cope-Selby, Understanding and engineering beneficial plant-microbe interactions: Plant growth promotion in energy crops. *Plant Biotechnol. J.* **12**, 1193–1206 (2014).
27. M. Bonhomme *et al.*, A local score approach improves GWAS resolution and detects minor QTL: Application to *Medicago truncatula* quantitative disease resistance to multiple *Aphanomyces euteiches* isolates. *Heredity* **123**, 517–531 (2019).
28. L. A. Mueller, P. Zhang, S. Y. Rhee, AraCyc: A biochemical pathway database for Arabidopsis. *Plant Physiol.* **132**, 453–460 (2003).
29. P. Schläpfer *et al.*, Genome-wide prediction of metabolic enzymes, pathways, and gene clusters in plants. *Plant Physiol.* **173**, 2041–2059 (2017).
30. C. Simillion, R. Liechti, H. E. L. Lischer, V. Ioannidis, R. Bruggmann, Avoiding the pitfalls of gene set enrichment analysis with SetRank. *BMC Bioinformatics* **18**, 151 (2017).
31. K. Charmpi, B. Ycart, Weighted Kolmogorov Smirnov testing: An alternative for Gene Set Enrichment Analysis. *Stat. Appl. Genet. Mol. Biol.* **14**, 279–293 (2015).
32. B. Brachi *et al.*, Linkage and association mapping of *Arabidopsis thaliana* flowering time in nature. *PLoS Genet.* **6**, e1000940 (2010).
33. S. Atwell *et al.*, Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* **465**, 627–631 (2010).
34. R. Donoso, *et al.*, Biochemical and genetic bases of indole-3-acetic acid (auxin phytohormone) degradation by the plant-growth-promoting rhizobacterium paraburkholderia phytofirmans PsJN. *Appl. Environ. Microbiol.* **83**, e01991-16 (2017).
35. H. Ganin *et al.*, Indole derivatives maintain the status quo between beneficial biofilms and their plant hosts. *Mol. Plant Microbe Interact.* **32**, 1013–1025 (2019).
36. J. W. Fahey, A. T. Zalcmann, P. Talalay, The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry* **56**, 5–51 (2001).
37. A. C. Huang *et al.*, A specialized metabolic network selectively modulates *Arabidopsis* root microbiota. *Science* **364**, eaau6389 (2019).
38. G. Castrillo *et al.*, Root microbiota drive direct integration of phosphate stress and immunity. *Nature* **543**, 513–518 (2017).
39. E. French, I. Kaplan, A. Iyer-Pascuzzi, C. H. Nakatsu, L. Enders, Emerging strategies for precision microbiome management in diverse agroecosystems. *Nat. Plants* **7**, 256–267 (2021).
40. O. M. Finkel, G. Castrillo, S. Herrera Paredes, I. Salas González, J. L. Dangl, Understanding and exploiting plant beneficial microbes. *Curr. Opin. Plant Biol.* **38**, 155–163 (2017).
41. K. R. Foster, J. Schluter, K. Z. Coyte, S. Rakoff-Nahoum, The evolution of the host microbiome as an ecosystem on a leash. *Nature* **548**, 43–51 (2017).
42. J. L. Morgan, A. E. Darling, J. A. Eisen, Metagenomic sequencing of an in vitro-simulated microbial community. *PLoS One* **5**, e10209 (2010).
43. J. Amani, R. Kazemi, A. R. Abbasi, A. H. Salmanian, A simple and rapid leaf genomic DNA extraction method for polymerase chain reaction analysis. *Iran. J. Biotechnol.* **9**, 69–71 (2011).
44. M. K. Chelius, E. W. Triplett, The diversity of archaea and bacteria in association with the roots of *Zea mays L. Microb. Ecol.* **41**, 252–263 (2001).
45. T. J. White, S. Bruns, S. Lee, J. Taylor, "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics" in *PCR Protocols: A Guide to Methods and Applications*, M. A. Innis, D. H. Gelfand, J. J. Sninsky, T. J. White, Eds. (Academic, 1990), pp. 315–322.
46. J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, P. D. Schloss, Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112–5120 (2013).
47. W. A. Walters *et al.*, PrimerProspector: De novo design and taxonomic analysis of barcoded polymerase chain reaction primers. *Bioinformatics* **27**, 1159–1161 (2011).
48. T. Samarakoon, S. Y. Wang, M. H. Alford, Enhancing PCR amplification of DNA from recalcitrant plant specimens using a trehalose-based additive. *Appl. Plant Sci.* **1**, 1200236 (2013).
49. N. Rohland, D. Reich, Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
50. J. G. Caporaso *et al.*, Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621–1624 (2012).
51. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10 (2011).
52. B. Brachi, microbiota_paper, data for "Plant genetic effects on microbial hubs impact fitness across field trials." GitLab. https://forgemia.inra.fr/bbrachi/microbiota_paper. Deposited 10 August 2020.
53. U. Kõljalg *et al.*, Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* **22**, 5271–5277 (2013).
54. C. Quast *et al.*, The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
55. P. Dixon, VEGAN, a package of R functions for community ecology. *J. Veg. Sci.* **14**, 927–930 (2003).
56. M. J. Anderson, T. J. Willis, Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology* **84**, 511–525 (2003).
57. D. Bates, M. Maechler, B. Bolker, S. Walker, Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* (2015).
58. J. Aitchison, The Statistical analysis of compositional data. *J. R. Stat. Soc. B* **44**, 365–374 (1982).
59. K.-A. K.-A. Lê Cao, I. González, S. Déjean, I. González, *Unravelling "omics" Data with the R Package mixOmics* (HAL, 2012).
60. T. Zhao, H. Liu, K. Roeder, J. Lafferty, L. Wasserman, The huge package for high-dimensional undirected graph estimation in R. *J. Mach. Learn. Res.* **13**, 1059–1062 (2012).
61. G. Csárdi, T. Nepusz, The igraph software package for complex network research. *InterJournal Complex Syst.* **1695**, 1–9 (2006).
62. G. Pau, F. Fuchs, O. Sklyar, M. Boutros, W. Huber, EBImage–An R package for image processing with applications to cellular phenotypes. *Bioinformatics* **26**, 979–981 (2010).
63. A. E. McCaig, S. J. Grayston, J. I. Prosser, L. A. Glover, Impact of cultivation on characterisation of species composition of soil bacterial communities. *FEMS Microbiol. Ecol.* **35**, 37–48 (2001).
64. C. Bartoli *et al.*, In situ relationships between microbiota and potential pathobiota in *Arabidopsis thaliana*. *ISME J.* **12**, 2024–2038 (2018).

65. M. Baym *et al.*, Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**, e0128036 (2015).
66. B. Bushnell, BBDuk. Jt Genome Inst. https://jgi. doe. gov/data-and-tools/bbtools/bb-tools-userguide/bbduk-guide/. Accessed 25 August 2020.
67. T. Seemann, Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
68. A. Bankevich *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
69. T. L. Karasov *et al.*, *Arabidopsis thaliana* and *Pseudomonas* pathogens exhibit stable associations over evolutionary timescales. *Cell Host Microbe* **24**, 168–179.e4 (2018).
70. G. Bradski, The OpenCV Library. *Dr. Dobbs J. Softw. Tools* **120**, 122–125 (2000).
71. X. Zhou, M. Stephens, Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
72. J. Yang *et al.*, Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
73. B. Brachi, H. Whitehurst, Data for "Plant genetic effects on microbial hubs impact host fitness in repeated field trials." Zenodo. https:/doi.org/10.5281/zenodo.6783090. Deposited 30 June 2022.
74. B. Brachi, H. Whitehurst, Microbial sequence data for "Plant genetic effects on microbial hubs impact host fitness in repeated field trials." NCBI BioProject. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA707473. Deposited 30 June 2022.

CHAPTER 3

ASSESSMENT OF GYRASE-B CLASSIFICATION IN LEAF MICROBIOME COMMUNITIES

3.1 Abstract

Characterizing the abundance and taxonomy of microbial community members informs phenotypic studies assessing microbes' effects on plant host health. One method for high-throughput, affordable characterization employs amplicon sequencing of marker genes, such as the 16S ribosomal subunit, which can then be taxonomically classified using reference databases consisting of amplicon sequences and respective taxonomic lineages. However, the 16S gene can occur in multiple copies within a genome and is prone to both recombination and horizontal transfer, resulting in low levels of concordance. Thus, using 16S amplicons to characterize microbiome taxonomies confounds microbial diversity analysis and introduces technical artifacts in phylogenetic studies. We use one plant microbiome data set in combination with a new isolate collection from wild *Arabidopsis thaliana* plants to assess the applicability of the single-copy gyrase subunit β (*gyrB*) marker gene to characterize naturally-occurring endophytic communities. We demonstrate that *gyrB* allows for the identification of more, distinct strains compared to 16S and demonstrates tighter correlation between genetic and genomic distances. Using our new data from isolates collected from wild *A. thaliana* plants, we provide a curated *gyrB* database that we demonstrate improves taxonomic resolution of one published data set. Continued collaboration to expand the *gyrB* database for use in the field will allow for increased precision in future microbiome studies.

3.2 Introduction

Plants host hundreds to thousands of distinct bacteria that affect host health, from increasing plant tissue growth rates to incurring localized cell death (Compant et al. 2019). Investigating how taxonomic distributions among microbiomes correlate with host phenotype will allow for the identification of tractable components of the microbiome community. The microbiome may be taxonomically characterized through a) the culturing and sequencing of bacteria from the plant tissue, b) metagenomic sequencing of the entire microbiome, or c) amplicon sequencing of essential bacterial genes. In practicality, isolating microbes from plant tissue remains technically challenging due to the varying media substrates, temperatures, and environments required to propagate the growth of unknown natural isolates, while metacommunity sequencing proves impractical due to the high proportion of host DNA "contaminating" the microbiome samples (although see examples for microbial DNA enrichment applied pre and post DNA extraction in Ikeda et al. (2009) and Heravi et al. (2020), respectively). However, using universal DNA primers to amplify, sequence, then taxonomically identify bacteria, allowing researchers to characterize microbiomes with efficiency.

The 16S ribosomal RNA gene, an essential microbial gene required for protein synthesis, is one of the first marker genes used for broad microbial phylogenetic classification (Pace, Sapp, and Goldenfeld 2012; Lane et al. 1985). In the 1970s, Carl Woese, George Fox and others rigorously sequenced the 16S ribosomal gene across thirteen microbial taxa, laying the foundation for taxonomic identification of microbes through a single gene (Woese and Fox 1977; Glöckner et al. 2017). With the advent of high-throughput sequencing, researchers continued to build and use 16S databases to characterize complex microbial communities (Quast et al. 2013). However, the tightly constrained gene evolves slowly and thus the DNA sequences often cannot differentiate closely related species. Recent phylogenetic studies also demonstrate that the 16S gene is susceptible to both horizontal gene transfer and recombination, resulting in low levels of concordance (Hassler et al. 2022). Additionally, its varying

copy levels may introduce bias into taxonomic community analysis (e.g. Větrovský and Baldrian 2013), although computational tools can help control for copy numbers (Perisin et al. 2016). Using a different, single-copy gene with higher divergence than 16S would theoretically increase statistical power when performing analysis reliant on taxonomic grouping including alpha/beta-diversity comparisons.

Yamamoto and Harayama (1995) expounded the application of another marker gene, DNA gyrase subunit-β. The high rates of evolution observed in *gyrB* distinguishes closely related species. Barret et al. (2015) pioneered its application in complex, seed microbial community analysis through the development of degenerate primers. They showed that *gyrB* discriminates amplicon sequence variants (ASVs) to the species level, with a genetic distance of 0.02 among amplicons identifying distinct species with high precision (F1=.959). Less than 10% of ASVs were unclassified at the family taxonomic level. However, subsequent research using *gyrB* in various study systems yielded varying levels of precision. One study found that 16S was more accurate than *gyrB* for community profiling microbial communities in neonatal care facilities (Martineau et al. 2018). Another study reviewing food microbiomes found that *gyrB* was substantially advantageous in tracking subspecies-level changes and measuring β-diversity (Poirier et al. 2018). Both Poirier et al. (2018) and Martineau et al. (2018) observe comparable performance between 16S and *gyrB* in most samples when considering phylum-level community characterization. Thus, the increased taxonomic resolution of the *gyrB* gene, compared to 16S, makes it a promising candidate for future microbiome composition studies but requires validation in each study system.

Few studies apply *gyrB* amplicon profiling to the model organism *A. thaliana's* microbial communities. However, in one comprehensive study, Bartoli et al. (2018) used *gyrB* amplicon sequence to examine the taxonomic diversity of *A. thaliana* leaf and root communities harboring putative pathogens. They observed several interesting ecological interactions, including notable shifts in diversity that were dependent on the presence of the pathobiota,

season, and tissue type. Still, 20-40% of all included operational taxonomic units (OTUs) for a given data point were unclassified at the family taxonomic level. Improving the taxonomic resolution in microbiome studies would increase statistical power and potentially enhance the ability to observe ecological interactions among taxonomic groups. The varying levels of success in *gyrB* applications are, in part, shaped by the *gyrB* reference databases used. In this context, databases house the amplicon sequences and associated lineages which inform classifiers during amplicon data processing. These classifiers assign taxonomies to all amplicon sequences observed. At the time of this writing, there are no known publicly-available, curated *gyrB* databases formatted for standard amplicon processing pipelines which include the co-amplified, paralogous *parE* gene sequences (although see Ramírez-Sánchez et al. 2022 for an example of a published database).

Here, we investigate the applicability of the *gyrB* amplicon in *A. thaliana* leaf endophytes by 1) comparing the efficiency of a new *gyrB* database in taxonomic classification compared to two other *gyrB* databases using a published data set, Bartoli et al. (2018), 2) comparing the taxonomic resolution afforded by 16S(v5-v7) compared to *gyrB* using bacterial isolates collected from fifty, wild *A. thaliana* plants; and 3) estimating the prevalence and abundance of *gyrB* paralogs in the collected isolates. Taken together, we demonstrate that *gyrB* is effective in identifying microbes in this system, but special care should be taken in filtering out paralogs and applying analysis dependent on grouping taxonomies below the phyla level. We provide guidance on reconstructing taxonomic amplicon databases from public resources and applying them to novel systems. Our database, raw reads, and scripts are available online at https://github.com/hlwhitehurst/gyrB-database-analysis.

3.3 Results

3.3.1 Development of a new, larger gyrB *database improves taxonomic identification*

We compiled a database of 53,239 trimmed and filtered *gyrB* sequences and respective lineages using data publicly available through the Joint Genomes Institute (JGI) and National Center for Biotechnology Information (NCBI) (See Materials and Methods). Dr. Matthieu Barret graciously provided a working version of their *gyrB* reference database, which we used for benchmark comparisons. Importantly, their database included complementary records to our own, including 4,373 records of the *parE* gene, a paralog of *gyrB*, and records for twenty families that were underrepresented in our database (∼2,500 records). We extracted and incorporated these records into our own for the subsequent analysis, resulting in 55,218 database records.

To test the effectiveness of our methods in building the *gyrB* database, we used the Bartoli et al. (2018) OTU table of microbes associated with *A. thaliana*. The data set includes 7,549 distinct *gyrB* OTU sequences from 1,901 *A. thaliana* biological samples. We used our *gyrB* database (WH) in comparison to the results obtained through the Barret database (BM) and the published taxonomy results by Bartoli et al. (2018) (BC). The WH and BM databases included *parE* sequences, and the WH database included the largest number of records (Table 3.1). For the WH and BM analyses, we used the respective databases with DADA2 Callahan et al. (2016), and the integrated native Bayesian classifier, to assign taxonomic lineages to the OTU sequences. Our results demonstrate that the WH database results in lower levels of "unclassified" OTUs at every taxonomic level and with comparable confidence levels to the BC and BM database results (Figure 3.3). The WH and BM databases report similar abundances for the top twenty families, indicating that the WH database mostly improves on the low-abundance or rare taxa. Notably, the addition of the paralog sequences identified

by Barret proves particularly beneficial as 4-7% of all distinct OTU sequences in the Bartoli et al. (2018) dataset were identified as paralogs (accounting for approximately 2% of all OTU counts).

Database	n. ref sequences	*parE* included?	reference
BC	30,525	No	Bartoli et al (2018)
BM	38,929	Yes	correspondence, Barret et al (2015)
WH	53,239	Yes	this paper

Table 3.1: List of *gyrB* databases used in comparing taxonomic classifications of OTU sequences provided in the Bartoli et al. (2018) data

3.3.2 parE *co-amplifies with* gyrB *and is unequally represented among and within taxonomic groups*

If the *gyrB* paralog, *parE*, is equally represented across the *A. thaliana* bacteria, the *parE* counts could theoretically be included in community profiles. However, if the paralog is not equally represented amongst bacteria, we hypothesized that *parE* would introduce bias to specific taxonomic groups and therefore should be removed before community analysis. Here we performed whole genome sequencing on 105 bacteria isolates collected from wild *A. thaliana* plants in Sweden to examine the frequency and ratios of *parE* genes in our study system.

We first reviewed the putative *parE* gene sequences which might co-amplify with *gyrB* using the degenerate primers designed by Barret et al. (2015). We mapped the *gyrB* primers to the whole genome assembly and extracted the respective amplicon products using Geneious. Of the 105 genome assemblies that we analyzed, 89 contained putative *gyrB* amplicons expected with the Barret et al. (2015) primers (one genome reporting two copies). Nearly all of those isolates (88) had one distinct *parE* sequence (although we observed one isolate with two putative copies).

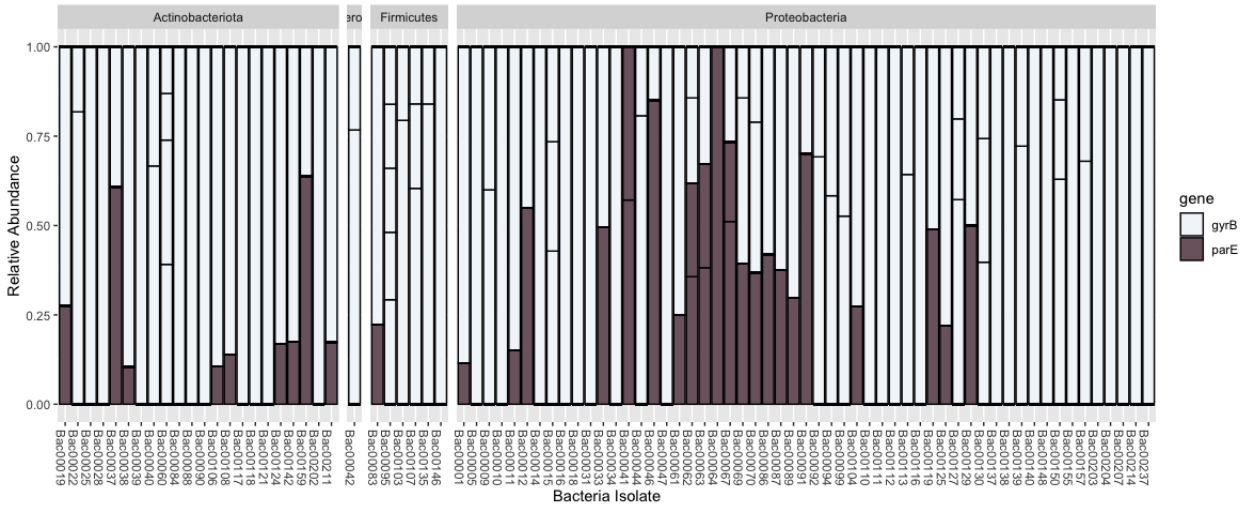

Figure 3.1: *parE* coamplifies with *gyrB* and varies among families. We observed differing efficiencies of *parE* amplicons co-amplifying with *gyrB* primers (Barret et al. 2015) in 105 bacterial strains collected from wild *A. thaliana* leaf communities.

We then performed *gyrB* amplicon sequencing on the microbial isolates' DNA to determine the empirical ratios and occurrences of *parE* co-amplification using the Barret et al. (2015) primers. We observed *parE* amplicons in 33 of the expected 88 genomes (Figure 3.1). The *parE* amplicons are not restrained to specific phylum as observed in food microbial communities (Poirier et al. 2018), and the portion of amplicon sequences assigned to *parE* varied across bacteria strains. Our results thus indicate that the putative off-target amplicons cannot be reliably filtered using *a priori* expectations extracted from short-read whole genome data or specific taxonomic groups in *A. thaliana* leaf communities.

3.3.3 gyrB *sequence distances better correlate to genomic distance compared to 16S*

Some common analyses of microbial community composition incorporate taxonomic grouping or phylogenetic distances, such as Bray-Curtis dissimilarity and Unifrac, respectively. However, the phylogenetic signal of 16S is weak compared to core genome phylogenies and other single-copy gene phylogenies (Hassler et al. 2022). To supplement our isolate sequences, we

included publicly available genomes of isolates collected from *A. thaliana*. We then mapped the 16S and *gyrB* primers to all genomes and extracted the flanked amplicon sequence. We removed any genomes that were low quality (redundancy >10% or completeness <90%, determined by anvi'o (Eren et al. 2015)) or for which we did not successfully extract either a 16S or *gyrB* sequence, resulting in 303 genomes for analysis. We generated whole-genome phylogenies using all of the single-copy genes (n=71) which were present in every genome (see scripts for all genes used). For genetic distances, we aligned the sequences for the 16S and *gyrB* amplicons and quantified Jukes-Cantor distances as compared to the whole genome phylogenetic distances. Pairwise comparison of amplicon sequence distances show amplicon genetic distances positively correlated with genomic distances, but *gyrB* (R=0.89, p<0.001) was more strongly correlated compared 16S (R=0.83, p<0.001)(Figure 3.2). When we mapped the genetic distances to the phylogenetic distances, we visualize some strains were more similar to distant clades than closer ones, supporting the trend described in Hassler et al. (2022).

3.3.4 gyrB *primers better identify prevalent* Plantibacter *bacteria compared to 16S*

In addition to comparing the 16S and *gyrB* genetic distances compared to genomic distance, we also compared the phylogenetic classifications of the isolates using 16S compared to *gyrB*. Surprisingly, the *gyrB* amplicon classified more isolates as Microbacteriaceae than 16S (172 and 566 isolates, respectively). Of the 566 isolates identified as Microbacteriaceae by the *gyrB* amplicon, 394 isolates did not report any 16S sequences in the processed data set. We examined the whole genomes of five Microbacteriacea for further insights. Interestingly, the Microbacteriaceae in the genus *Plantibacter* all showed one mismatch in both the forward and reverse primer regions. Thus, the 16S primers seem to be less efficient at amplifying *Plantibacter* isolates compared to the *gyrB* primers.

3.4 Discussion

We demonstrate similar or improved taxonomic identification confidence values and higher percentage of sequences identified across all taxonomic hierarchies compared to other unpublished datasets. One distinct difference between our database and the BC and BM databases is in the number of reference sequences used. The WH database generated for this paper benefits from several years of new data compared to the other databases: our database included approximately 1.3-1.7 times as many reference sequences (Table 3.1). We anticipate that the ongoing sequencing data contributions to public data sets will continue to expand and improve the building of *gyrB* reference databases. We provide scripts and the nucleotide database as guidance to develop specialized tools and to offer an initial benchmark for future *gyrB* database development. We also provide insights into the inconsistent prevalence of the paralogs in microbial leaf isolates, illustrating the importance of including *parE* references sequences in community composition studies using *gyrB*.

In addition to improved taxonomical classification, we demonstrate that *gyrB* genetic distance correlates better with genomic distance when compared to 16S. However, the correlation is not completely linear. The value of *gyrB* is most assuredly the ability to distinguish closely related bacterial strains that likely have identical 16S sequences. The incorporation of taxonomic hierarchies and taxonomic groupings in microbiome community analysis should be exercised with caution given the imperfect correlation between genetic and genomic distances. One surprising result of this study is that the 16S primers used in this study are inefficient in amplifying the 16S marker gene region of a prevalent *Plantibacter* bacteria. The 16S primers 799F and 1492R were designed to amplify v5-v7 of the 16S gene, reducing off-target amplification of *A. thaliana* mitochondrial DNA compared to primers for other variable regions of the marker gene (Chelius and Triplett 2001; Horton et al. 2014). The primers under amplify the *Plantibacter* genus, mostly likely due to two mismatches in the primer-binding regions. Our findings illustrate the potential for biases introduced by con-

strained primers, especially in microbial communities studying natural, unknown isolates. Verifying the efficiency of primers in each study system through multiple marker genes and genomic data proves an essential step in future microbiome studies.

3.5 Materials and Methods

All data was analyzed using R version 4.1.1 (R Core Team 2021) unless otherwise noted.

3.5.1 Generating WH gyrB *database*

JGI export To build our local database, we extracted all readily-available *gyrB* sequences from the Joint Genomes Institute (JGI) using TIGRFAM 01059 for both "finished" and "permanent drafts" of all available genomes. JGI imposes limitations on the number of sequences permitted per download, and so downloads were performed in batches. See scripts for additional details. We concurrently extracted lineage information by adding genes to the scaffold cart and selecting the lineage option. We cross-tabulated gene records with lineages using custom code.

NCBI export We exported *gyrB* sequences from the National Center for Biotechnology Information (NCBI) using the nuccore search term "gyrB". We cross tabulated taxonomic IDs with sequences by downloading the taxonomy reference files from NCBI and using custom scripts incorporating the TaxonKit (Shen and Ren 2021) and seqkit (Shen et al. 2016) commandline tools.

Combining data and formatting To improve specificity of our database, we then mapped the *gyrB* primers to each reference sequence and trimmed the flanking regions, removing any reads that did not have matches or were < 210 bases long. We removed all records with no lineage data. We formatted some taxonomic groups that had universally ambigu-

ous classifications. M. Barret provided their working *gyrB* database version (v5, Sept 15 2022, personal correspondence). After an initial comparison, we identified five families in which the BM database outperformed the WH database. We extracted and incorporated sequences from those families (∼2,500 records total) into the WH database BM along with the 4,373 records of the *parE* gene. All database compilation scripts are published online (https://github.com/hlwhitehurst/gyrB-database-analysis).

3.5.2 Comparing database classifications

We accessed the Bartoli et al. (2018) OTU table through personal correspondence with Dr. Claudia Bartoli, which included taxonomic assignments and bootstrap confidences for each distinct sequence. For comparative analysis, we used the DADA2 naive Bayesian classifier (Callahan et al. 2016; Wang et al. 2007) using standard parameters while extracting bootstrap values (minboot=50, outputBootstraps = TRUE).

3.5.3 Bacteria isolate collection and DNA extraction

Plant tissue collection We collected two leaves from ten plants at five locations across Sweden, totaling one-hundred leaf samples. Leaves were surface sterilized in the field. Rosette leaves were individually cut from the plant with sterile scissors and rinsed with double distilled water. The leaf was rinsed in 70% EtOH for 3-5 seconds, then placed in a fresh, sterile tube on dry ice. We then added sterile 300 ul 20% glycerol to each tube and hand macerated. Samples were stored at -80°C until plating.

Microbial isolate collection and propagation For isolate collection, we prepped six distinct media types, modified from Bai et al. (2015), which were intentionally selected for diverse substrates to grow a variety of bacteria. Media types included R2A, Minimal media containing Methanol, Tryptic Soy Agar, Tryptone Yeast extract Glucose Agar, Yeast Extract

Manitol Agar, and; 0.1 Tryptic Soy Agar(Appendix A, Table 6.1. In a sterile biological safety hood, 45 ul of the macerated glycerol substrate was plated onto each media type using wide-mouth filtered pipette tips and subsequently spread. Plates were allowed to dry for approximately 10 minutes before sealing with parafilm and placing in the incubator at 28°C. Plates were checked every 2-3 days for new colony growth for up to three weeks. New colonies inoculated into 300ul Nutrient Broth, loosely sealed, and placed on a 280 RPM (28°C) shaker until turbid, approximately 1-4 days.

DNA Extraction Liquid cultures were spun down 6600 RCF for 10 mins and decanted. The following DNA extraction was performed using custom scripts on the Tecan Freedom Evo liquid handler, and we performed a double enzymatic digest on all samples to increase yields from both gram-positive and gram-negative microbes. First, samples were incubated for 30 mins with 350 U Ready-Lyse Lysozyme and 245 U RNase A (QIAGEN, Germantown, MD) in 250ul TES (10 mM Tris-HCl pH 8, 1 mM EDTA, 100 mM NaCl). We added 2 mg/mL Proteinase K in 250ul TES + 2% SDS and incubated for 4-8 hours at 55°C. The SDS-protein complex was precipitated using 0.3 volume 5M NaCl, then briefly centrifuged to pellet. The clear supernatant was pipetted into a clean, sterile plate. The eluted DNA was further purified using in-house Solid Phase Reversible Immobilization (SPRI) beads (Rohland and Reich 2012), which contain 0.1% SpeedBead Carboxylate-Modified Magnetic Particles (Hydrophobic) (e.g. Cytiva, prod.# : 65152105050250, rinsed in TE buffer), 18% PEG-8000 (w/v), 1M NaCl, 10mM Tris-HCl, ph 8.0, 1 mM EDTA, pH 8.0). SPRI clean up was performed as described in (Rohland and Reich 2012), with the following modifications: beads were added at 0.5x sample DNA elution to bead volume, performing two ethanol rinses using 80% EtOH, and eluting the final genomic material in molecular-grade, sterile water.

3.5.4 Whole genome shotgun library preparation, sequencing, and processing

We performed whole genome sequencing on 237 of the nearly 6000 bacteria isolates. We selected the isolates based on unique 16S and/or *gyrB* amplicons with multiple isolates from highly abundant families within the sample set.

Library Preparation and Sequencing Standard Whole-genome library preps were generated and cleaned on the subset of the bacteria isolates' DNA using reduced volume reactions of 5 ul for tagmentation which were directly used in 12.5ul PCR reactions for adapter ligation (Lamble et al. 2013) and subsequently size selected and cleaned using the SPRI beads described above using 0.8:1 beads to sample volumes. Libraries were combined into three distinct pools and sequenced on the Illumina MiSeq (PE300 v3 chemistry) or the NovaSeq (PE100, SP flowcell). Genome assembly and quality analysis Reads were trimmed for adapters using BBtools (Bushnell, Rood, and Singer 2017) using standard parameters. Genomes were assembled using SPAdes (standard parameters, including run options "-k 21,33,55,77 –isolate") (Bankevich et al. 2012). The assembled genomes were annotated using standard anvi'o (Eren et al. 2015) pipelines and selecting the NCBI COGS option. See scripts for all anvi'o processing steps. Of the 237 genomes sequenced, 105 genomes met quality and completeness thresholds for subsequent analysis for this study (>100,000 bp n50, <10% redundancy, >90% complete as quantified by QUAST (Gurevich et al. 2013) and anvi'o (Eren et al. 2015)).

3.5.5 Amplicon sequencing and analysis

Amplicon sequencing constructs were designed using the modular oligos structure described in (Illumina 2013) which uses a two-step PCR protocol (See Appendix C for table of oligos). We co-amplified 16S and *gyrB* marker genes.

PCR1

The initial PCR used marker-gene primers with internal indexes [Bartoli 2018] and an Illumina adapter overhang on the 5' ends, resulting in the construct: 5'-[adapter overhang]-[internal index]-[16S or *gyrB* primer]-3'. PCRs were performed using the Kapa HiFi Kit (KK 2502) in 15 ul reactions comprised of 4.1 ul molecular grade water, 3.0 ul KAPA HiFi Fidelity Buffer (5X), 0.5 ul dNTP mix, 0.4 ul KAPA HiFi HotStart DNA Polymerase, 1.5 ul [1.6 mM] 16S F primer + 1.5 ul [6.6 mM] *gyrB* forward primers and reverse primers each. PCR thermocycler program consisted of an initial denaturing step for 5 min (98°C), followed by 29 cycles of a 30 second denature (98°C), 30 second annealing (61°C), 45 second extension (72°C), then a final extension phase for 5 min (72°C) before holding at 4°C. Due to high amounts of primer-dimers formed by the high concentration of degenerate *gyrB* primers, samples were size-selected using SPRI beads as described above, modified to use a bead to sample ratio 0.8:1.

PCR2

and Sequencing PCR primer constructs used standard Illumina indexes and adapter sequences as described in (Illumina 2013). PCR reactions comprised of: 3.1 ul molecular grade water, 3.0 KAPA HiFi Fidelity Buffer (5X), 0.5 ul dNTP mix, 0.4 ul KAPA HiFi HotStart DNA Polymerase, 1 ul [5uM] forward primer + 1 ul [5uM] reverse primers, and 5 ul of PCR1 product. PCR thermocycler program consisted of an initial denaturing step for 5 min (98°C), followed by 12 cycles of a 30 second denature (98°C), 30 second annealing (55°C), 40 second extension (72°C), then a final extension phase for 5 min (72°C) before holding at 4°C. Samples were pooled by plate and size selected using SPRI beads as described above, using 0.8:1 bead to sample ratio and re-eluted into half the starting volume using molecular grade water. Samples were size selected for 300- to 700-bp products using the 1.5% Agrose BluePippen kit (Sage Science, Beverly, MA, USA). Samples were sequenced across

two sequencing runs on the Illumina MiSeq (V3 PE300 chemistry).

3.5.6 Phylogenetic tree construction for amplicon sequences

We downloaded all bacteria genomes generated from *A. thaliana* isolates publicly available through JGI IMG database (Price, Dehal, and Arkin 2010) (Taxonomy Domain = "Bacteria" AND Environmental Classification Host Name = "Arabidopsis" OR Environmental Classification Isolate = "Arabidopsis"). The JGI IMG genomes (n=398) were combined with the 237 genomes generated in this research, and the 16S and *gyrB* primer sequences (Appendix C: Table 8.1) were mapped to each genome and the amplicon sequences were extracted using BBTools (Bushnell, Rood, and Singer (2017), see scripts). Genomes were removed from analysis if they reported lower quality as described previously (completeness <90% or redundancy <10%, quantified by anvi'o) (Eren et al. 2015), or if either the 16S or *gyrB* sequences were not successfully extracted through primer mapping, resulting in 303 genomes. Whole genomes phylogenies were built off of all shared single copy genes (n=71) using anvi'o ("anvi-gen-phylogenomic-tree" option) (Eren et al. 2015), which employs FastTree (Price, Dehal, and Arkin 2010). The tree was rooted using ape version 5.7 (Paradis, Claude, and Strimmer 2004).

For the genetic distances, the amplicon sequences were aligned using ClustalW and extracted the distances using the R packages msa (Bodenhofer et al. 2015) and phanghorn (Schliep 2011) using the default Jukes-Cantor option.

3.5.7 Data Availability

A portion of these data were produced by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337; operated under Contract No. DE-AC02-05CH11231) in collaboration with the user community (Price, Dehal, and Arkin 2010).

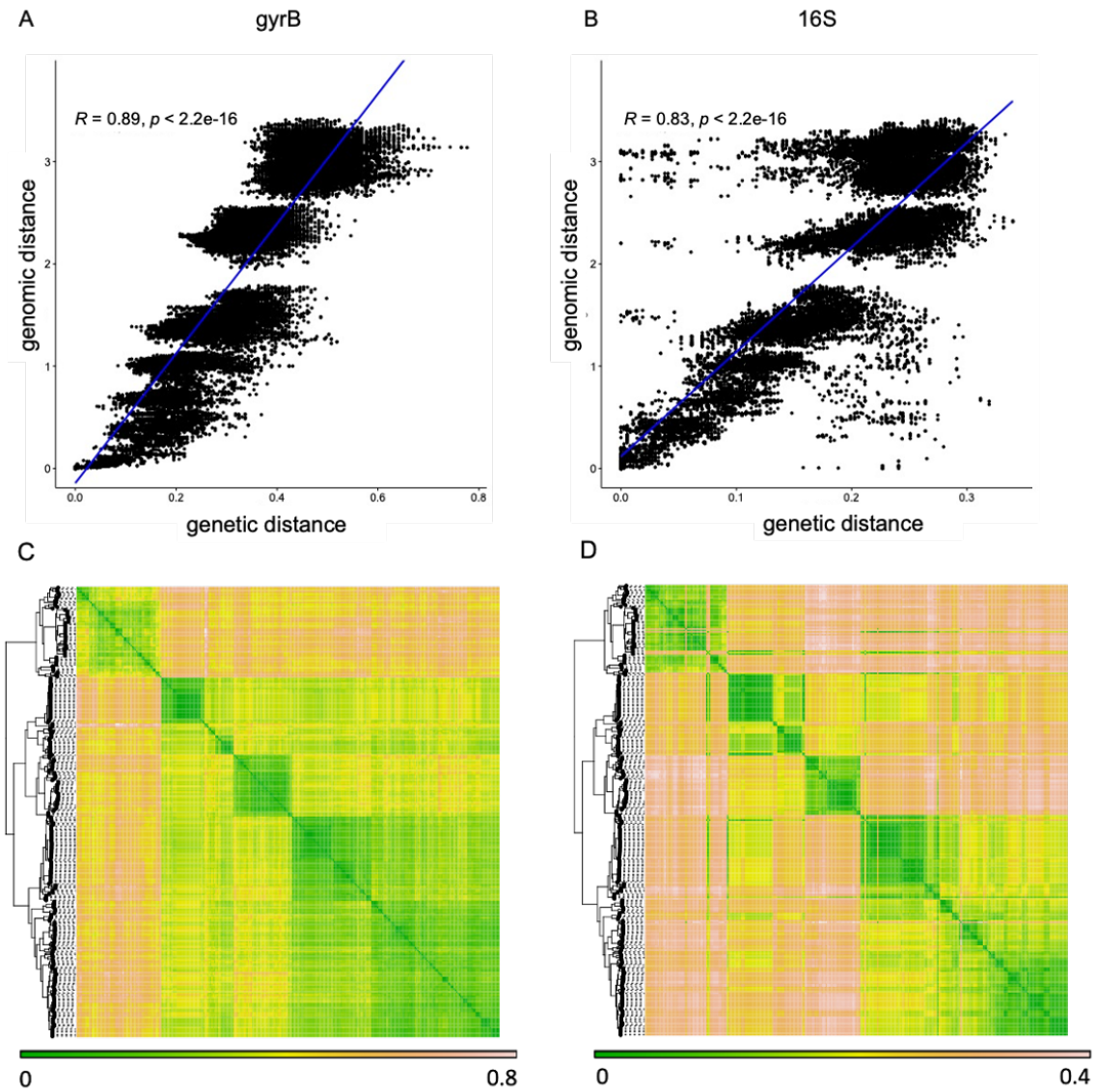

Figure 3.2: Comparing 16S and *gyrB* genetic distances compared to whole-genome phylogenetic distances. Whole genome phylogenies were built from 71 shared, single copy core genes in from 303 bacteria genomes sequenced from *A. thaliana* plant isolates. The pair-wise Jukes-Cantor distances of 16S (A) and *gyrB* (B) amplicon are positively correlated to the phylogenetic distances. When mapping pair-wise 16S (C) and *gyrB* (D) distances onto whole-genome phylogenetic trees (shown in the same order as both rows and columns), some 16S amplicons sequences are more similar to inter-clade bacteria rather than intra-clade.

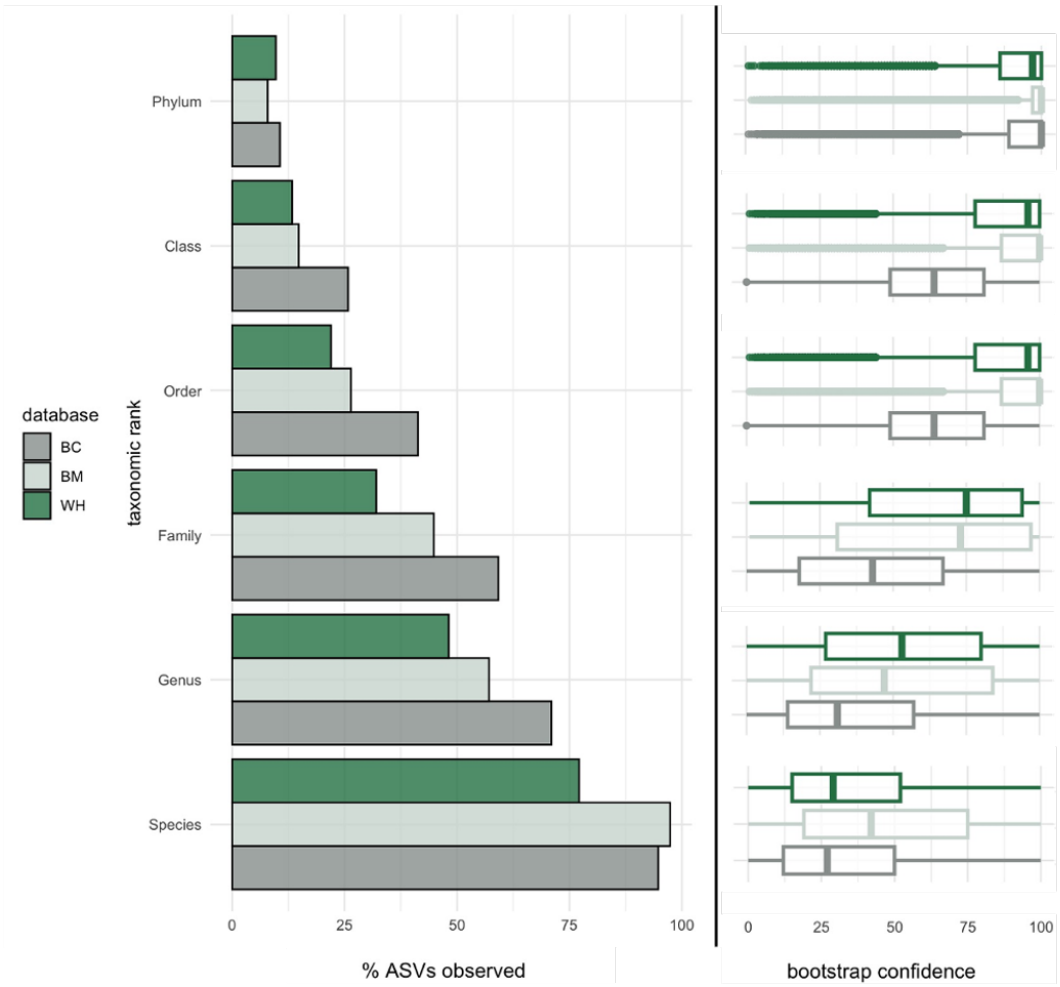

Figure 3.3: Comparing unclassified, distinct OTUs from Bartoli et al. (2018) using three distinct databases. We compared the resulting unclassified percentages from the published data (BC) compared to the Barret (BM) and the Whitehurst database (WH). Image B shows the collection of bootstrap confidences for the OTUs classified for each the respective taxonomic level and database.

CHAPTER 4

HERITABILITY OF THE MICROBIOME

4.1 Introduction

Microbial communities (microbiomes) are well documented to affect plant host health through a variety of interactions ranging from beneficial to pathogenic (Jones and Dangl 2006; Van Wees, Van der Ent, and Pieterse 2008; Vogel et al. 2021). The spectrum of potential host effects can be altered by emergent and dynamic microbe-microbe, host-microbe, and abiotic effects (Brader et al. 2017). Host-microbiome interactions are further complicated by the vast taxonomic diversity of the microbial community: a single plant often contains hundreds to thousands of distinct microbial species (Vorholt 2012; Bodenhausen, Horton, and Bergelson 2013; Brachi et al. 2022), each with their own spectrum of potential host effects. Thus, characterizing the phenotypes of these complex microbiomes proves a technically challenging task. Despite the challenges, quantifying the association between variation in the microbial associates across host genotypes is a necessary first step in identifying host-tractable components of the microbiome that will facilitate understanding of microbes' susceptibility to host defenses, emergent disease phenotypes, and influential host metabolic pathways and physiology (Grieneisen et al. 2021).

Broad-sense heritability (H2) is one such metric to better resolve the host genotypic effect on microbiome phenotypes. H2 here is quantified as the proportion of variance among microbiome taxonomic abundances attributable to host genotype (and is distinct from heritability as used in quantitative genetics: see Beilsmith et al. 2019 and Wagner 2021 for reviews). Measurable heritability of a substantial portion of all microbes has been reported in a wide variety of hosts, including plants (e.g. Horton et al. 2014; Walters et al. 2018; Brown et al. 2021), insects (e.g. Wu et al. 2021), and mammals (Grieneisen et al. 2021). Frequent and measurable amounts of broad-sense heritability in microbiome studies indicates

that the host genotype is ecologically important in shaping the microbial community taxa and abundances, but associations with genotypes are not necessarily resultant of host genetic variability (Henry and Bergelson 2023). For example, vertical transmission of microbes through seeds collected from different environments could create a maternal genotype effect unrelated to host-genetic control of associated microbes (Shahzad et al. 2018). Heritability estimates are therefore useful in identifying candidate microbes influenced by host-genotype but host-genetic control must be independently tested.

Genome Wide Association studies (GWAS) can help elucidate genetic components driving heritability by using single-nucleotide polymorphism (SNP) data to perform association tests, quantifying effects of genetic variation on the observed microbial phenotype. The inherent magnitude of multiple tests in GWAS requires controlling for false positives such as through Bonferroni correction. Avoiding uninformative phenotype testing can reduce testing thereby preventing unnecessarily further constraining the threshold for significance. In this way, heritability estimates also offer a systematic way to assess large groups of microbes for candidates to use in GWAS in order to identify host-genes shaping the microbiome.

Arabidopsis thaliana is a robust study system to examine broad-sense heritability and genetic control of the microbiome due to the availability of both highly homozygous, nearly clonal seed stocks of over 1000 distinct genotypes (accessions) with respective genomic sequence data. Horton et al. 2014 performed the first heritability and GWAS study of microbes in *A. thaliana* leaves using 196 genotype accessions replicated in the field. They performed heritability estimates on both community traits and the presence of particular bacteria through marker gene sequencing of the leaf phyllosphere. Host genetic variance explained 46% of species richness when analyzing the top one hundred abundant species. By using Gene Ontology (GO) annotation of putatively significant host genes, they observed that the most highly significant genes were associated with defense genes, while species richness was associated with enrichment in the viral regulation gene ontological category. The

few other studies examining natural microbial communities in *A. thaliana* also report heritability and associated genetic control of the microbiome (Bergelson et al. 2021; Brachi et al. 2022), and the observation is recapitulated in additional plant study systems (Deng et al. 2021). However, the plants' environment, phenotypic plasticity, developmental stage, and tissue type also play a significant role in microbial diversity and abundance: the confounding effects of these variables with host genotype are not well understood (Wagner 2021).

In one empirical study, Brachi et al. (2022) performed replicated field trials of 200 distinct *A. thaliana* accessions and their associated endophytic leaf microbiome phenotypes, across two years and four locations in Sweden. Communities were quantified by the relative abundances of microbial taxa using the 16S (v5-7) ribosomal subunit marker gene in bacteria. Heritability was quantified as the proportion of variance for a given bacteria's abundance that is attributable to the host genotype compared to the total variance measured (Box 1, Equation 4.4-4.5), and significant host genotype effects were observed for 10-22% of all leaf endophytes.

In Brachi et al. (2022), the phenotype is the variance of a given microbe's abundance among distinct host genotypes. If some proportion of a given microbial phenotype is attributed to host genotype, the estimated host genotype effects (α_k, Box 1 Equation 4.4) can be decomposed into distinct intercept values for each genotype called BLUPs (Best Linear Unbiased Predictors) for the focal microbe (Bernardo 2020). In this way, we can use microbial abundance data to identify microbe candidates that are influenced by host genotype, then use the estimated host genotype effects in subsequent GWAS to identify host gene-candidates driving the observed broad-sense heritability. In the GWAS performed by Brachi et al. (2022), only two significant SNPs were identified among all heritable ASVs, perhaps a surprising observation given the high broad-sense heritability reported among the microbes.

The phenomenon of few identifiable causal host gene candidates, but significant heritability, is not unique to this experiment or study system. The so-called "missing-heritability"

problem is well documented, but the explanations are broad and poorly defined, including biological (e.g. epistatic effects, complex traits, unidentified biological interactions) and technical (e.g. variance introduced during sample processing) explanations (Sandoval-Motta et al. 2017). For example, the Brachi et al. (2022) experiment used the canonical 16S (v5-7) marker gene, which has a low correlation between genetic distance and genomic distance, effectively grouping distinct bacteria into arbitrary taxonomic units (Hassler et al. 2022). The taxonomic grouping may mask an association of distinct bacteria to the host genotype if grouped bacteria have differing host-effects. The use of a less constrained gene in which distinct amplicon sequences better correlate with biologically distinct bacteria may increase estimates of host genotype effects and host gene candidates for individual bacterial taxonomic groups. We refer to these groups by their distinct amplicon sequences, called amplicon sequence variants (ASVs).

Moreover, heritability estimates may have upward bias and low accuracy due to potential non-normal distribution of host genotype effects, which can result from incomplete models (see Schielzeth et al. 2020 for a practical review). As previously suggested (Chen et al. 2018), microbial interactions are one potentially influential factor contributing to variance and masking host-genotype effects when omitted from phenotypic variance estimations. While host physiology shapes the ability of a microbe to successfully propagate in the face of host metabolites and immune responses, so do the co-occurring microbes. For example, microbes can encourage propagation of co-occurring microbes through the production of beneficial secondary metabolites, releasing effectors to suppress plant immunity response, and production of quorum molecules. Conversely, microbes can restrict growth of co-occurring microbes through competition of resources, production of antimicrobial molecules, and inducing plant immune responses (Finkel et al. 2020; Chen et al. 2018; Trivedi et al. 2020).

BOX 1: Heritability of host-associated microbes

Broad-sense heritability We first consider the definition of heritability, a term capturing the phenotype variance attributable to the host genotype:

$$H^2 = \frac{V_g}{V_p} \tag{4.1}$$

Given that:

$$V_g = V_a + V_d + V_i \tag{4.2}$$

$$V_p = V_g + V_e + V_{g*e} \tag{4.3}$$

Where V_g is the genotype variance, a summation of additive (V_a), dominance (V_d), and epistatic effects V_i. V_p represents the phenotype variance, a summation of genetic (V_g), environment (V_e) and gene by environment effects (V_{g*e}) (Nyquist 1991).

Microbe phenotypes For microbiome data, we can use log-transformed and ASV count tables to quantify the variance in the ASV (i.e. the phenotype) as:

$$Y_{ik} \sim a_k + \epsilon_{ik} \tag{4.4}$$

Where Y_ik is the log-transformed abundance data for an OTU i, and $a_k \approx N(0, \sigma_a^2)$ is a random intercept representing the kth host genotype, and $\epsilon_k \sim N(0, \sigma_e^2)$ captures the residual variance. We refer to this as the Host Genotype Only (HGO) model. The genotype effects and residual variance can then be applied to broad-sense heritability (H^2) *Equation 4.1*:

$$H^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2} \tag{4.5}$$

Therefore, H^2 is the mean variance assigned to the host-accession genotype as a proportion of the mean variance of the host genotype and the residual variance, which is conceptually comprised of all non-genotype variance including environment and gene-by-environmental variance as delineated in *Equation 4.3*.

Incorporating microbe-microbe interactions

$$Y_{ik} \sim a_k + \beta_j X_{ij} + \epsilon_{ik} \tag{4.6}$$

We can incorporate a variable for interacting microbes in Equation 3.6, where all variables are the same as defined for Equation 3.4 with the addition of X_{jk}, the log-transformed abundance for ASV$_j$, and β_j, the fixed effect regression coefficient of ASVj. ASV$_j$ was identified as the strongest-interacting ASVs with the focal ASVi, as determined by networks derived from MB inverse-covariance SPIEC-EASI methods. We refer to this as the ASV Included Variable (AIV) model.

Microbe-microbe interactions are complex, and the influence of microbe-microbe interactions compared to host-microbe interactions is not well understood. By including estimations of the fixed effects of co-occurring microbes, we may improve the accuracy of the variation at-

tributed to host genotypes, stabilizing the distributions of effects and improving the accuracy of the model estimates (Schielzeth et al. 2020).

Here, we explore two techniques to improve quantification of host-genotype effects on microbes and the candidate host genes driving those effects. First, we evaluate the impacts of increasing precision in bacterial identification. We use the *gyrB* marker gene, shown in Chapter 3 of this thesis to both have greater taxonomic resolution and a tighter correlation between genetic and genomic distance, thus allowing for the identification of more, taxonomically-distinct bacteria groups compared to 16S v5-v7. Second, we incorporate microbe-microbe interactions into analysis for heritability and host-genotype effects to better resolve the host genotype effects on each microbe. In particular, we test the effects of adding a covariate to account for microbial interactions to improve accuracy on the quantified genotypic effects. For each distinct bacterial taxa group, we identify the strongest putative interacting bacteria using co-occurrence networks and add the interacting bacteria as a fixed effect in the linear mixed model (Equation 4.6).

4.2 Results

Sample processing and taxonomic classification of plant microbiomes

We re-sequenced 1182 leaf microbiome samples from (Brachi et al. 2022) using 16S v5-7 and *gyrB* (Barret et al. 2015) primers (Methods). In brief, the samples were from 200 Swedish Arabidopsis genotypes planted in a randomized block field experiment in Ullstorp, Sweden in 2011 as established seedlings. Plants overwintered and were collected immediately after the first snowmelt. Five or six replicates per genotype were surface sterilized, freeze-dried and stored at -80°C until processing. DNA was extracted using a double enzyme-digest to include gram-positive bacteria, followed by a phenol-chloroform DNA isolation protocol. Given that several years had passed from the original sample collection and this-reanalysis, we purified

the DNA with Solid Phase Reversible Immobilization (SPRI) prior to amplification while also removing small degraded fragments (Rohland and Reich 2012). We performed dual-amplification of *gyrB* and 16Sv5-v7 of all samples using in-house designed two-step PCR (Illumina 2013). Sequences were trimmed for primers and adapters. Reads were further de-noised, clustered into amplicon sequence variants (ASVs), and assigned taxonomies using DADA2 (Callahan et al. 2016). We filtered out extremely rare ASVs before performing additional analysis, keeping ASVs with more than ten reads observed in more than six samples.

As expected, the *gyrB* identified more, distinct ASVs with 912 *gyrB* ASVs considered compared to 583 16S ASVs. Taxonomic profiling of the plant communities shows similar relative abundances at the family level when considering all samples, but with *gyrB* showing a higher percentage of Sphingomonadacea and a lower percentage of Oxalobacteracea compared to 16Sv5-7. Many of the ASVs observed here are uncommon, with the majority of ASVs observable in less than 20% of the samples (Figure 4.1).

4.2.1 gyrB *and OTU interaction covariates*

We next used our *gyrB* based ASV dataset to examine the effects of including a microbial interaction covariate in the genotypic effects and heritability estimates. We performed a SPIEC-EASI network analysis on the OTU abundance tables, which were transformed using the Additive Log Ratio (ALR). The ALR method uses an internal reference in each sample as opposed to the sample mean used in the center log ratio transformation (CLR) (see Quinn et al. 2019 for a review). The internal references used in this experiment were 16S and *gyrB* synthetic plasmids containing the respective primer binding sites flanking a synthetic "spike" sequence that was approximately the same length as the average bacterial amplicon sequence (Methods). The spike plasmids were added in equal amounts to each sample and co-amplified with the natural isolates' DNA. For each focal ASV, we identified the strongest

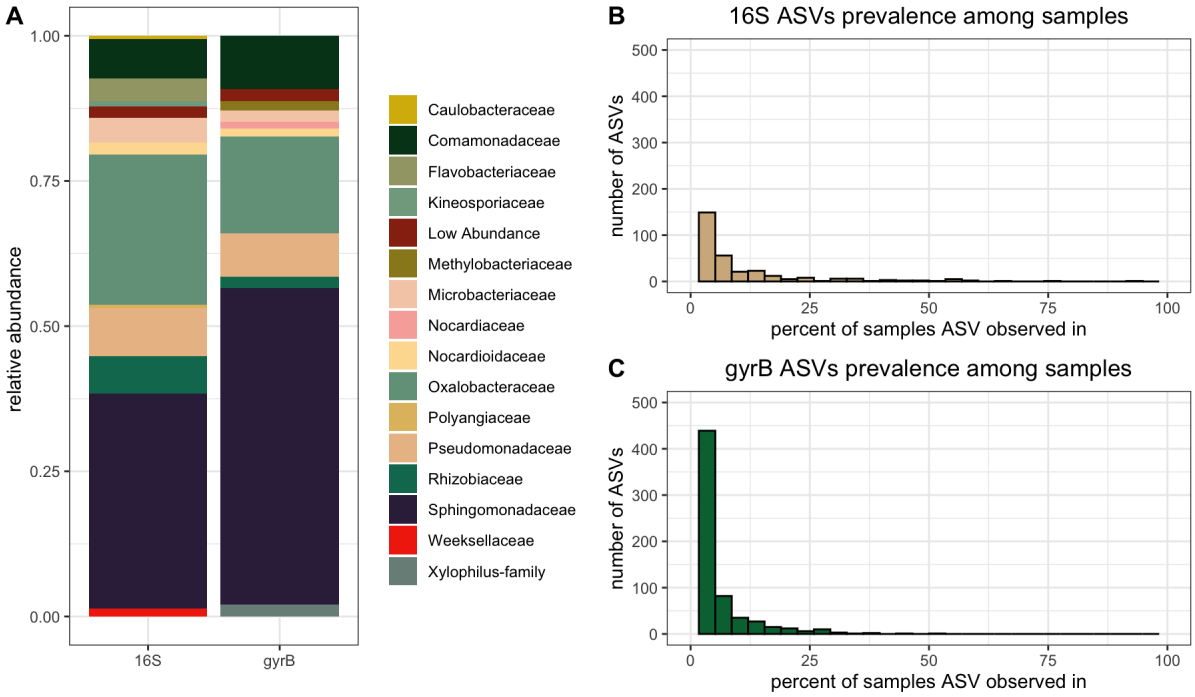

Figure 4.1: Comparing 16S and *gyrB* phyllosphere taxonomic community profiles: Microbiome taxonomic composition of 1182 *A. thaliana* leaf endophyte communities show similar profiles when comparing 16S and *gyrB* amplicons **(A)**. The *gyrB* methods identified more total microbial ASVs, but when reviewing the prevalence of 16S **(B)** and *gyrB* ASVs **(C)**, the majority of all ASVs were observed in less than 20% of all plant samples.

putative interacting ASV by ranking ASVs by interaction strength. We then performed a series of tests to compare the Host Genotype Only (HGO) model to the model with host genotype and ASV Interaction coVariate (AIV) model.

Here, we performed pairwise comparisons of the AIV and HGO models for each of the 912 *gyrB* ASVs by performing traditional null-hypothesis significance testing of the nested models using log-ratio tests of the models refit with ML. First, we tested the effects of randomly selecting an ASV for the fixed effect covariate (Equation 4.6) compared to the HGO model to evaluate if adding a covariate variable improved the fit by chance. When we randomly selected an ASV among putatively non-interacting ASVs for the covariate for the models, the random ASV covariate was identified as significant in only 7% of the model comparisons of the AIV to HGO models (61 of 912 tests,$\mathrm{P}_{Bonferroni} < 5e^{-5}$). After confirming that randomly

selected ASVs do not reliably improve the fit of the model, we tested the incorporation of the ASV interaction covariate informed by the SPIEC-EASI network. For each focal ASV, we identified the strongest, putatively interacting ASV for the fixed effect variable (Equation 4.6). The ASV covariate in the AIV model was identified as significant in 91% of the pairwise comparisons of the AIV and HGO models (831 of 912 tests, $P_{Bonferroni} < .05/912$).

We also estimated the Akaike information criterion (AIC) (Burnham and Anderson 2004) for both models using lmerTest (Kuznetsova, Brockhoff, and Christensen 2017) in R . In pairwise comparisons of ASVs using the two models, the AIV model had substantially higher levels of empirical support than the HGO model in 876 of the 912 cases (Δ_{AIC}=AIC_{HGO}-AIC_{AIV}, $\Delta_{AIC} > 2$). The average broad sense heritability estimate for the AIV model was significantly lower than HGO (Supplementary Figure 4.7, Wilcoxon Test $p<0.05$). Schielzeth et al. 2020 illustrate that the inclusion of missing covariates improves models by stabilizing the distributions of the parameters and biased estimates. We examined the model residuals and estimates to see the trends among our models. When reviewing the statistics of individual ASV models, we did see shifts towards more normal distributions of the residuals and the host genotype effects in AIV models (BLUPs) (Figure 4.2).

Additionally, when we compiled the model estimates for all ASVs between the two models, we saw AIV models reported smaller variability in the estimated residual variances and group (host genotype) variances, indicating that the ASV covariate is stabilizing the distributions as expected. Taken together, the model comparisons and shifts in the model parameter distributions indicate that incorporating the SPIEC-EASI informed ASV covariate improves our ability to model count data for any given ASV compared to omitting or randomly selecting ASVs. We used the AIV model for all subsequent analysis.

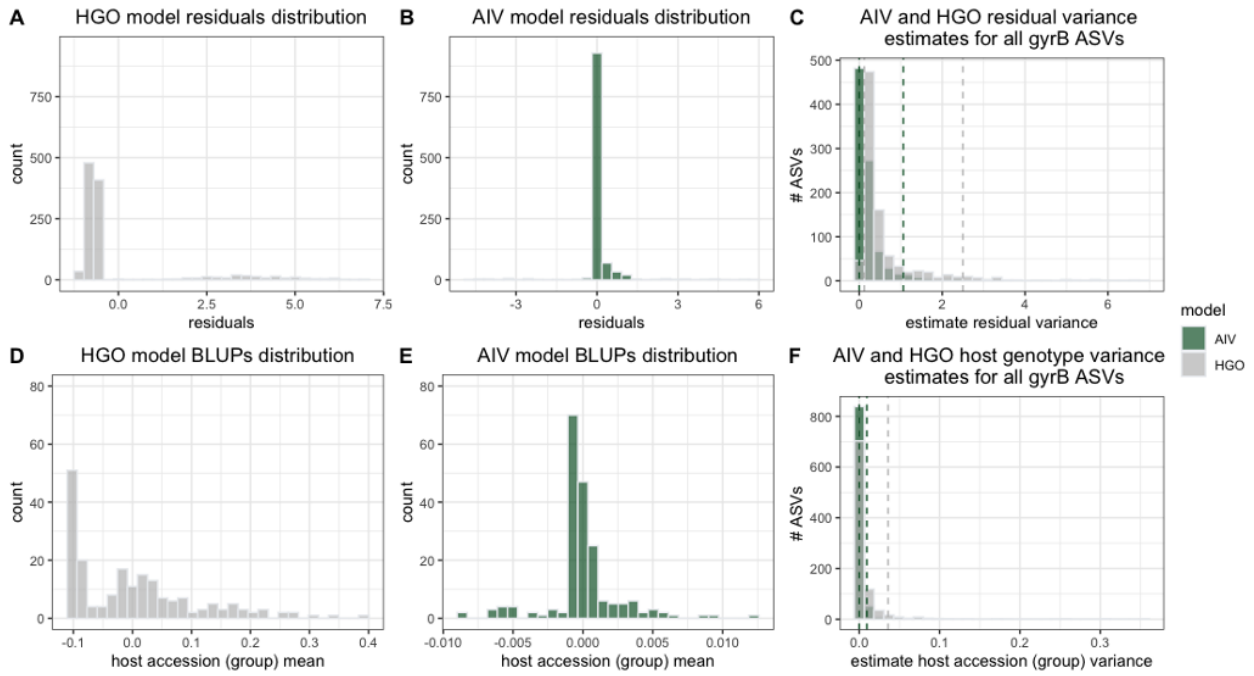

Figure 4.2: AIV model shifts parameters compared to 16S. The model residual and BLUP distributions of a given ASV visually seem to shift to a more normal distribution when comparing the HGO and AIV models, as seen here with Pseudomonas_16 (A, B, D, E). When comparing all 912 *gyrB* ASVs, the model point estimates for the residual variance (C) and genotype variance (F) reduce in variability when including the ASV interaction covariate.

4.2.2 Heritability varies among and within families

We first calculated heritability estimates for all ASVs in the 16S and *gyrB* data sets using the AIV model. Heritability was determined as the percent variance attributed to the genotype (Equation 4.5). Heritability estimates ranged from 0-10%. Surprisingly, heritability estimates from *gyrB* and 16S datasets were generally comparable, with some differences in variance (Figure 4.3).

We identified heritable ASVs as those that reported significant effects for the host genotype in the AIV models (log ratio test, $p < .05$). We identified 29 heritable *gyrB* ASVs and 34 heritable 16S ASVs. The Sphingomonas family reported the largest number of heritable ASVs for both 16S and *gyrB*, which is also the family with the greatest number of distinct ASVs in both the 16S and *gyrB* datasets.

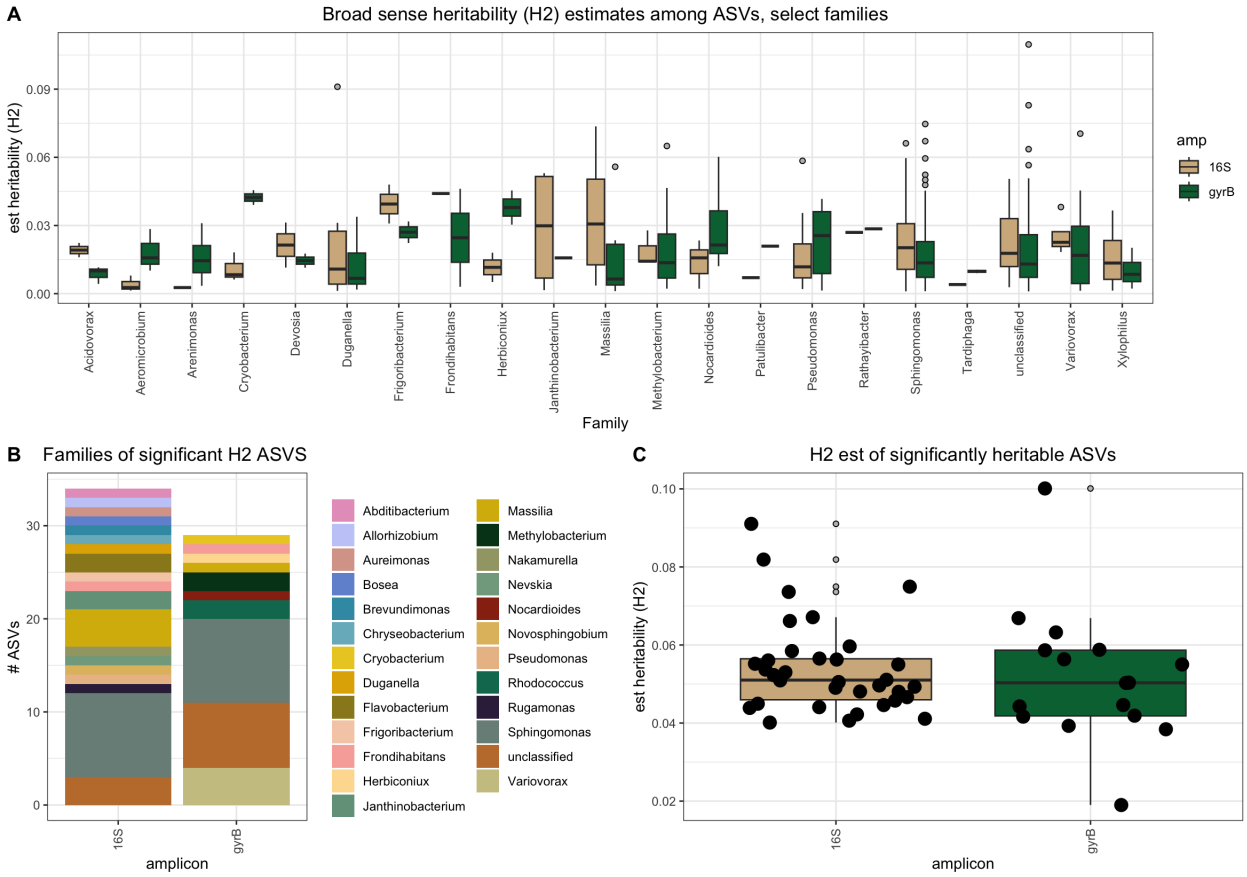

Figure 4.3: Broad-sense heritability estimates vary within and among families. Taxonomic grouping, here captured through 16S and *gyrB*, also influences the heritability estimates observed. Heritability estimates determined on ALR transformed data. There is no consistent difference between 16S and *gyrB* estimates of broad-sense heritability (H2) among families (A), and different families were identified as heritable (B). Of ASVs identified as significantly heritable, the H2 estimates were comprable between 16S and *gyrB*(C).

4.2.3 GWAS candidates

We next performed GWAS for each of the 63 heritable ASVs identified using the 16S and *gyrB* AIV model using GEMMA, accounting for population structure using a centered genetic relatedness matrix (Zhou and Stephens 2012). Of the heritable *gyrB* ASVs, 22 of 29 reported significant SNPs while only 13 of the 34 heritable 16S ASVs reported significant SNPs (Wald Test, $P_{Bonferroni} < .05/1{,}035{,}775$)(Table 4.1). We then looked at the proximity of the *gyrB* and 16S significant SNPs to known, annotated host genes. Significant SNPs identified in

16S and *gyrB* GWAS summaries				
amplicon	total ASVs	heritable ASVs	ASVs with significant GWAS SNPs	n distinct host gene candidates
16S	681	34	13	50
gyrB	931	29	22	287

Table 4.1: ASV summaries and GWAS results for 16S and *gyrB*

A. thaliana GWAS may occur up to hundreds of kilobases away from the casual host gene due to variance in the concordance of the host genotypes or allelic heterogeneity (e.g., Gloss et al. (2022)). We used a sliding window of 10kb among all significant SNPs and identified 92 distinct groups among the *gyrB* SNP candidates, while we identified 13 distinct groups among the 16S SNP candidates

We reviewed all gene candidates within 500 bp of significant SNPs, identifying 50 and 287 gene candidates from the 16S and *gyrB* significant SNPs, respectively (Appendix D). The 16s and *gyrB* host gene candidates were mostly distinct from each other, having only three candidate genes in common: AT5G35410, AT5G16830, and AT5G16840. AT5G35410 codes for a protein involved in the plant's potassium regulation. The AT5G16830 gene codes for a protein in the SYP20 family of protein receptors. AT5G16840 binds to a known fungal elicitor and contributes to ROS mediated plant immune responses. We performed a PANTHER Gene Ontology (GO) enrichment analysis (Mi, Muruganujan, and Thomas 2013; Thomas et al. 2022) between the 16S and *gyrB* gene groups for all genes within 500bp of significant SNPs but observed no statistically significant enrichment for any functional groups in either of the data sets (Supplementary Figure 4.4).

4.3 Discussion

Microbial community interactions are acknowledged to play a role in host–genotypic effects on microbes but are typically studied as whole microbiome community phenotypes, such as taxonomic relative abundances or disease symptoms (e.g. Beilsmith et al. 2019; He et al. 2021. To decompose the community interactions, we turn to individual ASVs. While several studies in *A. thaliana* and other hosts have identified and quantified individual heritability of individual ASVs, they omit microbe-microbe interactions (Deng et al. 2021; Brachi et al. 2022). Here we show that incorporating microbe-microbe interactions increases our ability to assess heritability and host-gene candidates driving microbial variation by improving model fits. By improving the model fits, we subsequently better stabilize the distribution of ASV model parameters, which should improve the precision of the model estimates (Schielzeth et al. 2020).

The precise mechanisms driving improvement in estimates of heritability with consideration of microbe-microbe interactions are technically and biologically conflated. Technically, incorporation of the microbe-microbe interactions might partially capture the between sample count variances. We show that our network-informed microbial interactions were successful in incorporating microbe-microbe interactions in the model, while randomly selected microbes improved the model only rarely. When incorporating the ASV interacting covariate, the estimated environmental covariate stayed relatively consistent between the HGO and AIV models for the majority of heritable ASVs, both in the 16S and *gyrB* datasets (Supplementary Figures 4.5-4.6). Thus, the improvement of fit of AIV models compared to the HGO models is likely attributable to the stabilization of the parameter distributions facilitated through the addition of the interacting ASV covariate, resulting in more precise heritability and BLUP estimates.

Interestingly, we also show that finer delineation of taxonomic grouping (through the use of less constrained marker gene, *gyrB*) results in similar heritability estimates among

families compared to the more constrained marker gene (16S v5-7). We hypothesized that 16S grouping may be masking host-genetic effects by grouping functionally distinct strains with presumably distinct host-genetic interactions. However, the broad-sense heritabilities between the *gyrB* and 16S across distinct families proved similar. Artificially high rates of heritability may partly result from sampling error due to sparse data sets, especially in datasets with skewed or bimodal distributions (Schielzeth et al. 2020). Our data are indeed sparse, with the majority of ASVs in less than 20% of the samples in both datasets.

We found that 16S reported a significant, moderately strong correlation between prevalence and heritability (p = $1.52e^{-16}$, $P_{Spearman}$=.33, Supplementary Figure 4.5) while *gyrB* did not (p = .09, $P_{Spearman}$=-.06, Supplementary Figure 4.5). This could indicate that the finer ASV grouping defined by *gyrB* stabilizes the parameter distributions by partitioning out distinct ASVs groups with distinct modes. More generally, the finer taxonomic resolution provided through *gyrB*, compared to 16S, mitigates bimodal distributions of model estimates through more precise grouping of the composite microbes and improves the ability of the linear mixed model to fit the data and estimate heritability. Thereby, the *gyrB* grouping is less biased in sparse data sets because the distribution of the parameters is more often better stabilized than the 16S count data (Schielzeth et al. 2020).

The ability to better resolve ASVs using *gyrB*, and its subsequent hypothesized improvement in estimating genotype variance, was also validated when we consider the number of host gene candidates, as well as the number of ASVs identified as heritable. In particular, 76% of the heritable *gyrB* ASVs reported a significant SNP, whereas only 38% of heritable 16S ASVs reported a significant SNP in the GWAS. This was true despite the fact that more 16S ASVs were identified as heritable compared to *gyrB* (5% and 3% of the 16S and *gyrB* ASVs, respectively). While we identified more 16S gene candidates than the initial Brachi et al. (2022) analysis, we recapitulate the general trend of heritable ASVs lacking host-gene candidates driving heritability. Using *gyrB* ASVs looks to mitigate this issue, but candidate

genes need to be empirically validated.

We identified 338 host gene-candidates within 500 bp of significant SNPs ($P_{Bonferroni} < 1e^{-6}$). Interestingly, we did not identify the same GWAS SNPs as Brachi et al. (2022), who identified SNPs using thresholds more conservative than the thresholds employed in this study. One reason for the disconnect between the original analysis of these samples and our reanalysis is that the sequencing platforms have advanced. At the time of the Brachi et al. 2022 analysis, sequences were neither long enough nor of sufficiently high quality to merge reads for the targeted amplicon region. Thus, only the forward reads were used and distinct 16S sequences were grouped into single OTUs due to technical limitations. For example, the most abundant OTU amplicon sequence (OTU B1) in the 2018 Ullstorp analysis was found as a perfect sub-string match to 34 distinct ASV amplicon sequences in the current analysis.

While the missing heritability phenomenon is typically observed among microbiome studies, some studies also reported high numbers of GWAS gene candidates. In one comprehensive study on host-genetic control of the *Sorghum bicolor* rhizobium community and distinct bacteria taxa (OTUs), Deng et al. 2021 report high levels of heritability of both community composition and OTUs. They looked for GWAS candidates that explained community compositions and found a single prominent peak. They then identified bacteria that also had SNPs called in the same 1.5Mb region as the peak of interest. They found approximately 40 OTUs with significant SNPs in the region of interest, indicating that there are many GWAS candidates across the entire host genome.

In conclusion, we illustrate the value of incorporating microbe-microbe interactions in the estimations of host genotype effects and broad-sense heritability estimates. We observe over 300 significant gene candidates as potentially significant host genes shaping leaf microbiomes among *A. thaliana* candidates. While the genes should be empirically tested, the inclusion of ecologically significant interactions allows for the improved identification of host genes for future studies on host-genetic influence over associated microbes.

4.4 Materials and Methods

4.4.1 Sample Collection and DNA extraction

Arabidopsis thaliana rosette samples were collected and processed as previously described, see Brachi et al. 2022 for detailed methods. Briefly, we identified 200 *A. thaliana* accessions (inbred genotypes) collected from Sweden for which we had whole genome-sequences. Six replicates of each accession were started in trays, thinned, and transferred to field plots in Ullstorp Sweden in 2011 using a randomized block design. Plants overwintered and were collected in the spring 2012, a few days post snowmelt. Plant rosettes were stored in separate envelopes on dry ice and moved to -80°C storage until processing. The chloroform-phenol DNA extraction employed a double enzymatic digestion of Proteinase K and Lysozyme to ensure yields from gram-positive bacteria.

The DNA aliquots extracted in 2012 were re-purified and size selected for fragments >1000bp for the current analysis. We performed a Solid Phase Reversible Immobilization (SPRI) bead clean up on the samples prior to amplification (Rohland and Reich 2012). Beads were added using a 0.6:1 bead to sample ratio, vortexed, flash spun, incubated 5 minutes at room temperature, then set 3 minutes on the magnetic stand at room temperature. The supernatant was subsequently withdrawn, followed by two 80% EtOH washes, re-eluted in the same starting volume using molecular grade water, vortexed, flash spun, and incubated for another 5 minutes on the magnetic stand at room temperature. The supernatant was withdrawn and dispensed into a clean plate.

4.4.2 PCR Amplification and Library Prep

We co-amplified 16S and *gyrB* marker genes for bacterial taxonomic identification for each plant sample. We used the oligo structure described in (Illumina 2013), which uses a two-step PCR protocol.

PCR1: The initial PCR used marker-gene primers with internal indexes (Bartoli et al. 2018; Chelius and Triplett 2001) for *gyrB* and 16S primers, respectively) and an Illumina adapter overhang on the 5' ends, resulting in the construct (5'-[adapter overhang][internal index]-[16S or *gyrB* primer]-3'). See Appendix C full construct sequences. We used Kapa HiFi Kit (Roche, Basel, Switzerland, #KK 2502) from the same lot for all PCRs. PCRs were performed in triplicate using 15 ul reactions comprised of 4.1 ul molecular grade water, 3.0 ul KAPA HiFi Fidelity Buffer (5X), 0.5 ul dNTP mix, 0.4 ul KAPA HiFi HotStart DNA Polymerase, 1.5 ul [6.6 mM] F+R 16S primer mix, 1.5 ul [1.6mM] F+R *gyrB* primers mix, 0.25uL [.014 nM] 16S spike plasmid, 0.25uL [.014 nM] *gyrB* spike plasmid. PCR thermocycler program consisted of an initial denaturation step for 5 min (98°C), followed by 29 cycles of a 30 second denature (98°C), 30 second annealing (61°C), 45 second extension (72°C), then a final extension phase for 5 min (72°C) before holding at 4°C. Triplicate replicates were pooled and processed together for the remainder of the protocol. Due to high amounts of primer dimers formed by the degenerate *gyrB* primers, samples were size-selected using SPRI beads as described above, modified to use a bead to sample ratio 0.9:1.

PCR2: PCR primer constructs used standard Illumina indexes and adapter sequences as described in (Illumina 2013). PCR reactions comprised of: 3.1 ul molecular grade water, 3.0 KAPA HiFi Fidelity Buffer (5X), 0.5 ul dNTP mix, 0.4 ul KAPA HiFi HotStart DNA Polymerase, 1.5 ul [5 mM] forward primer, 1.5 ul [5mM] reverse primer, and 5 ul of PCR1 product. PCR thermocycler program consisted of an initial denaturing step for 5 minutes (98°C), followed by 12 cycles of a 30 second denature (98°C), 30 second annealing (55°C), 40 second extension (72°C), then a final extension phase for 5 min (72°C) before holding at 4°C. Samples were pooled by plate and size selected using SPRI beads as described above, using 0.8:1 bead to sample ratio and re-eluted into half the starting volume using molecular grade water. Samples were size selected for 300-bp to 700-bp product using the BluePippen (Sage Science, Beverly, MA, USA) 1.5% agarose kit (SAG-CDF1510) . Samples were sequenced

across two sequencing suns on the Illumina MiSeq V3 PE300 chemistry (Illumina, San Diego CA, #MS-102-3003) at the New York University Genomics Core.

4.4.3 Data Analysis

All data was analyzed using R version 4.1.1 (R Core Team 2021) unless otherwise noted.

Sequencing data preprocessing

For sequencing data For sequence data preprocessing, we used BBTools, a suite of command line tools using kmer based algorithms (Bushnell, Rood, and Singer 2017). Because amplicons for a given sample were co-amplified and both amplicon products used the same Illumina and internal barcodes, sequencing data was first sorted into *gyrB* and 16S datasets. We mapped and sorted the reads to 16S or *gyrB* primer sequences using BBtools seal (parameters copyundefined restrictleft=25 k=14 hdist=1 kpt=t), trimmed for adapters using BBtools bbduk (ktrim=r k=23 mink=11 hdist=1 tpe tbo), sorted by internal barcodes using BBtools seal (match=first k=5 restrictleft=7), and trimmed primers using BBtools seal (copyundefined ktrim=l restrictleft=40 k=13 mink=11 hdist=1).

ASV tallies

We used FIGARO (Weinstein et al. 2019) to identify optimal maximum error rates and trim positions for 16S and *gyrB* amplicons for each flowcell to retain the highest percentage of reads; however, *gyrB* parameters required additional refinement due to the short amplicon sizes being outside the scope of figaro parameters. See scripts for specific parameter settings. Using DADA2 (Callahan et al. 2016), we trimmed reads by size, estimated error rates (learnErrors), generated dereplicated sequences, applied the DADA2 algorithm, merged reads (maxMismatch =0, minOverlap = 10), then removed chimeras. We then collapsed identical reads into the same group if there was read length variation or a shift in

alignment using DADA2 “collapseNoMismatch”: effectively, shorter read sequence counts could be absorbed into a group representing a longer read containing the identical sequence. We then combined ASV tables for each amplicon across both runs and used the DADA2 naïve baysian classifier in combination with the SILVA 16S database, nr99_v138.1 (Quast et al. 2013) or an in house *gyrB* database, each modified to contain the spike sequence, for 16S and *gyrB* identification, respectively.

ASV normalization, transformation, and network analysis

All reads were initially filtered to keep only sequences observed with more than three reads in more than six samples. Read counts were then normalized computationally by starting DNA input as quantified through Quant-iT Picogreen (Invitrogen, Waltham MA #P7589) on the Tecan Spark 10M. We performed normalization on the data sets by generating a normalization vector based on the samples' DNA concentration. We performed the ALR transformation through the vegan package (Dixon 2003) using the spike reads as the reference vector, and a pseudo count of one. The SPEIC-EASI package (Kurtz et al. 2015) was locally modified to forgo the program's requisite CLR transformation, but was otherwise processed using default parameters. Prior to linear mixed modeling, we scaled the transformed data.

AIV Model analysis

For model development, we asserted a "bottom up" approach. We used the ALR transformed ASV data in the AIV and HGO models (defined in Box 1) using lme4 (Bates et al 2015) . We used anova function in R to perform log ratio test on the pairwise comparison of the respective ASV in the AIV and HGO models ($P_{Bonferroni} < 0.05/912$).

Heritability estimates and Heritable ASVS

Broad sense heritability estimates used the AIV models for all ASVs' ALR transformed counts, which were then input into the respective models. We extracted the variance of the host-genotype and dividing by the sum of all variance in the model using "varcompmer" in the HLMdiag R package (Loy and Hofmann 2014). To identify heritable ASVs candidates, we used the AIV model with and without the host genotype random intercept covariate, using the anova function to perform log ratio tests to quantify the significance of the host-genotype. We identified heritable ASV candidates as those with the host genotype effect greater than 0, and with $p<0.05$.

Genome Wide Association Mapping

We ran GWAS for each heritable ASV identified within the four data sets using the methods described in (Brachi et al. 2022), which drew upon SNP data collected by the 1001 Genome Project (The 1001 Genomes Consortium 2016). We extracted the Best Linear Unbiased Predictors (BLUPs) from the host genotype random intercept for the ASV using lme4 (Bates et al. 2015) in R. Thus, the phenotype for each GWA performed used the estimated effects of the 200 host genotypes per heritable ASV. We used GEMMA (Zhou and Stephens 2012) to estimate SNP-based (pseudo)heritability while accounting for population structure by including a genetic relatedness matrix (gk -1). We determined SNP significance by Wald Tests (-lmm 1) with $p_{Bonferroni} < .05/103{,}000{,}000$ SNPs.

4.5 Supplementary

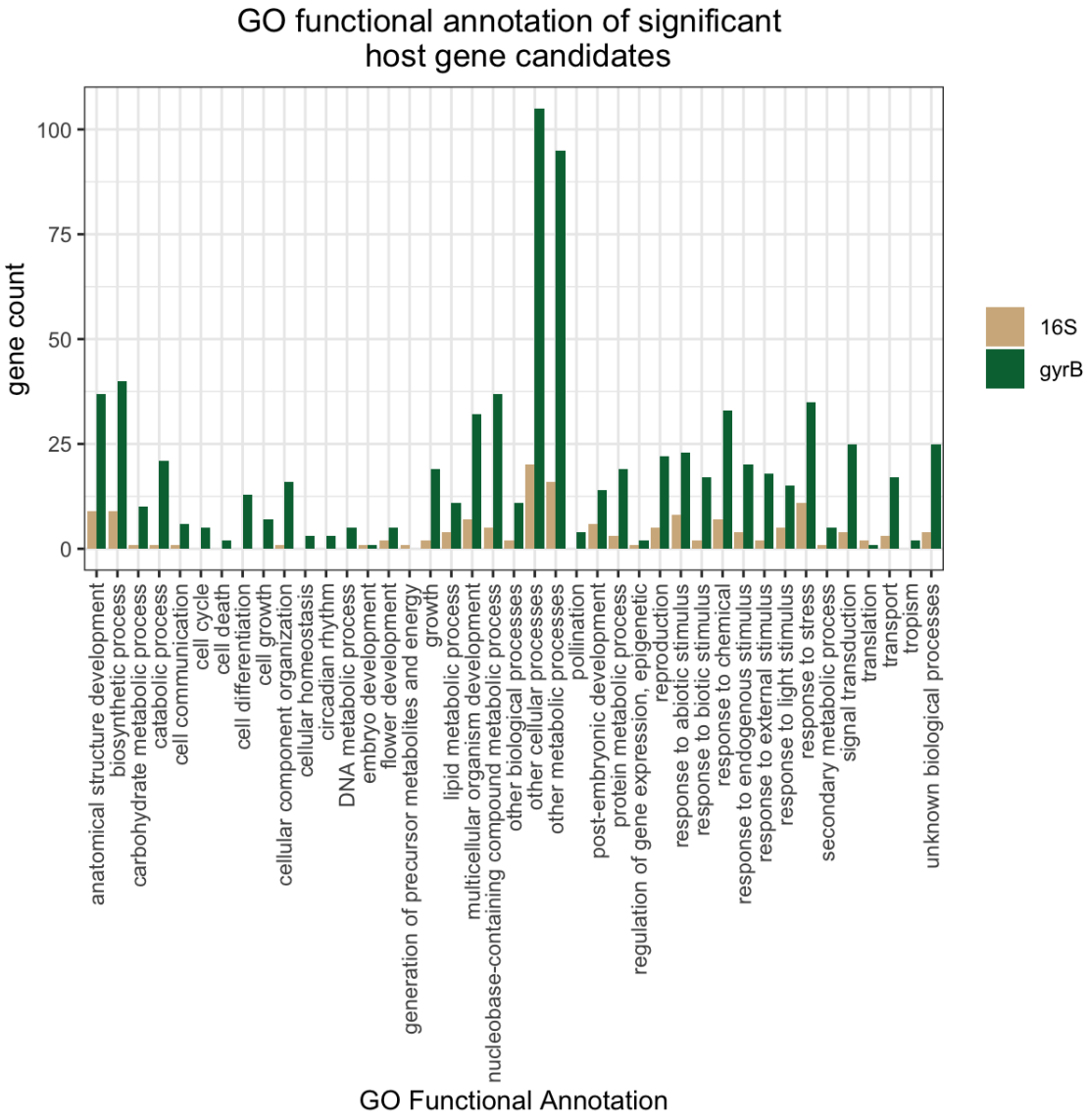

Figure 4.4: GO annotation of *A. thaliana* gene candidates influencing heritability 16S and *gyrB* ASVs. There was no statistically significant enrichment among functional groups assigned to the 16S and *gyrB* candidate genes.

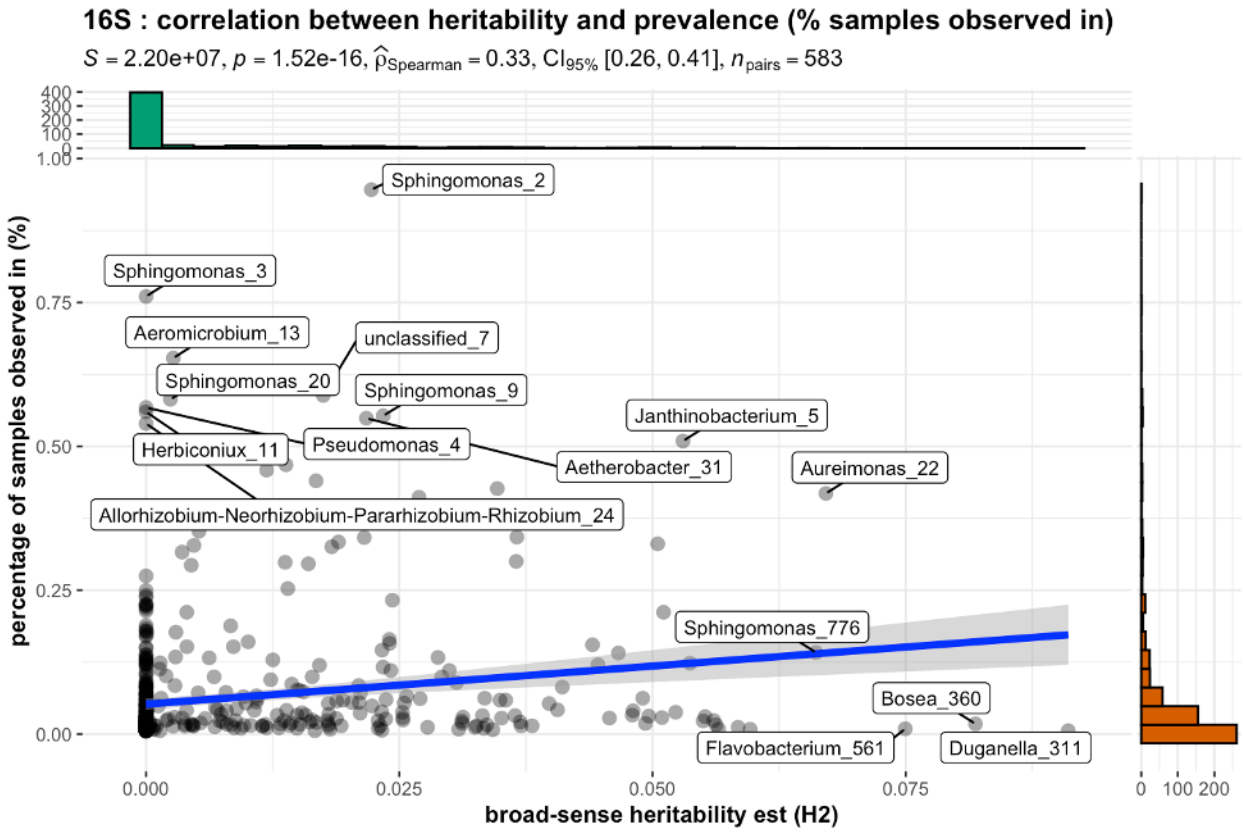

Figure 4.5: Summary statistics and visualization of the correlation between prevalence and broad-sense heritability (nonparametric testing) for 16S ASVs

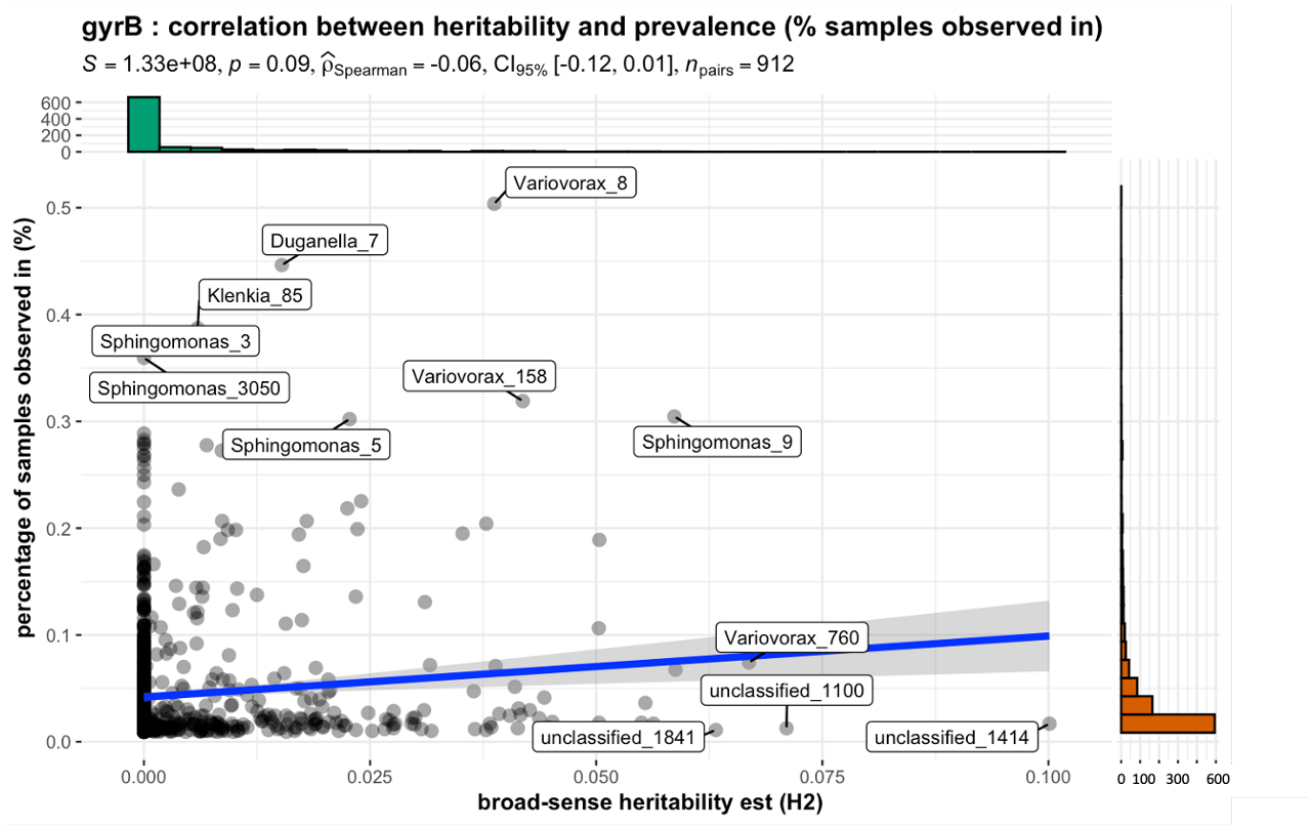

Figure 4.6: Summary statistics and visualization of the correlation between prevalence and broad-sense heritability (nonparametric testing) *gyrB* ASVs)

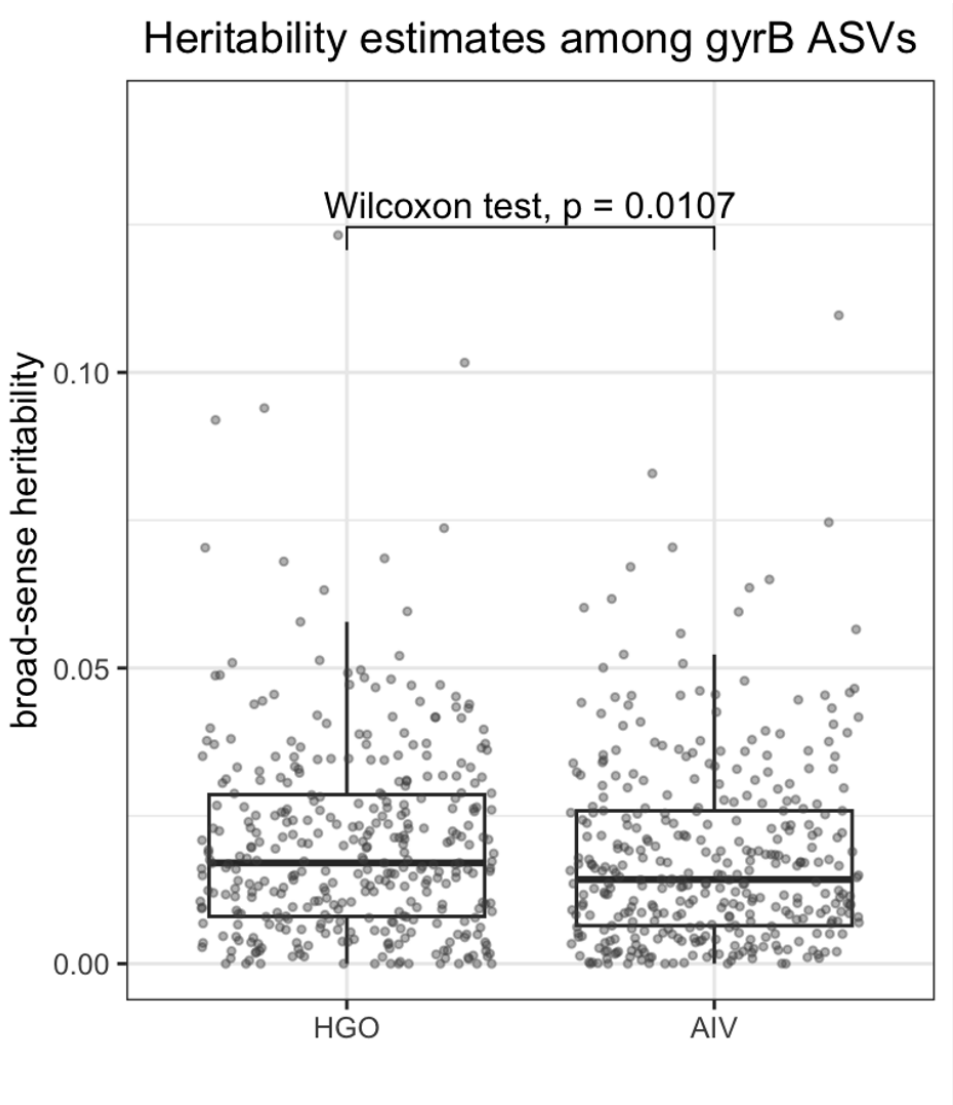

Figure 4.7: Broad-sense heritability estimates are significantly lower using the AIV model with *gyrB* ASVs (paired Wilcoxon test $p<0.05$)

CHAPTER 5

CONCLUSION

Microbiomes alter the host phenotype; they are composed of hundreds to thousands of distinct species co-occurring within and on host tissue, and are reciprocally shaped by host physiology (Brader et al. 2017). Due to small amounts of biological material and technical processing artifacts, noise in these systems can exceed the magnitude of their biological correlations, making significant ecological interactions difficult to detect. The initial objective of my dissertation aimed to improve estimates for host genotype effects on microbes using increased taxonomic precision and absolute counts, with the ability to verify through empirical testing using natural isolates. I expanded my analysis to include microbe-microbe effects to refine estimates on heritability of the microbiome.

My three research chapters progress to show how the collection of isolates from natural *Arabidopsis thaliana* plants provided the microbial "living library" required to build and validate a *gyrB* taxonomic database for the system. I then successfully apply the database to complex communities in *A. thaliana*. Through ongoing analysis of these datasets, I also identify microbial ecological interactions as a promising variable to help explain trends in differential microbial abundance data. I specifically show that the inclusion of microbe-microbe interactions in *gyrB* defined communities provides a more informative model to asses host-genotype effects on microbes found in *A. thaliana* associated endophyte communities.

The Chapter 2 describes the extensive collection of natural isolates from *Arabidopsis thaliana* leaves. I used both *gyrB* and 16S variable regions to taxonomically identify the bacteria. These isolates provided the biological samples required for my future experiments and allowed me to validate the *gyrB* taxonomic identification methods before applying the method to complex communities. For decades, *gyrB* had been reported as a single copy gene highly reliable in phylogenetic identification (Watanabe et al. 2001) and has more recently been shown to successfully characterize complex seed and plant communities using

novel primers (Barret et al. 2015; Bartoli and Roux 2017). This precedence mismatched my own observations in my lab work; even for single isolates, I often observed dual amplicon sequences when using the published primers, and the sequences were not taxonomically identifiable using the reportedly reliable taxonomic database methods previously described in the literature. This led to substantial troubleshooting in the lab as I tried to identify possible contaminants and missteps in my data processing methods. Later, I would find the work done by Poirier et al. 2018 describing the *parE* co-amplification using *gyrB* primers, nicely explaining my laboratory confusion.

I was eventually able to expound the *gyrB* methods and build a database for the field which resulted in more precise bacteria taxonomic identification. Still, this experimental bottleneck centered on *gyrB* development led to an extension of the isolate project and my subsequent research. I intermittently paused the collection of new isolates as I tried to troubleshoot, and I regret not being able to consolidate the collection of isolates to a single, uninterrupted amount of time. Biological samples of microbial communities change over time in storage as bacteria become non-viable. Laboratory variables also change and influence microbial propagation: media ingredients, consumables, personnel, and instruments. With changing starting materials and variables, I did not feel confident doing more specific analysis of the bacteria isolate communities between leaves, plants, and populations. Still, the effort put forth to collect natural bacterial isolates in Chapter 2 proved rewarding. We collected 12 of 43 microbiome “hubs" previously identified in Swedish *A. thaliana* leaf communities. Moreover, I generated whole-genome sequences for one of those isolates, *Brevundimonas strain B38*, and empirically validated computationally inferred interactions, showing the microbe's positive effect on host plant growth in single-inoculation experiments.

I apply the *gyrB* database built in Chapter 3 to characterize complex microbial communities in Chapter 4. I discuss our initial hypothesis that *gyrB* would significantly increase the heritability estimates compared to 16S, but end up finding that the estimates were compara-

ble. While I argue that the validity of the *gyrB* estimates is more likely biologically relevant, there is also the interesting variable of taxonomic scale and levels of ecological interactions. If a single locus polymorphism in a microbe confers a differential host effect, then *gyrB* is more likely than 16S to tract that variation due to the distinct concordances described in Chapter 3. Alternatively, consider the possibility that species within a genus have equal host effects but compete with each other for colonization of the plant and have stochastic success. Decomposing this genus group into distinct strains for the heritability analysis would result in a loss of statistical power, and so broadening the taxonomic grouping, say through 16S identification, would likely be the better option in identifying host-genotype interactions. Examining these taxonomic groupings in further detail would be a way to identify important genes versus conserved complex gene traits influencing host-microbe interactions.

In addition to the use of *gyrB* to characterize communities in Chapter 4, I incorporate microbe-microbe interactions. The inclusion of microbe-microbe interactions improves model fits to the data compared to host-genotype only model methods. This idea was born from my extensive fixation on these data after initial investigations proved less insightful. Primarily, I initially hypothesized that using synthetic spike-in oligos would significantly improve the model fit estimates because I would be able to employ more precise statistics using additive-log ratio transformations (ALR) compared to the canonical center-log ratio (CLR), specifically for rare microbes. I did see some improvements anecdotally (e.g. mildly higher significance estimates in the GWAS components). Overall, the data analysis results were similar enough that I elected to only use the ALR data analysis and pivoted towards exploring the role of microbe-microbe interactions in heritability estimates. The minimal information gained through ALR at the cost of sequencing is also noteworthy, but could theoretically be the result of my relatively low sequencing coverage. The impact of coverage on this conclusion would have to be tested with higher coverage data: another potential interest for future researchers.

While microbe-microbe interactions are a promising component of host-genotype effects on the microbiome, I believe there is more to be understood in the experimental technical components (albeit less exciting in the terms of developing ecological theory). These microbial ecosystems are inherently complex. Each microbe shows stochastic biological variance in the abundance data, but we as researchers introduce substantial stochastic technical variance that is imperceptible through canonical plate or batch effects. For example, in the Brachi et al. (2022) paper included in Chapter 1, and revisited in Chapter4, there are specific wells in the sample plates that did not generate sequencing data across all sequencing library preparations, presumably an issue with the multi-channel pipette used on the liquid-handling unit. While this issue is at least blatant enough to be captured through read count analysis quality checks, it is easy to imagine more subtle, consistent bias influencing the study system that are not so clearly identifiable. In sparse data sets, these biases have the potential to influence the data in significant ways due to sampling error. Empirical testing through isolates proves to be a necessary component to better understand the technical artifacts embedded in the data analysis. To verify inferred ecological interactions and technical artifacts, synthetic community development and microbial competition experiments in planta using natural isolates would be a natural progression of the research outlined in this dissertation.

To truly recapitulate an *A. thaliana* natural leaf microbiome, one would need to expand beyond bacteria to include other prominent microbes, including oomycetes, fungi, and viruses. The data reviewed across this thesis are limited in that they focus only on bacteria in the leaf endophytic communities. To support future researchers in broadening the scope of empirical studies, I also collected oomycetes and fungi from the same plant leaves used for the bacteria collection. The fungal samples are preserved but yet to be identified. The bacteria isolate collection also presumably contains oomycetes, which would be a subset of the isolates that did not successfully amplify *gyrB* sequences. Future researchers could verify the identity of the oomycetes and fungal samples and combine them with bacteria in

synthetic microbiome studies to better understand natural *A. thaliana* microbiomes in their full complexity.

CHAPTER 6

APPENDIX A: ISOLATION MEDIA

The media listed below were taken from Bai et al. (2015). Agar was added for media plates and omitted for liquid cultures.

MEDIA	COMPONENTS	AMOUNT
TSB (pH: 7.2)	Casein (pancreatic digest)	17 g
	Soya peptone (papaic digest)	3 g
	NaCl	5 g
	K2HPO	2.5 g
	Dextrose	2.5 g
	(Agar)	20 g
	diH2O	to 1000 mL
1/10 TSB (pH: 7.2)	Casein (pancreatic digest)	1.7 g
	Soya peptone (papaic digest)	0.3 g
	NaCl	0.5 g
	K2HPO	0.25 g
	Dextrose	0.25 g
	(Agar)	20 g
	diH2O	to 1000 mL
TYG (pH: 7.0)	Tryptone	1 g
	Yeast extract	1 g
	D-glucose	0.5 g
	KCl	6.34 g
	NaCl	1.2 g
	MgSO4 7H2O	0.25 g
	K2HPO4	0.13 g
	CaCl2 2H2O	0.22 g

MEDIA	COMPONENTS	AMOUNT
TYG *cont'd*	K2SO4	0.17 g
	Na2SO4	2.4 g
	NaHCO3	0.5 g
	Na2CO3	0.09 g
	Fe EDTA	0.07 g
	(Agar)	20 g
	diH2O	to 1000 mL
YEM (pH: 7)	Yeast extract	0.5 g
	Mannitol	5 g
	K2HPO4	0.5 g
	MgSO4 7H2O	0.2 g
	NaCl	0.1 g
	(Agar)	20 g g
	diH2O	to 1000 mL
R2A (pH: 7.2)	Casein acid hydrolysate	0.5 g
	Yeast extract	0.5 g
	Proteose peptone	0.5 g
	Dextrose	0.5 g
	Starch	0.5 g
	K2HPO4	0.3 g
	MgSO4	0.024 g
	Sodium pyruvate (4°C)	0.3 g
	(Agar)	15 g
	diH2O	to 1000 mL
PDA + tet (pH 5.2)	Potato Infusion Powder	4 g
	Dextrose	20 g
	Agar	15 g

MEDIA	COMPONENTS	AMOUNT
PDA + tet *cont'd*	diH2O	to 1000 mL
	+ 1 mL tetracycline (12 mg/ml) after autoclave	
MEA + tet (pH: 4.7)	Malt Extract (BD Difco)	15 g
	(Agar)	20 g
	diH2O	to 1000 mL
	+ 1 mL tetracycline (12 mg/ml) after autoclave	
	! LIGHT SENSITIVE !	
MMM (pH: 7.1)	NH4Cl	1.62 g
	MgSO4 7H2O	0.2 g
	K2HPO4	2.4 g
	NaH2PO4 2H2O	1.1 g
	Methanol	5 ml
	Na2EDTA 2H2O	15 mg
	FeSO4 7H2O	3.0 mg
	ZnSO4 7H2O	4.5 mg
	CoCl2 6H2O	3.0 mg
	MnCl2	0.64 mg
	H3BO3	1.0 mg
	Na2MoO4 2H2O	0.4 mg
	CuSO4 5H2O	0.3 mg
	CaCl2 2H2O	3.0 mg
	(Agar)	15 g
	diH2O	to 1000 mL

Table 6.1: Media Recipes for bacteria isolate collections

CHAPTER 7

APPENDIX B: HUBS FOUND IN ISOLATE COLLECTION

The following table provides the isolate names from Chapter 1 and corresponding hub names from Brachi et al. (2022) along with the identified taxonomy. Isolates are available in the Bergelson lab freezer stocks.

Isolate	Hub	Taxa
TJOR_4-1_069	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
TJOR_7-1_146	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
ULL_4-2_042	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_075	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_7-2_167	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_1-2_R1	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_1-2_R2	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_1-2_R3	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_4-1_R3	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_1-2_S3	B1	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
A_1-2_019	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_321	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_406	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_329	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_428	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
A_5-2_272	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_301	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_308	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_013	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_334	B12	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_8-1_032	B13	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
VAR_5-2_206	B13	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
VAR_5-2_308	B13	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
A_5-2_124	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_329	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
VAR_5-2_307	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_310	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_316	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_319	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_126	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_333	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_309	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
A_5-2_332	B18	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Rhizobiaceae;Pararhizobium-Rhizobium
VAR_3-1_Y15	B25	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
TJ2_10-1_012	B25	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
TJOR_4-1_036.a	B25	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
A_4-1_121	B25	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
VAR_5-2_304	B25	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
VAR_7-2_u083	B25	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA

Isolate	Hub	Taxa
VAR_5-2_029	B25	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA
TJOR_7-1_120	B26	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_5-2_145	B26	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_5-2_146	B26	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_5-2_264	B26	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
A_10-1_086	B26	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_2-1_172	B28	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Polaromonas
VAR_2-1_177	B28	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Polaromonas
VAR_5-1_106	B28	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Polaromonas
VAR_9-1_Y11	B28	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Polaromonas
VAR_5-2_100	B28	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Polaromonas
VAR_1-2_R82	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_5-1_R84	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_5-1_R88	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
ULL_9-1_R8	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_8-1_R48	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_5-2_394	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_5-2_008	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_5-2_386	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_5-2_331	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
VAR_5-2_001	B38	Bacteria;Proteobacteria;Alphaproteobacteria;Caulobacterales;Caulobacteraceae;Brevundimonas
ULL_8-2_006	B40	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Xanthobacteraceae;Tardiphaga
VAR_5-2_409	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
A_6-1_039.a	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
VAR_5-2_269	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
TJ2_7-2_015	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
A_6-1_021.a	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
A_5-2_053	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
TJ2_7-2_003	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
VAR_5-2_271	B5	Bacteria;Proteobacteria;Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Variovorax
TJOR_7-1_027	B53	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
TJOR_7-1_101	B53	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_7-2_194	B53	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_403	B53	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
TJOR_7-1_347	B53	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_376	B53	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-2_302	B53	Bacteria;Proteobacteria;Alphaproteobacteria;Sphingomonadales;Sphingomonadaceae;Sphingomonas
VAR_5-1_117	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
TJOR_7-1_113	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium

Isolate	Hub	Taxa
VAR_5-2_364	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_7-2_177	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_7-2_191	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_5-2_334	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_5-2_412	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_5-2_427	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_7-2_182	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium
VAR_5-2_274	B58	Bacteria;Proteobacteria;Alphaproteobacteria;Rhizobiales;Beijerinckiaceae;Methylobacterium

Table 7.1: Hub Isolate Candidates: Isolates matching hub 16S sequences from Brachi et al. (2022) with >99% identity match using BLAST

CHAPTER 8

APPENDIX C: OLIGOS FOR ILLUMINA SEQUENCING

8.1 Primers for Illumina 16S and *gyrB* sequencing

This section describes the oligos designed for amplification and sequencing of marker genes in bacteria. The design mostly uses the amplicon library methods described in Illumina (2013). The adapters required for sequencing are large, so the adapters are ligated in two PCR steps. PCR 1 amplifies the marker gene and incorporates a portion of the adapter. PCR2 amplifies from the adapter region of the PCR1 oligos and extends to add the remaining adapter oligos.

We differ from Illumina (2013) in that we include inline indexes on the 5' of the primer region to increase multiplexing capabilities. Inline barcodes were adapted from Bartoli et al. (2018).

primer name	orientation	inline barcode
t1 forward	F	GACTAC
t2 forward	F	CTGGTT
t3 forward	F	ACTCGA
t4 forward	F	TGCTGT
t1 reverse	R	AAGGCC
t2 reverse	R	GTCAGG
t3 reverse	R	CCTCTT
t4 reverse	R	TCGTAG

Table 8.1: PCR1 Inline barcodes. Inline barcodes used with gene specific amplification primers in Illumina library prep.

amplicon	orientation	primer name	primer seq
16S	F	16S-799F	AACMGGATTAGATACCCKG
16S	R	16S-1193R	ACGTCATCCCCACCTTCC
gyrB	F	*gyrB* F	MGNCCNGSNATGTAYATHGG
gyrB	R	*gyrB* R	ACNCCRTGNARDCCDCCNGA

Table 8.2: PCR1 amplicon primers. gene specific amplification for Illumina library preparation.

8.1.1 PCR1 oligos

Barcoded primers for 16S, *gyrB* amplification (PCR1) for Illumina Sequencing:

Forward amplification primer description:

1. 5' Illumina overlap region
2. forward inline barcode (Table 8.1)
3. gene specific forward primer (Table 8.2)

Final forward oligos PCR1 combining 5'-(1)(2)(3)-3':

TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG <inline barcode> <primer seq>

Reverse amplification primer description:

1. 3' Illumina overlap region
2. inline forward barcode (Table 8.1)
3. gene specific reverse primer (Table 8.2)

Final reverse oligos PCR1 combining 5'-(1)(2)(3)-3':

GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG <inline barcode> <primer seq>

8.1.2 PCR2 oligos

Primers for library indexing (PCR2) on the Illumina Platform:

P5 (forward) indexing primer description:

1. 5' Illumina adapter
2. i5 index (Table 8.3)
3. PCR 1 overlap region

Final reverse oligos PCR1 combining 5'-(1)(2)(3)-3':

AATGATACGGCGACCACCGAGATCTACAC <i5 index> TCGTCGGCAGCGTC

P7 (reverse) indexing primer description:

1. 3' Illumina adapter
2. i7 index (Table 8.3)
3. PCR 1 overlap region

Final reverse oligos PCR1 combining 5'-(1)(2)(3)-3':

CAAGCAGAAGACGGCATACGAGAT <i7 index> GTCTCGTGGGCTCGG

i5 name	i5 index	i7 name	i7 index
SA501	ATCGTACG	NA701	AACTCTCG
SA502	ACTATCTG	NA702	ACTATGTC
SA503	TAGCGAGT	NA703	AGTAGCGT
SA504	CTGCGTGT	NA704	CAGTGAGT
SA505	TCATCGAG	NA705	CGTACTCA
SA506	CGTGAGTG	NA706	CTACGCAG
SA507	GGATATCT	NA707	GGAGACTA
SA508	GACACCGT	NA708	GTCGCTCG
		NA709	GTCGTAGT
SB501	CTACTATA	NA710	TAGCAGAC
SB502	CGTTACTA	NA711	TCATAGAC
SB503	AGAGTCAC	NA712	TCGCTATA
SB504	TACGAGAC		
SB505	ACGTCTCG	NB701	AAGTCGAG
SB506	TCGACGAG	NB702	ATACTTCG
SB507	GATCGTGT	NB703	AGCTGCTA
SB508	GTCAGATA	NB704	CATAGAGA
		NB705	CGTAGATC
SC501	ACGACGTG	NB706	CTCGTTAC
SC502	ATATACAC	NB707	GCGCACGT
SC503	CGTCGCTA	NB708	GGTACTAT
SC504	CTAGAGCT	NB709	GTATACGC
SC505	GCTCTAGT	NB710	TACGAGCA
SC506	GACACTGA	NB711	TCAGCGTT
SC507	TGCGTACG	NB712	TCGCTACG
SC508	TAGTGTAG		
SD501	AAGCAGCA		
SD502	ACGCGTGA		
SD503	CGATCTAC		
SD504	TGCGTCAC		
SD505	GTCTAGTG		
SD506	CTAGTATG		
SD507	GATAGCGT		
SD508	TCTACACT		

Table 8.3: Illumina indices used in library amplification.

CHAPTER 9

APPENDIX D: HOST GENE CANDIDATES DRIVING HERITABILITY OF MICROBES

Here we provide *Arabidopsis thaliana* gene candidates identified through Genome Wide Association studies using the AIV model (see Chapter 4).

	amplicon	Locus Identifier	Gene Model Name	Gene Model Description	Gene Model Type
1	16S	AT1G27150	AT1G27150.1	Tetratricopeptide repeat (TPR)-like superfamily protein;(source:Araport11)	protein_coding
2	16S	AT1G27160	AT1G27160.1	valyl-tRNA synthetase / valine-tRNA ligase-like protein;(source:Araport11)	protein_coding
3	16S	AT1G31320	AT1G31320.1	LOB domain-containing protein 4;(source:Araport11)	protein_coding
4	16S	AT1G31330	AT1G31330.1	Encodes subunit F of photosystem I.	protein_coding
5	16S	AT1G50620	AT1G50620.1	A specific subunit of MINU1/2-associated SWI/SNF (MAS) complexes (PMID:36189880 and PMID:36471048).	protein_coding
6	16S	AT1G51940	AT1G51940.1	Encodes a LysM-containing receptor-like kinase. Induction of chitin-responsive genes by chitin treatment is not blocked in the mutant. Based on protein sequence alignment analysis, it has a typical RD signaling domain in its catalytic loop and possesses autophosphorylation activity.It is required for the suppression of defense responses in absence of pathogen infection or upon abscisic acid treatment. Loss-of-function mutants display enhanced resistance to Botrytis cinerea and Pectobacterium carotovorum. Its expression is repressed by pathogen infection and biological elicitors and is induced abscisic acid.Expression is strongly repressed by elicitors and fungal infection, and is induced by the hormone abscisic acid (ABA). Insertional mutants show increased expression of PHYTOALEXIN-DEFICIENT 3 (PAD3), enhanced resistance to Botrytis cinerea and Pectobacterium carotovorum infection and reduced physiological responses to ABA, suggesting that LYK3 is important for the cross-talk between signaling pathways activated by ABA and pathogens (PMID:24639336).	protein_coding
7	16S	AT1G56490	AT1G56490.1	pseudogene of Pectin lyase-like superfamily protein;(source:Araport11)	pseudogene
8	16S	AT1G60270	AT1G60270.1	beta glucosidase 6;(source:Araport11)	protein_coding
9	16S	AT1G79280	AT1G79280.2	Encodes a 237-kDA protein with similarity to vertebrate Tpr, a long coiled-coil proteins of nuclear pore inner basket filaments. It is localized to the inner surface of the nuclear envelope and is a component of the nuclear pore-associated steps of sumoylation and mRNA export in plants. Mutations affect flowering time regulation and other developmental processes. Probably acts in the same pathway as ESD4 in affecting flowering time, vegetative and inflorescence development.	protein_coding
10	16S	AT1G79310	AT1G79310.1	Encodes a putative metacaspase. Arabidopsis contains three type I MCP genes (MCP1a-c) and six type II MCP genes (MCP2a?f): AtMCP1a/At5g64240, AtMCP1b/At1g02170, AtMCP1c/At4g25110, AtMCP2a/At1g79310, AtMCP2b/At1g79330, AtMCP2c/At1g79320, AtMCP2d/At1g79340, AtMCP2e/At1g16420, AtMCP2f/At5g04200.	protein_coding
11	16S	AT2G01430	AT2G01430.1	ATHB17 is a member of the HD-Zip transcription factor family. It is expressed most strongly in roots at different stages of development and induced by ABA, paraquat, drought, and NaCl treatments. Loss of function mutants are more sensitive to salt and drought stress.The protein is nuclear localized and has been shown to bind to the promoter of SIG5 and other genes.	protein_coding
12	16S	AT2G44540	AT2G44540.1	glycosyl hydrolase 9B9;(source:Araport11)	protein_coding
13	16S	AT3G04240	AT3G04240.1	Protein O-GlcNAc transferase. Together with SPY functions to competitively regulate RGA1 (At2g01570).	protein_coding
14	16S	AT3G26618	AT3G26618.1	eukaryotic release factor 1-3;(source:Araport11)	protein_coding
15	16S	AT3G26920	AT3G26920.1	FBD / Leucine Rich Repeat domains containing protein;(source:Araport11)	protein_coding
16	16S	AT3G29187	AT3G29187.1	transposable_element_gene;(source:Araport11);non-LTR retrotransposon family (LINE), has a 9.2e-08 P-value blast match to GB:NP_038602 L1 repeat, Tf subfamily, member 18 (LINE-element) (Mus musculus);(source:TAIR10)	transposable_element_gene
17	16S	AT3G45775	AT3G45775.1	transposable_element_gene;(source:Araport11);copia-like retrotransposon family, has a 9.5e-189 P-value blast match to GB:AAA57005 Hopscotch polyprotein (Ty1_Copia-element) (Zea mays);(source:TAIR10)	transposable_element_gene
18	16S	AT4G08500	AT4G08500.1	Encodes a member of the A1 subgroup of the MEKK (MAPK/ERK kinase kinase) family. MEKK is another name for Mitogen-Activated Protein Kinase Kinase Kinase (MAPKKK or MAP3K). This subgroup has four members: At4g08500 (MEKK1, also known as ARAKIN, MAP3Kb1, MAPKKK8), At4g08480 (MEKK2, also known as MAP3Kb4, MAPKKK9), At4g08470 (MEKK3, also known as MAP3Kb3, MAPKKK10) and At4g12020 (MEKK4, also known as MAP3Kb5, MAPKKK11, WRKY19). Nomenclatures for mitogen-activated protein kinases are described in Trends in Plant Science 2002, 7(7):301. Mediates cold, salt, cadmium and wounding stress signalling. Phosphorylates MEK1.	protein_coding
19	16S	AT4G08690	AT4G08690.1	Sec14p-like phosphatidylinositol transfer family protein;(source:Araport11)	protein_coding
20	16S	AT4G13840	AT4G13840.1	HXXXD-type acyl-transferase family protein;(source:Araport11)	protein_coding
21	16S	AT4G14165	AT4G14165.1	F-box family protein-like protein;(source:Araport11)	protein_coding
22	16S	AT4G17230	AT4G17230.1	Encodes a scarecrow-like protein (SCL13). Member of GRAS gene family. Regulated by heat shock.	protein_coding
23	16S	AT4G30100	AT4G30100.1	P-loop containing nucleoside triphosphate hydrolases superfamily protein;(source:Araport11)	protein_coding
24	16S	AT5G01240	AT5G01240.1	Encodes LAX1 (LIKE AUXIN RESISTANT), a member of the AUX1 LAX family of auxin influx carriers. Required for the establishment of embryonic root cell organization.	protein_coding
25	16S	AT5G01260	AT5G01260.2	Carbohydrate-binding-like fold;(source:Araport11)	protein_coding
26	16S	AT5G01270	AT5G01270.2	Encodes CPL2, a carboxyl-terminal domain (CTD) phosphatase that dephosphorylates CTD Ser5-PO4 of the RNA polymerase II complex. Regulates plant growth, stress and auxin responses.	protein_coding
27	16S	AT5G01280	AT5G01280.1	Encodes a microtubule-associated protein.	protein_coding
28	16S	AT5G01290	AT5G01290.1	mRNA capping enzyme family protein;(source:Araport11)	protein_coding
29	16S	AT5G09660	AT5G09660.4	encodes a microbody NAD-dependent malate dehydrogenase encodes an peroxisomal NAD-malate dehydrogenase that is involved in fatty acid beta-oxidation through providing NAD to the process of converting fatty acyl CoA to acetyl CoA.	protein_coding
30	16S	AT5G19460	AT5G19460.1	nudix hydrolase homolog 20;(source:Araport11)	protein_coding
31	16S	AT5G35410	AT5G35410.1	encodes a member of the CBL-interacting protein kinase family, is a regulatory component controlling plant potassium nutrition	protein_coding
32	16S	AT5G59250	AT5G59250.1	Encodes a chloroplast localized H+/glucose antiporter.	protein_coding
33	gyrB	AT1G11360	AT1G11360.1	Adenine nucleotide alpha hydrolases-like superfamily protein;(source:Araport11)	protein_coding
34	gyrB	AT1G11410	AT1G11410.4	S-locus lectin protein kinase family protein;(source:Araport11)	protein_coding
35	gyrB	AT1G11420	AT1G11420.1	Member of the plant-specific DUF724 protein family. Arabidopsis has 10 DUF724 proteins.	protein_coding
36	gyrB	AT1G20130	AT1G20130.1	GDSL-motif esterase/acyltransferase/lipase. Enzyme group with broad substrate specificity that may catalyze acyltransfer or hydrolase reactions with lipid and non-lipid substrates.	protein_coding
37	gyrB	AT1G20240	AT1G20240.1	SWI-SNF-related chromatin binding protein;(source:Araport11)	protein_coding
38	gyrB	AT1G21080	AT1G21080.3	DNAJ heat shock N-terminal domain-containing protein;(source:Araport11)	protein_coding
39	gyrB	AT1G21890	AT1G21890.1	nodulin MtN21-like transporter family protein	protein_coding
40	gyrB	AT1G29840	AT1G29840.1	alpha/beta-Hydrolases superfamily protein;(source:Araport11)	protein_coding
41	gyrB	AT1G30790	AT1G30790.1	F-box and associated interaction domains-containing protein;(source:Araport11)	protein_coding
42	gyrB	AT1G30795	AT1G30795.1	Glycine-rich protein family;(source:Araport11)	protein_coding
43	gyrB	AT1G30860	AT1G30860.1	RING/U-box superfamily protein;(source:Araport11)	protein_coding
44	gyrB	AT1G30900	AT1G30900.1	VACUOLAR SORTING RECEPTOR 6;(source:Araport11)	protein_coding
45	gyrB	AT1G31540	AT1G31540.2	Disease resistance protein (TIR-NBS-LRR class) family;(source:Araport11)	protein_coding
46	gyrB	AT1G34044	AT1G34044.1	pseudogene of 50S ribosomal protein L34;(source:Araport11)	pseudogene
47	gyrB	AT1G34060	AT1G34060.1	Pyridoxal phosphate (PLP)-dependent transferases superfamily protein;(source:Araport11)	protein_coding
48	gyrB	AT1G34065	AT1G34065.1	S-adenosylmethionine carrier 2;(source:Araport11)	protein_coding
49	gyrB	AT1G34110	AT1G34110.1	Leucine-rich receptor-like protein kinase family protein;(source:Araport11)	protein_coding

50	gyrB	AT1G34130	AT1G34130.1	Encodes homolog of yeast STT3, a subunit of oligosaccharyltransferase.	protein_coding
51	gyrB	AT1G34140	AT1G34140.1	polyadenylate-binding protein, putative / PABP, putative, non-consensus splice donor TA at exon 1; similar to polyadenylate-binding protein (poly(A)-binding protein) from (Triticum aestivum) GI:1737492, (Nicotiana tabacum) GI:7673355, {Arabidopsis thaliana} SP:P42731; contains InterPro entry IPR000504: RNA-binding region RNP-1 (RNA recognition motif) (RRM). Only member of the class IV PABP family.	protein_coding
52	gyrB	AT1G35750	AT1G35750.1	Encodes a member of the Arabidopsis Pumilio (APUM) proteins containing PUF domain (eight repeats of approximately 36 amino acids each). PUF proteins regulate both mRNA stability and translation through sequence-specific binding to the 3' UTR of target mRNA transcripts.	protein_coding
53	gyrB	AT1G35840	AT1G35840.1	transposable_element_gene;(source:Araport11);copia-like retrotransposon family, has a 0. P-value blast match to dbj|BAA78426.1| polyprotein (AtRE2-1) (Arabidopsis thaliana) (Ty1_Copia-element);(source:TAIR10)	transposable_element_gene
54	gyrB	AT1G49840	AT1G49840.1	glutamyl-tRNA (Gln) amidotransferase subunit A (DUF620);(source:Araport11)	protein_coding
55	gyrB	AT1G56430	AT1G56430.1	Encodes a protein with nicotianamine synthase activity.	protein_coding
56	gyrB	AT1G56450	AT1G56450.1	20S proteasome beta subunit PBG1 (PBG1) mRNA, complete cds	protein_coding
57	gyrB	AT1G60380	AT1G60380.1	NAC (No Apical Meristem) domain transcriptional regulator superfamily protein;(source:Araport11)	protein_coding
58	gyrB	AT1G62810	AT1G62810.1	Encodes COPPER AMINE OXIDASE1 (CuAO1). Contributes to abscisic acid- and polyamine-induced nitric oxide biosynthesis and abscisic acid signal transduction.	protein_coding
59	gyrB	AT1G64060	AT1G64060.1	Interacts with AtrbohD gene to fine tune the spatial control of ROI production and hypersensitive response to cell in and around infection site.	protein_coding
60	gyrB	AT1G65540	AT1G65540.1	LETM1-like protein;(source:Araport11)	protein_coding
61	gyrB	AT1G68600	AT1G68600.1	aluminum activated malate transporter family protein;(source:Araport11)	protein_coding
62	gyrB	AT1G76420	AT1G76420.1	Identified in an enhancer trap line; member of the NAC family of proteins. Expressed at the boundary between the shoot meristem and lateral organs and the polar nuclei in the embryo sac. Together with CUC2-DA1-UBP15 part of a regulatory module which controls the initiation of axillary meristems, thereby determining plant architecture. Regulates axillary meristem initiation by directly binding to the DA1 promoter.	protein_coding
63	gyrB	AT1G77210	AT1G77210.1	AtSTP14 belongs to the family of sugar transport proteins (AtSTPs)in volved in monosaccharide transport. Heterologous expression in yeast revealed that AtSTP14 is the transporter specifc for galactose and does not transport other monosaccharides such as glucose or fructose.	protein_coding
64	gyrB	AT1G77230	AT1G77230.1	Tetratricopeptide repeat (TPR)-like superfamily protein;(source:Araport11)	protein_coding
65	gyrB	AT1G77240	AT1G77240.1	AMP-dependent synthetase and ligase family protein;(source:Araport11)	protein_coding
66	gyrB	AT1G77250	AT1G77250.1	RING/FYVE/PHD-type zinc finger family protein;(source:Araport11)	protein_coding
67	gyrB	AT2G07050	AT2G07050.1	Involved in the biosynthesis of brassinosteroids. Catalyzes the reaction from epoxysqualene to cycloartenol.	protein_coding
68	gyrB	AT2G14700	AT2G14700.1	hypothetical protein;(source:Araport11)	protein_coding
69	gyrB	AT2G26360	AT2G26360.1	Mitochondrial substrate carrier family protein;(source:Araport11)	protein_coding
70	gyrB	AT2G42200	AT2G42200.1	Encodes a putative transcriptional regulator that is involved in the vegetative to reproductive phase transition. Expression is regulated by MIR156b. SPL activity nonautonomously inhibits initiation of new leaves at the shoot apical meristem. Overexpression of SPL9 (rSPL9) promoted the expression of C-REPEAT BINDING FACTOR 2 (CBF2) and hereafter enhanced the freezing tolerance.	protein_coding
71	gyrB	AT2G42365	AT2G42365.1	Natural antisense transcript overlaps with AT2G42360 and AT2G42370;(source:Araport11)	antisense_long_noncoding_rna
72	gyrB	AT2G42370	AT2G42370.1	hypothetical protein;(source:Araport11)	protein_coding
73	gyrB	AT2G42380	AT2G42380.2	Encodes a member of the BZIP family of transcription factors. Forms heterodimers with the related protein AtbZIP61. Binds to G-boxes in vitro and is localized to the nucleus in onion epidermal cells.	protein_coding
74	gyrB	AT2G42388	AT2G42388.1	other_RNA;(source:Araport11)	other_rna
75	gyrB	AT2G42390	AT2G42390.1	kinase C substrate, heavy chain-like protein;(source:Araport11)	protein_coding
76	gyrB	AT2G43070	AT2G43070.4	SIGNAL PEPTIDE PEPTIDASE-LIKE 3;(source:Araport11)	protein_coding
77	gyrB	AT2G43080	AT2G43080.1	Encodes a prolyl-4 hydroxylase that can hydroxylate poly(L-proline),the collagen model peptide (Pro-Pro-Gly)10 and other proline rich peptides.	protein_coding
78	gyrB	AT2G43100	AT2G43100.1	Small subunit, which together with IPMI SSU2, IPMI SSU3 and IPMI LSU1, is a member of heterodimeric isopropylmalate isomerase (IPMI). Together with IPMI SSU3 participates in the Met chain elongation pathway.	protein_coding
79	gyrB	AT2G43110	AT2G43110.1	U3 containing 90S pre-ribosomal complex subunit;(source:Araport11)	protein_coding
80	gyrB	AT2G44290	AT2G44290.1	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein;(source:Araport11)	protein_coding
81	gyrB	AT3G01420	AT3G01420.1	Encodes an alpha-dioxygenase involved in protection against oxidative stress and cell death. Induced in response to Salicylic acid and oxidative stress. Independent of NPR1 in induction by salicylic acid. The mRNA is cell-to-cell mobile.	protein_coding
82	gyrB	AT3G05400	AT3G05400.1	Major facilitator superfamily protein;(source:Araport11)	protein_coding
83	gyrB	AT3G06500	AT3G06500.1	Encodes an alkaline/neutral invertase which localizes in mitochondria. It may be modulating hormone balance in relation to the radicle emergence. Mutants display severely reduced shoot growth and reduced oxygen consumption. Mutant root development is not affected as reported for A/N-InvA mutant (inva) plants. The mRNA is cell-to-cell mobile.	protein_coding
84	gyrB	AT3G06510	AT3G06510.2	Encodes a protein with beta-glucosidase and galactosyltransferase activity, mutants show increased sensitivity to freezing. Though it is classified as a family I glycosyl hydrolase, it has no hydrolase activity in vitro.	protein_coding
85	gyrB	AT3G06530	AT3G06530.4	ARM repeat superfamily protein;(source:Araport11)	protein_coding
86	gyrB	AT3G06540	AT3G06540.1	Encodes a cytoplasmic Rab escort protein that preferentially binds the GDP-bound form of Rab and stimulates geranylgeranylation of various Rab GTPases in Arabidopsis extracts in vitro.	protein_coding
87	gyrB	AT3G06545	AT3G06545.1	transmembrane protein;(source:Araport11)	protein_coding
88	gyrB	AT3G06550	AT3G06550.2	Encodes a homolog of the protein Cas1p known to be involved in polysaccharide O-acetylation in Cryptococcus neoformans. Mutants show reduced cell wall polysaccharide acetylation and increased resistance to Botrytis cinerea. The protein is expressed in the Golgi and is involved in the acetylation of xylan during secondary wall biosynthesis.	protein_coding
89	gyrB	AT3G06560	AT3G06560.1	Encodes a poly(A) polymerase. Located in the cytoplasm.	protein_coding
90	gyrB	AT3G06570	AT3G06570.1	Galactose oxidase/kelch repeat superfamily protein;(source:Araport11)	protein_coding
91	gyrB	AT3G06580	AT3G06580.1	Encodes a protein with galactose kinase activity. The gene was shown to complement the yeast Agal1 mutant defective in the galactokinase gene GAL1.	protein_coding
92	gyrB	AT3G11540	AT3G11540.1	Contains a tetratricopeptide repeat region, and a novel carboxy-terminal region. SPY acts as both a repressor of GA responses and as a positive regulation of cytokinin signalling. SPY may be involved in reducing ROS accumulation in response to stress. Regulates root hair patterning independently of 2 gibberellin signalling. Together with SEC functions to competitively regulate RGA1 (At2g01570). Negative regulator of trichome branching.	protein_coding

93	gyrB	AT3G11580	AT3G11580.3	SOD7 encodes nuclear localized B3 DNA binding domain and a transcriptional repression motif. Belongs to the RAV gene family. Functions in regulation of seed size and binds to and represses KLU. Transcription repressor involved in regulation of inflorescence architecture. Required for axillary meristem formation and acts by repression of CUC2/CUC3.Based on expression patterns, it is not required for stem cell specification during embryo shoot apical meristem initiation.	protein_coding
94	gyrB	AT3G11590	AT3G11590.1	golgin family A protein;(source:Araport11)	protein_coding
95	gyrB	AT3G11591	AT3G11591.1	bric-a-brac protein;(source:Araport11)	protein_coding
96	gyrB	AT3G11600	AT3G11600.1	One of two plant specific paralogs of unknown function. Interacts with GL2. GIR1/GIR2 loss of function resembles gl2 lof mutations.	protein_coding
97	gyrB	AT3G11850	AT3G11850.1	myosin-binding protein (Protein of unknown function, DUF593);(source:Araport11)	protein_coding
98	gyrB	AT3G13380	AT3G13380.1	Similar to BRI, brassinosteroid receptor protein.	protein_coding
99	gyrB	AT3G24730	AT3G24730.1	mRNA splicing factor, thioredoxin-like U5 snRNP;(source:Araport11)	protein_coding
100	gyrB	AT3G26614	AT3G26614.1	transposable_element_gene;(source:Araport11);non-LTR retrotransposon family (LINE), has a 3.0e-39 P-value blast match to GB:NP_038607 L1 repeat, Tf subfamily, member 9 (LINE-element) (Mus musculus);(source:TAIR10)	transposable_element_gene
101	gyrB	AT3G28880	AT3G28880.1	serine/threonine-protein phosphatase 6 regulatory ankyrin repeat subunit;(source:Araport11)	protein_coding
102	gyrB	AT3G43190	AT3G43190.1	Encodes a protein with sucrose synthase activity (SUS4).	protein_coding
103	gyrB	AT3G43200	AT3G43200.1	pseudogene of target of trans acting-siR480/255 protein;(source:Araport11)	pseudogene
104	gyrB	AT3G43205	AT3G43205.1	transposable_element_gene;(source:Araport11);copia-like retrotransposon family, has a 9.5e-21 P-value blast match to GB:AAD12998 pol polyprotein (Ty1_Copia-element) (Zea mays);(source:TAIR10)	transposable_element_gene
105	gyrB	AT3G43210	AT3G43210.1	Encodes a kinesin TETRASPORE. Required for cytokinesis in pollen. In mutants, all four microspore nuclei remain within the same cytoplasm after meiosis.	protein_coding
106	gyrB	AT3G43220	AT3G43220.1	Phosphoinositide phosphatase family protein;(source:Araport11)	protein_coding
107	gyrB	AT3G43230	AT3G43230.1	FYVE domain-containing protein; autophagy adaptor that directly interacts with the autophagosome marker ATG8 and localizes on both membranes of the autophagosome.	protein_coding
108	gyrB	AT3G43240	AT3G43240.1	Interacts with CHR11, CHR17, and RTL1, several known subunits of ISWI. JA biosynthesisis is positively regulated by this chromatin remodeling complex, thereby promoting stamen filament elongation.	protein_coding
109	gyrB	AT3G43250	AT3G43250.1	coiled-coil protein (DUF572);(source:Araport11)	protein_coding
110	gyrB	AT3G43260	AT3G43260.1	deoxyhypusine protein;(source:Araport11)	protein_coding
111	gyrB	AT3G47210	AT3G47210.1	hypothetical protein (DUF247);(source:Araport11)	protein_coding
112	gyrB	AT3G47500	AT3G47500.1	Dof-type zinc finger domain-containing protein, identical to H-protein promoter binding factor-2a GI:3386546 from (Arabidopsis thaliana). Interacts with LKP2 and FKF1, but its overexpression does not change flowering time under short or long day conditions.	protein_coding
113	gyrB	AT3G48250	AT3G48250.1	Encodes a pentatricopeptide repeat protein implicated in splicing of intron 1 of mitochondrial nad7 transcripts.	protein_coding
114	gyrB	AT3G48430	AT3G48430.1	Relative of Early Flowering 6 (REF6) encodes a Jumonji N/C and zinc finger domain-containing protein that acts as a positive regulator of flowering in an FLC-dependent pathway. REF6 mutants have hyperacetylation of histone H4 at the FLC locus. REF6 interacts with BES1 in a Y2H assay and in vitro. REF6 may play a role in brassinoteroid signaling by affecting histone methylation in the promoters of BR-responsive genes. It is most closely related to the JHDM3 subfamily of JmjN/C proteins. The mRNA is cell-to-cell mobile.	protein_coding
115	gyrB	AT3G48440	AT3G48440.1	Zinc finger C-x8-C-x5-C-x3-H type family protein;(source:Araport11)	protein_coding
116	gyrB	AT3G52870	AT3G52870.1	IQ calmodulin-binding motif family protein;(source:Araport11)	protein_coding
117	gyrB	AT3G52880	AT3G52880.2	Encodes a peroxisomal monodehydroascorbate reductase, involved in the ascorbate-glutathione cycle which removes toxic H2O2	protein_coding
118	gyrB	AT3G52890	AT3G52890.1	KCBP-interacting protein kinase interacts specifically with the tail region of KCBP	protein_coding
119	gyrB	AT3G52905	AT3G52905.1	Polynucleotidyl transferase, ribonuclease H-like superfamily protein;(source:Araport11)	protein_coding
120	gyrB	AT3G55190	AT3G55190.1	alpha/beta-Hydrolases superfamily protein;(source:Araport11)	protein_coding
121	gyrB	AT3G55380	AT3G55380.2	May function together with UBC7 and UBC13 in the plant READ pathway, required in plant responses to multiple stress conditions.	protein_coding
122	gyrB	AT3G55510	AT3G55510.1	Encodes a regulator of floral determinacy in that interacts with both nucleolar and nucleoplasmic proteins.	protein_coding
123	gyrB	AT3G60240	AT3G60240.4	protein synthesis initiation factor 4G (EIF4G). A mutation in this gene (cum2-1) results in decreased accumulation of CMV coat protein in upper, uninoculated leaves. Likely affects cell-to-cell movement of the virus, also affects TCV multiplication.	protein_coding
124	gyrB	AT3G60270	AT3G60270.1	Cupredoxin superfamily protein;(source:Araport11)	protein_coding
125	gyrB	AT3G60290	AT3G60290.1	2-oxoglutarate (2OG) and Fe(II)-dependent oxygenase superfamily protein;(source:Araport11)	protein_coding
126	gyrB	AT3G60300	AT3G60300.2	RWD domain-containing protein;(source:Araport11)	protein_coding
127	gyrB	AT3G60310	AT3G60310.1	Component of Fanconi anemia (FA) complex.	protein_coding
128	gyrB	AT3G60330	AT3G60330.1	H[+]-ATPase 7;(source:Araport11)	protein_coding
129	gyrB	AT3G60340	AT3G60340.1	alpha/beta-Hydrolases superfamily protein;(source:Araport11)	protein_coding
130	gyrB	AT3G60350	AT3G60350.1	ARABIDILLO-2 and its homolog, ARABIDILLO -1, are unique among Arabidopsis Arm-repeat proteins in having an F-box motif and fall into a phylogenetically distinct subgroup from other plant Arm-repeat proteins Similar to arm repeat protein in rice and armadillo/beta-catenin repeat family protein / F-box family protein in Dictyostelium. ARABIDILLO-2 promote lateral root development. Mutant plants form fewer lateral roots, while ARABIDILLO-2-overexpressing lines produce more lateral roots than wild-type seedlings.	protein_coding
131	gyrB	AT3G60360	AT3G60360.1	embryo sac development arrest 14;(source:Araport11)	protein_coding
132	gyrB	AT4G00695	AT4G00695.3	Spc97/Spc98 family of spindle pole body (SBP) component;(source:Araport11)	protein_coding
133	gyrB	AT4G08620	AT4G08620.1	Encodes a high-af?nity sulfate transporter. Contains STAS domain. Expressed in roots and guard cells. Up-regulated by sulfur deficiency.	protein_coding
134	gyrB	AT4G08630	AT4G08630.1	fas-binding factor-like protein;(source:Araport11)	protein_coding
135	gyrB	AT4G08650	AT4G08650.1	transposable_element_gene;(source:Araport11);hypothetical protein;(source:TAIR10)	transposable_element_gene
136	gyrB	AT4G08660	AT4G08660.1	transposable_element_gene;(source:Araport11);Mutator-like transposase family, has a 3.0e-74 P-value blast match to O22273 / 233-373 Pfam PF03108 MuDR family transposase (MuDr-element domain);(source:TAIR10)	transposable_element_gene
137	gyrB	AT4G16710	AT4G16710.1	glycosyltransferase family protein 28;(source:Araport11)	protein_coding
138	gyrB	AT4G17700	AT4G17700.1	hypothetical protein;(source:Araport11)	protein_coding
139	gyrB	AT4G17710	AT4G17710.1	Encodes a homeobox-leucine zipper family protein belonging to the HD-ZIP IV family.	protein_coding
140	gyrB	AT4G17713	AT4G17713.1	Encodes a defensin-like (DEFL) family protein.	protein_coding
141	gyrB	AT4G17718	AT4G17718.1	Encodes a defensin-like (DEFL) family protein.	protein_coding
142	gyrB	AT4G18440	AT4G18440.1	L-Aspartase-like family protein;(source:Araport11)	protein_coding
143	gyrB	AT4G18460	AT4G18460.1	D-Tyr-tRNA(Tyr) deacylase family protein;(source:Araport11)	protein_coding
144	gyrB	AT4G19035	AT4G19035.1	Encodes a member of a family of small,secreted, cysteine rich protein with sequence similarity to the PCP (pollen coat protein) gene family.	protein_coding
145	gyrB	AT4G24550	AT4G24550.2	Encodes a component of the AP4 complex and is involved in vacuolar sorting of storage proteins.	protein_coding

146	gyrB	AT4G24560	AT4G24560.1	Encodes a ubiquitin-specific protease. There is no evidence for a phenotype in ubp16-1 mutants, however, double mutant analysis with ubp15 mutants reveals a role for UBP16 in plant development and cell proliferation.	protein_coding
147	gyrB	AT4G24610	AT4G24610.2	pesticidal crystal cry8Ba protein;(source:Araport11)	protein_coding
148	gyrB	AT4G26555	AT4G26555.1	Encodes a chloroplast lumen-targeted immunophilin that plays a role in the acclimation of plants under photosynthetic stress conditions, probably by regulating PsaL stability.	protein_coding
149	gyrB	AT4G26560	AT4G26560.1	Encodes calcineurin B-like protein 7 (CBL7).Interacts with and modulates the activity of the PM ATPase AHA2.	protein_coding
150	gyrB	AT4G26570	AT4G26570.2	member of AtCBLs (Calcineurin B-like Calcium Sensor Proteins); interacts with CIPK3/9/23/26 resulting in phosphorylation of their downstream targets, influencing the maintenance of cellular magnesium homeostasis.	protein_coding
151	gyrB	AT4G35010	AT4G35010.1	putative beta-galactosidase (BGAL11 gene)	protein_coding
152	gyrB	AT4G35020	AT4G35020.1	A member of ROP GTPase gene family; Encodes a Rho-like GTP binding protein.	protein_coding
153	gyrB	AT5G05560	AT5G05560.2	Encodes a subunit of the Arabidopsis thaliana E3 ubiquitin ligase complex that plays a synergistic role with APC4 both in female gametogenesis and in embryogenesis.	protein_coding
154	gyrB	AT5G09980	AT5G09980.1	elicitor peptide 4 precursor;(source:Araport11)	protein_coding
155	gyrB	AT5G09995	AT5G09995.2	transmembrane protein;(source:Araport11)	protein_coding
156	gyrB	AT5G13820	AT5G13820.1	Encodes a protein that specifically binds plant telomeric DNA repeats. It has a single Myb telomeric DNA-binding (SANT) domain in C-terminus that prefers the sequence TTTAGGG. Single Myb Histone (SMH) gene family member.	protein_coding
157	gyrB	AT5G15050	AT5G15050.1	Encodes GlcAT14B. Has glucuronosyltransferase activity adding glucuronic acid residues to beta-1,3- and beta-1,6-linked galactans.	protein_coding
158	gyrB	AT5G16180	AT5G16180.1	Promotes the splicing of chloroplast group II introns. Splices atpF introns.	protein_coding
159	gyrB	AT5G16840	AT5G16840.2	Binds to ACD11 and fungal elicitor RxLR207. Regulates ROS mediated defense response.	protein_coding
160	gyrB	AT5G16850	AT5G16850.1	Encodes the catalytic subunit of telomerase reverse transcriptase. Involved in telomere homeostasis. Homozygous double mutants with ATR show gross morphological defects over a period of generations. TERT shows Class II telomerase activity in vitro, indicating that it can initiate de novo telomerase synthesis on non-telomeric DNA, often using a preferred position within the telomerase-bound RNA. Loss of function mutants have reduced telomere length in roots and over a period of generations, decreasing root meristem function.	protein_coding
161	gyrB	AT5G16860	AT5G16860.1	Tetratricopeptide repeat (TPR)-like superfamily protein;(source:Araport11)	protein_coding
162	gyrB	AT5G16870	AT5G16870.1	Peptidyl-tRNA hydrolase II (PTH2) family protein;(source:Araport11)	protein_coding
163	gyrB	AT5G16900	AT5G16900.1	Leucine-rich repeat protein kinase family protein;(source:Araport11)	protein_coding
164	gyrB	AT5G22060	AT5G22060.1	Co-chaperonin similar to E. coli DnaJ	protein_coding
165	gyrB	AT5G25060	AT5G25060.1	Part of SWAP1-SFPS-RRC1 splicing factor complex which modulates pre-mRNA splicing to promote photomorphogenesis.	protein_coding
166	gyrB	AT5G28892	AT5G28892.1	transposable_element_gene;(source:Araport11);pseudogene, similar to putative helicase, blastp match of 40%25 identity and 7.3e-142 P-value to GP|14140296|gb|AAK54302.1|AC034258_20|AC034258 putative helicase {Oryza sativa (japonica cultivar-group)};(source:TAIR10)	transposable_element_gene
167	gyrB	AT5G29020	AT5G29020.1	transposable_element_gene;(source:Araport11);similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G60930.1);(source:TAIR10)	transposable_element_gene
168	gyrB	AT5G30276	AT5G30276.1	transposable_element_gene;(source:Araport11);gypsy-like retrotransposon family (Athila), has a 3.0e-99 P-value blast match to gb|AAL06419.1|AF378075_1 reverse transcriptase (Athila4) (Arabidopsis thaliana) (Gypsy_Ty3-family);(source:TAIR10)	transposable_element_gene
169	gyrB	AT5G35111	AT5G35111.1	pseudogene of Peroxidase superfamily protein;(source:Araport11)	pseudogene
170	gyrB	AT5G35130	AT5G35130.1	transposable_element_gene;(source:Araport11);pseudogene, similar to putative helicase, predicted proteins from different species;(source:TAIR10)	transposable_element_gene
171	gyrB	AT5G35142	AT5G35142.1	transposable_element_gene;(source:Araport11);gypsy-like retrotransposon family (Athila), has a 1.5e-102 P-value blast match to GB:CAA57397 Athila ORF 1 (Arabidopsis thaliana);(source:TAIR10)	transposable_element_gene
172	gyrB	AT5G35145	AT5G35145.1	transposable_element_gene;(source:Araport11);pseudogene, hypothetical protein, PF03078: ATHILA ORF-1 family;(source:TAIR10)	transposable_element_gene
173	gyrB	AT5G35148	AT5G35148.1	transposable_element_gene;(source:Araport11);gypsy-like retrotransposon family (Athila), has a 2.5e-89 P-value blast match to GB:CAA57397 Athila ORF 1 (Arabidopsis thaliana);(source:TAIR10)	transposable_element_gene
174	gyrB	AT5G35150	AT5G35150.1	transposable_element_gene;(source:Araport11);CACTA-like transposase family (Ptta/En/Spm), has a 1.6e-26 P-value blast match to At5g29026.1/8-244 CACTA-like transposase family (Ptta/En/Spm) (CACTA-element) (Arabidopsis thaliana);(source:TAIR10)	transposable_element_gene
175	gyrB	AT5G35170	AT5G35170.1	adenylate kinase family protein;(source:Araport11)	protein_coding
176	gyrB	AT5G35180	AT5G35180.4	ENHANCED DISEASE RESISTANCE protein (DUF1336);(source:Araport11)	protein_coding
177	gyrB	AT5G35190	AT5G35190.2	proline-rich extensin-like family protein;(source:Araport11)	protein_coding
178	gyrB	AT5G35200	AT5G35200.1	ENTH/ANTH/VHS superfamily protein;(source:Araport11)	protein_coding
179	gyrB	AT5G35210	AT5G35210.1	Encodes a chloroplast envelope-bound plant homeodomain (PHD) transcription factor with transmembrane domains that functions in multiple retrograde signal pathways. The proteolytic cleavage of PTM occurs in response to retrograde signals and amino-terminal PTM accumulates in the nucleus, where it activates ABI4 transcription in a PHD-dependent manner associated with histone modifications.	protein_coding
180	gyrB	AT5G35240	AT5G35240.1	transposable_element_gene;(source:Araport11);pseudogene, similar to putative transposable element, blastp match of 47%25 identity and 9.3e-52 P-value to GP|13122426|dbj|BAB32907.1||AP003047 putative transposable element {Oryza sativa (japonica cultivar-group)};(source:TAIR10)	transposable_element_gene
181	gyrB	AT5G35320	AT5G35320.1	DBH-like monooxygenase;(source:Araport11)	protein_coding
182	gyrB	AT5G35330	AT5G35330.1	Protein containing methyl-CpG-binding domain.Has sequence similarity to human MBD proteins.	protein_coding
183	gyrB	AT5G35331	AT5G35331.1	transposable_element_gene;(source:Araport11);non-LTR retrotransposon family (LINE), has a 7.8e-44 P-value blast match to GB:BAA20419 reverse transcriptase (LINE-element) (Mus musculus);(source:TAIR10)	transposable_element_gene
184	gyrB	AT5G35332	AT5G35332.1	transposable_element_gene;(source:Araport11);pseudogene, hypothetical protein, similar to putative reverse transcriptase;(source:TAIR10)	transposable_element_gene
185	gyrB	AT5G35336	AT5G35336.1	transposable_element_gene;(source:Araport11);pseudogene, similar to SAE1-S9-protein, similar to putative non-LTR retroelement reverse transcriptase;(source:TAIR10)	transposable_element_gene
186	gyrB	AT5G35344	AT5G35344.1	transposable_element_gene;(source:Araport11);pseudogene, hypothetical protein;(source:TAIR10)	transposable_element_gene
187	gyrB	AT5G35348	AT5G35348.1	transposable_element_gene;(source:Araport11);pseudogene, hypothetical protein;(source:TAIR10)	transposable_element_gene
188	gyrB	AT5G35353	AT5G35353.1	transposable_element_gene;(source:Araport11);Mutator-like transposase family, has a 3.4e-44 P-value blast match to Q9SI25 /181-349 Pfam PF03108 MuDR family transposase (MuDr-element domain);(source:TAIR10)	transposable_element_gene
189	gyrB	AT5G35354	AT5G35354.1	transposable_element_gene;(source:Araport11);hAT-like transposase family (hobo/Ac/Tam3), has a 2.6e-18 P-value blast match to GB:AAD24567 transposase Tag2 (hAT-element) (Arabidopsis thaliana);(source:TAIR10)	transposable_element_gene
190	gyrB	AT5G35360	AT5G35360.3	Encodes biotin carboxylase subunit (CAC2).	protein_coding
191	gyrB	AT5G35370	AT5G35370.1	S-locus lectin protein kinase family protein;(source:Araport11)	protein_coding
192	gyrB	AT5G35375	AT5G35375.1	transmembrane protein;(source:Araport11)	protein_coding
193	gyrB	AT5G35380	AT5G35380.1	kinase with adenine nucleotide alpha hydrolases-like domain-containing protein;(source:Araport11)	protein_coding

194	gyrB	AT5G35390	AT5G35390.1	Encodes a member of the receptor-like kinase family of genes. In pollen tubes, it accumulates in the plasma membrane of the apical growing tip through the process of exocytosis.	protein_coding
195	gyrB	AT5G35400	AT5G35400.2	Enzyme for the pseudouridine (Ψ) to uridine (U) conversion.	protein_coding
196	gyrB	AT5G35405	AT5G35405.1	Encodes a ECA1 gametogenesis related family protein	protein_coding
197	gyrB	AT5G35413	AT5G35413.1	transposable_element_gene;(source:Araport11);non-LTR retrotransposon family (LINE), has a 1.6e-41 P-value blast match to GB:NP_038605 L1 repeat, Tf subfamily, member 30 (LINE-element) (Mus musculus);(source:TAIR10)	transposable_element_gene
198	gyrB	AT5G35416	AT5G35416.1	transposable_element_gene;(source:Araport11);non-LTR retrotransposon family (LINE), has a 1.2e-16 P-value blast match to GB:NP_038603 L1 repeat, Tf subfamily, member 23 (LINE-element) (Mus musculus);(source:TAIR10)	transposable_element_gene
199	gyrB	AT5G35980	AT5G35980.1	Encodes a dual specificity protein kinase which phosphorylates substract proteins on Ser/Thr and Tyr residues. Some substrates include annexin family proteins. YAK1 mutations suppress TOR deficiency in Arabidopsis and consequences of lst8 mutations. The YAK1 protein is phosphorylated by the TOR kinase.	protein_coding
200	gyrB	AT5G35995	AT5G35995.1	F-box/RNI-like superfamily protein;(source:Araport11)	protein_coding
201	gyrB	AT5G36001	AT5G36001.1	RING/U-box superfamily protein;(source:Araport11)	protein_coding
202	gyrB	AT5G36002	AT5G36002.1	Natural antisense transcript overlaps with AT5G36001.;(source:Araport11)	antisense_long_noncoding_rna
203	gyrB	AT5G36015	AT5G36015.1	transposable_element_gene;(source:Araport11);pseudogene, similar to OSJNBa0026J14.30, blastp match of 57%25 identity and 3.0e-64 P-value to GP|20146463|dbj|BAB89243.1||AP004231 OSJNBa0026J14.30 {Oryza sativa (japonica cultivar-group)};(source:TAIR10)	transposable_element_gene
204	gyrB	AT5G36080	AT5G36080.1	Myb/SANT-like DNA-binding domain protein;(source:Araport11)	protein_coding
205	gyrB	AT5G36100	AT5G36100.1	Protein of unknown function that is found on the surfaces of lipid droplets and may function to anchor the droplets to the plasma membrane.	protein_coding
206	gyrB	AT5G36110	AT5G36110.1	Encodes a member of the CYP716A subfamily of cytochrome P450 monooxygenases with triterpene oxidizing activity catalyzing C-28 hydroxylation of alpha-amyrin, beta-amyrin, and lupeol, producing uvaol, erythrodiol, and betulin, respectively. Additionally, it shows carboxylation activity for the C-28 position of alpha- and beta-amyrin.	protein_coding
207	gyrB	AT5G36140	AT5G36140.1	Encodes a member of the CYP716A subfamily of cytochrome P450 monooxygenases with triterpene oxidizing activity catalyzing C-28 hydroxylation of alpha-amyrin, beta-amyrin, and lupeol, producing uvaol, erythrodiol, and betulin, respectively.In particular, 22alpha-hydroxylation activity has been observed against alpha-amyrin. Should be merged with At5g36130.	protein_coding
208	gyrB	AT5G40595	AT5G40595.1	hypothetical protein;(source:Araport11)	protein_coding
209	gyrB	AT5G40600	AT5G40600.1	bromodomain testis-specific protein;(source:Araport11)	protein_coding
210	gyrB	AT5G42010	AT5G42010.1	Transducin/WD40 repeat-like superfamily protein;(source:Araport11)	protein_coding
211	gyrB	AT5G42220	AT5G42220.1	Ubiquitin-like superfamily protein;(source:Araport11)	protein_coding
212	gyrB	AT5G42750	AT5G42750.1	Encodes a plasma-membrane associated phosphoprotein that interacts directly with the kinase domain of BRI1 through the evolutionarily conserved C-terminal BIM motif binding to the C-lobe of the BRI1 kinase domain. It interferes with the interaction between BRI1 with its signalling partner, the plasma membrane localised LRR-receptor kinase BAK1 by inhibiting the transphosphorylation to keep BRI1 at a basal level of activity. It is phosphorylated by BRI1 at Ser270 & Ser274 and at tyrosine site Tyr211 and dissociates from plasma membrane to end up in the cytosol after phosphorylation. Its loss-of-function mutant shows higher sensitivity to BR treatment.	protein_coding
213	gyrB	AT5G42900	AT5G42900.1	Acts with COR28 as a key regulator in the COP1-HY5 regulatory hub by regulating HY5 activity to ensure proper skotomorphogenic growth in the dark and photomorphogenic development in the light.	protein_coding
214	gyrB	AT5G43100	AT5G43100.1	Eukaryotic aspartyl protease family protein;(source:Araport11)	protein_coding
215	gyrB	AT5G51820	AT5G51820.1	Encodes a plastid isoform of the enzyme phosphoglucomutase involved in controlling photosynthetic carbon flow. Effective petiole movement against the direction of the gravity requires functional PGM activity that is required for full development of amyloplasts.	protein_coding
216	gyrB	AT5G54010	AT5G54010.1	Encodes a flavonoid 3-O-glucoside:2″-O-glucosyltransferase that determines pollen-specific flavonol structure.	protein_coding
217	gyrB	AT5G54130	AT5G54130.2	Calcium-binding endonuclease/exonuclease/phosphatase family;(source:Araport11)	protein_coding
218	gyrB	AT5G55610	AT5G55610.1	isopentenyl-diphosphate delta-isomerase;(source:Araport11)	protein_coding
219	gyrB	AT5G55620	AT5G55620.1	hypothetical protein;(source:Araport11)	protein_coding
220	gyrB	AT5G55630	AT5G55630.1	Encodes AtTPK1 (KCO1), a member of the Arabidopsis thaliana K+ channel family of AtTPK/KCO proteins. AtTPK1 is targeted to the vacuolar membrane. Forms homomeric ion channels in vivo. Voltage-independent and Ca2+-activated K+ channel. Activated by 14-3-3 proteins. Vacuolar K+-conducting TPC1 and TPK1/TPK3 channels act in concert to provide for Ca2+- and voltageinduced electrical excitability to the central organelle of plant cells.	protein_coding
221	gyrB	AT5G55640	AT5G55640.1	Na-translocating NADH-quinone reductase subunit A;(source:Araport11)	protein_coding
222	gyrB	AT5G55660	AT5G55660.1	DEK domain-containing chromatin associated protein	protein_coding
223	gyrB	AT5G55670	AT5G55670.1	RNA-binding (RRM/RBD/RNP motifs) family protein;(source:Araport11)	protein_coding
224	gyrB	AT5G55690	AT5G55690.1	AGL47 MADS box gene.	protein_coding
225	gyrB	AT5G55700	AT5G55700.1	In vitro assay indicates no beta-amylase activity of BAM4. However mutation in BAM4 impairs starch breakdown. BAM4 may play a regulatory role.	protein_coding
226	gyrB	AT5G55710	AT5G55710.1	Encodes a component of the TIC (translocon at the inner envelope membrane of chloroplasts) protein translocation machinery mediating the protein translocation across the inner envelope of plastids. The Arabidopsis genome encodes four Tic20 homologous proteins, AT1G04940(Tic20-I), AT2G47840(Tic20-II), AT4G03320(Tic20-IV) and AT5G55710(Tic20-V).	protein_coding
227	gyrB	AT5G55720	AT5G55720.1	Pectin lyase-like superfamily protein;(source:Araport11)	protein_coding
228	gyrB	AT5G55820	AT5G55820.1	Encodes a plant ortholog of the inner centromere protein (INCENP), which is implicated in the control of chromosome segregation and cytokinesis in yeast and animals. Required for female gametophytic cell specification and seed development.	protein_coding
229	gyrB	AT5G56180	AT5G56180.1	encodes a protein whose sequence is similar to actin-related proteins (ARPs) in other organisms. Member of nuclear ARP family of genes.	protein_coding
230	gyrB	AT5G58140	AT5G58140.1	Membrane-bound protein serine/threonine kinase that functions as blue light photoreceptor in redundancy with PHO1. Involved in stomatal opening, chloroplast movement and phototropism. Mediates blue light-induced growth enhancements. PHOT1 and PHOT2 mediate blue light-dependent activation of the plasma membrane H+-ATPase in guard cell protoplasts. PHOT2 possesses two LOV (LOV1 and LOV2, for light-oxygen-voltage-sensing) domains involved in FMN-binding and a C-terminus forming a serine/threonine kinase domain. LOV2 acts as an inhibitor of phototropin kinase in the dark, and light cancels the inhibition through cysteine-FMN adduct formation. LOV1 in contrast acts as an attenuator of photoactivation. Localized to the Golgi apparatus under the induction of blue light. The mRNA is cell-to-cell mobile.	protein_coding
231	gyrB	AT5G58180	AT5G58180.1	member of YKT6 Gene Family, R-SNARE protein. Together with YKT61 interacts with SYP41 and are essential for membrane fusion.	protein_coding
232	gyrB	AT5G58680	AT5G58680.1	ARM repeat superfamily protein;(source:Araport11)	protein_coding
233	gyrB	AT5G58690	AT5G58690.1	phosphatidylinositol-speciwc phospholipase C5;(source:Araport11)	protein_coding

234	gyrB	AT5G59710	AT5G59710.1	Encodes a nuclear-localized NOT (negative on TATA-less) domain-containing protein that interacts with the Agrobacterium VirE2 protein and is required for Agrobacterium-mediated plant transformation. It likely facilitates T-DNA integration into plant chromosomes and may play a role as a transcriptional regulator. The mRNA is cell-to-cell mobile.	protein_coding
235	gyrB	AT5G59940	AT5G59940.1	Cysteine/Histidine-rich C1 domain family protein;(source:Araport11)	protein_coding
236	gyrB	AT5G62770	AT5G62770.1	membrane-associated kinase regulator, putative (DUF1645);(source:Araport11)	protein_coding
237	gyrB	AT5G64230	AT5G64230.1	1,8-cineole synthase;(source:Araport11)	protein_coding
238	gyrB	AT5G64240	AT5G64240.2	Encodes a type I metacaspase. Two Arabidopsis metacaspases, AT1G02170 (MC1) and AT4G25110 (MC2) antagonistically control programmed cell death in Arabidopsis. MC1 is a positive regulator of cell death and requires conserved caspase-like putative catalytic residues for its function. MC2 negatively regulates cell death. This function is independent of the putative catalytic residues. A third type I Arabidopsis metacaspase is MC3 (AT5g64240).	protein_coding
239	gyrB	AT3G60318	NA	NA	NA

BIBLIOGRAPHY

Albright, Michaeline B.N. et al. (2022). "Solutions in microbiome engineering: prioritizing barriers to organism establishment". In: *ISME Journal* 16.2, pp. 331–338. ISSN: 17517370. DOI: 10.1038/s41396-021-01088-5.

Bai, Yang et al. (2015). "Functional overlap of the Arabidopsis leaf and root microbiota". In: *Nature* 528.7582, pp. 364–369. ISSN: 14764687. DOI: 10.1038/nature16192.

Bankevich, Anton et al. (2012). "SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing". In: *Journal of Computational Biology* 19.5, pp. 455–477. ISSN: 10665277. DOI: 10.1089/cmb.2012.0021.

Barret, Matthieu et al. (2015). "Emergence shapes the structure of the seed microbiota". In: *Applied and Environmental Microbiology* 81.4, pp. 1257–1266. ISSN: 10985336. DOI: 10.1128/AEM.03722-14.

Bartoli, Claudia and Fabrice Roux (2017). "Genome-wide association studies in plant pathosystems: Toward an ecological genomics approach". In: *Frontiers in Plant Science* 8.May. ISSN: 1664462X. DOI: 10.3389/fpls.2017.00763.

Bartoli, Claudia et al. (2018). "In situ relationships between microbiota and potential pathobiota in Arabidopsis thaliana". In: *ISME Journal* 12.8, pp. 2024–2038. ISSN: 17517370. DOI: 10.1038/s41396-018-0152-7. URL: http://dx.doi.org/10.1038/s41396-018-0152-7.

Bates, Douglas et al. (2015). "Fitting linear mixed-effects models using lme4". In: *Journal of Statistical Software* 67.1. ISSN: 15487660. DOI: 10.18637/jss.v067.i01. arXiv: 1406.5823.

Beilsmith, Kathleen et al. (2019). "Genome-wide association studies on the phyllosphere microbiome: Embracing complexity in host–microbe interactions". In: *Plant Journal* 97.1, pp. 164–181. ISSN: 1365313X. DOI: 10.1111/tpj.14170.

Bergelson, Joy et al. (2021). "Assessing the potential to harness the microbiome through plant genetics". In: *Current Opinion in Biotechnology* 70, pp. 167–173. ISSN: 18790429. DOI: 10.1016/j.copbio.2021.05.007. URL: https://doi.org/10.1016/j.copbio.2021.05.007.

Bernardo, Rex (2020). *Breeding for Quantitative Traits in Plants*. 3rd ed. Woodbury: Stemma Press, pp. 229–256.

Bernardo-Cravo, Adriana P. et al. (2020). "Environmental Factors and Host Microbiomes Shape Host–Pathogen Dynamics". In: *Trends in Parasitology* 36.7, pp. 616–633. ISSN: 14715007. DOI: 10.1016/j.pt.2020.04.010. URL: https://doi.org/10.1016/j.pt.2020.04.010.

Bodenhausen, Natacha, Matthew W. Horton, and Joy Bergelson (2013). "Bacterial Communities Associated with the Leaves and the Roots of Arabidopsis thaliana". In: *PLoS ONE* 8.2. ISSN: 19326203. DOI: 10.1371/journal.pone.0056329.

Bodenhofer, Ulrich et al. (2015). "Msa: An R package for multiple sequence alignment". In: *Bioinformatics* 31.24, pp. 3997–3999. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv494.

Brachi, Benjamin et al. (2022). "Plant genetic effects on microbial hubs impact host fitness in repeated field trials". In: *Proceedings of the National Academy of Sciences of the United States of America* 119.30. ISSN: 10916490. DOI: 10.1073/pnas.2201285119.

Brader, Günter et al. (2017). "Ecology and Genomic Insights into Plant-Pathogenic and Plant-Nonpathogenic Endophytes". In: *Annual Review of Phytopathology* 55, pp. 61–83. ISSN: 15452107. DOI: 10.1146/annurev-phyto-080516-035641.

Brown, Shawn P. et al. (2021). "Correction to: Soil origin and plant genotype structure distinct microbiome compartments in the model legume Medicago truncatula (Microbiome, (2020), 8, 1, (139), 10.1186/s40168-020-00915-9)". In: *Microbiome* 9.1, pp. 1–17. ISSN: 20492618. DOI: 10.1186/s40168-021-01080-3.

Burnham, Kenneth P. and David R. Anderson (2004). "Multimodel inference: Understanding AIC and BIC in model selection". In: *Sociological Methods and Research* 33.2, pp. 261–304. ISSN: 00491241. DOI: 10.1177/0049124104268644.

Bushnell, Brian, Jonathan Rood, and Esther Singer (2017). "BBMerge – Accurate paired shotgun read merging via overlap". In: *PLoS ONE* 12.10, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0185056.

Callahan, Benjamin J. et al. (2016). "DADA2: High-resolution sample inference from Illumina amplicon data". In: *Nature Methods* 13.7, pp. 581–583. ISSN: 15487105. DOI: 10.1038/nmeth.3869.

Chelius, M. K. and E. W. Triplett (2001). "The diversity of archaea and bacteria in association with the roots of Zea mays L." In: *Microbial Ecology* 41.3, pp. 252–263. ISSN: 00953628. DOI: 10.1007/s002480000087.

Chen, Lianmin et al. (2018). "A system biology perspective on environment-host-microbe interactions". In: *Human Molecular Genetics* 27.R2, R187–R194. ISSN: 14602083. DOI: 10.1093/hmg/ddy137.

Chen, Lihong et al. (2016). "VFDB 2016: Hierarchical and refined dataset for big data analysis - 10 years on". In: *Nucleic Acids Research* 44.D1, pp. D694–D697. ISSN: 13624962. DOI: 10.1093/nar/gkv1239.

Compant, Stéphane et al. (2019). "A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application". In: *Journal of Advanced Research* 19, pp. 29–37. ISSN: 20901232. DOI: 10.1016/j.jare.2019.03.004.

Coyte, Katharine Z. and Seth Rakoff-Nahoum (2019). "Understanding Competition and Cooperation within the Mammalian Gut Microbiome". In: *Current Biology* 29.11, pp. 538–544. ISSN: 0000000000. DOI: 10.1016/j.cub.2019.04.017.Understanding.

Deng, Siwen et al. (2021). "Genome wide association study reveals plant loci controlling heritability of the rhizosphere microbiome". In: *ISME Journal* 15.11, pp. 3181–3194. ISSN: 17517370. DOI: 10.1038/s41396-021-00993-z.

Dillon, Marcus M. et al. (2019). "Molecular evolution of Pseudomonas syringae type III secreted effector proteins". In: *Frontiers in Plant Science* 10.April, pp. 1–18. ISSN: 1664462X. DOI: 10.3389/fpls.2019.00418.

Dixon, Philip (2003). "Computer program review VEGAN , a package of R functions for community ecology". In: *Journal of Vegetation Science* 14.6, pp. 927–930. URL: http://doi.wiley.com/10.1111/j.1654-1103.2002.tb02049.x.

Eren, A. Murat et al. (2015). "Anvi'o: An advanced analysis and visualization platformfor 'omics data". In: *PeerJ* 2015.10, pp. 1–29. ISSN: 21678359. DOI: 10.7717/peerj.1319.

Finkel, Omri M. et al. (2020). "A single bacterial genus maintains root growth in a complex microbiome". In: *Nature* 587.7832, pp. 103–108. ISSN: 14764687. DOI: 10.1038/s41586-020-2778-7. URL: http://dx.doi.org/10.1038/s41586-020-2778-7.

Foster, Kevin R. et al. (2017). "The evolution of the host microbiome as an ecosystem on a leash". In: *Nature* 548.7665, pp. 43–51. ISSN: 14764687. DOI: 10.1038/nature23292. URL: http://dx.doi.org/10.1038/nature23292.

Friedman, Jonathan and Eric J. Alm (2012). "Inferring Correlation Networks from Genomic Survey Data". In: *PLoS Computational Biology* 8.9, pp. 1–11. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1002687.

Garza, Daniel R. et al. (2018). "Towards predicting the environmental metabolome from metagenomics with a mechanistic model". In: *Nature Microbiology* 3.4, pp. 456–460. ISSN: 20585276. DOI: 10.1038/s41564-018-0124-8. URL: http://dx.doi.org/10.1038/s41564-018-0124-8.

Glöckner, Frank Oliver et al. (2017). "25 years of serving the community with ribosomal RNA gene reference databases and tools". In: *Journal of Biotechnology* 261.February,

pp. 169–176. ISSN: 18734863. DOI: 10.1016/j.jbiotec.2017.06.1198. URL: http://dx.doi.org/10.1016/j.jbiotec.2017.06.1198.

Gloss, Andrew D. et al. (2022). " Genome-wide association mapping within a local Arabidopsis thaliana population more fully reveals the genetic architecture for defensive metabolite diversity ". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 377.1855. ISSN: 0962-8436. DOI: 10.1098/rstb.2020.0512.

Grieneisen, Laura et al. (2021). "Gut microbiome heritability is nearly universal but environmentally contingent". In: *Science* 373.6551, pp. 181–186. ISSN: 10959203. DOI: 10.1126/science.aba5483.

Gurevich, Alexey et al. (2013). "QUAST: Quality assessment tool for genome assemblies". In: *Bioinformatics* 29.8, pp. 1072–1075. ISSN: 13674803. DOI: 10.1093/bioinformatics/btt086.

Hassler, Hayley B. et al. (2022). "Phylogenies of the 16S rRNA gene and its hypervariable regions lack concordance with core genome phylogenies". In: *Microbiome* 10.1, pp. 1–18. ISSN: 20492618. DOI: 10.1186/s40168-022-01295-y. URL: https://doi.org/10.1186/s40168-022-01295-y.

He, Xiaoqing et al. (2021). "Network mapping of root–microbe interactions in Arabidopsis thaliana". In: *npj Biofilms and Microbiomes* 7.1, pp. 1–10. ISSN: 20555008. DOI: 10.1038/s41522-021-00241-4. URL: http://dx.doi.org/10.1038/s41522-021-00241-4.

Henry, Lucas P. and Joy Bergelson (2023). "Evolutionary implications of host genetic control for engineering beneficial microbiomes". In: *Current Opinion in Systems Biology* 34.Box 1, p. 100455. ISSN: 24523100. DOI: 10.1016/j.coisb.2023.100455. URL: https://doi.org/10.1016/j.coisb.2023.100455.

Henry, Lucas P. et al. (2021). "The microbiome extends host evolutionary potential". In: *Nature Communications* 12.1, pp. 1–13. ISSN: 20411723. DOI: 10.1038/s41467-021-25315-x. URL: http://dx.doi.org/10.1038/s41467-021-25315-x.

Heravi, Fatemah Sadeghpour et al. (2020). "Host DNA depletion efficiency of microbiome DNA enrichment methods in infected tissue samples". In: *Journal of Microbiological Methods* 170.March. ISSN: 18728359. DOI: 10.1016/j.mimet.2020.105856.

Horton, Matthew W. et al. (2014). "Genome-wide association study of Arabidopsis thaliana leaf microbial community". In: *Nature Communications* 5.May, pp. 1–7. ISSN: 20411723. DOI: 10.1038/ncomms6320.

Ikeda, Seishi et al. (2009). "Development of a Bacterial Cell Enrichment Method and its Application to the Community Analysis in Soybean Stems". In: *Microbial Ecology* 58.4, pp. 703–714. ISSN: 00953628. DOI: 10.1007/s00248-009-9566-0.

Illumina (2013). "16S Metagenomic Sequencing Library". In: *Illumina.com* B, pp. 1–28. URL: http://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry{_}documentation/16s/16s-metagenomic-library-prep-guide-15044223-b.pdf.

Jones, J.D.G. and J.L. Dangl (2006). "The plant immune system". In: *Nature* 444.7117, pp. 323–329.

Karasov, Talia et al. (2019). "The relationship between microbial population size and disease in the Arabidopsis thaliana phyllosphere". In: DOI: 10.1101/828814.

Ke, Jing, Bing Wang, and Yasuo Yoshikuni (2021). "Microbiome Engineering: Synthetic Biology of Plant-Associated Microbiomes in Sustainable Agriculture". In: *Trends in Biotechnology* 39.3, pp. 244–261. ISSN: 18793096. DOI: 10.1016/j.tibtech.2020.07.008. URL: https://doi.org/10.1016/j.tibtech.2020.07.008.

Knights, Dan et al. (2014). "Complex host genetics influence the microbiome in inflammatory bowel disease". In: *Genome Medicine* 6.12, pp. 1–11. ISSN: 1756994X. DOI: 10.1186/s13073-014-0107-1.

Korte, Arthur and Farlow Ashley (2013). "The advantages and limitations of trait analysis with GWAS : a review Self-fertilisation makes Arabidopsis particularly well suited to GWAS". In: *Plant methods* 9.1, p. 29. ISSN: 1746-4811.

Kurtz, Zachary D. et al. (2015). "Sparse and Compositionally Robust Inference of Microbial Ecological Networks". In: *PLoS Computational Biology* 11.5, pp. 1–25. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1004226. arXiv: 1408.4158.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune H.B. Christensen (2017). "lmerTest Package: Tests in Linear Mixed Effects Models". In: *Journal of Statistical Software* 82.13, pp. 1–26. ISSN: 15487660. DOI: 10.18637/JSS.V082.I13.

Lamble, Sarah et al. (2013). "Improved workflows for high throughput library preparation using the transposome-based nextera system". In: *BMC Biotechnology* 13. ISSN: 14726750. DOI: 10.1186/1472-6750-13-104.

Lane, David J et al. (1985). "Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses (reverse transcriptase/dideoxynudeotide)". In: *Evolution* 82.October, pp. 6955–6959.

Li, Shao peng et al. (2019). "Niche and fitness differences determine invasion success and impact in laboratory bacterial communities". In: *ISME Journal* 13.2, pp. 402–412. ISSN: 17517370. DOI: 10.1038/s41396-018-0283-x. URL: http://dx.doi.org/10.1038/s41396-018-0283-x.

Loy, Adam and Heike Hofmann (2014). "HLMdiag: A suite of diagnostics for hierarchical linear models in R". In: *Journal of Statistical Software* 56.5, pp. 1–28. ISSN: 15487660. DOI: 10.18637/jss.v056.i05.

Marquet, Mike et al. (2022). "Evaluation of microbiome enrichment and host DNA depletion in human vaginal samples using Oxford Nanopore's adaptive sequencing". In: *Scientific Reports* 12.1, pp. 1–10. ISSN: 20452322. DOI: 10.1038/s41598-022-08003-8. URL: https://doi.org/10.1038/s41598-022-08003-8.

Martineau, Christine et al. (2018). "Serratia marcescens outbreak in a neonatal intensive care unit: New insights from next-generation sequencing applications". In: *Journal of Clinical Microbiology* 56.9. ISSN: 1098660X. DOI: 10.1128/JCM.00235-18.

Mi, Huaiyu, Anushya Muruganujan, and Paul D. Thomas (2013). "PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees". In: *Nucleic Acids Research* 41.D1, pp. 377–386. ISSN: 03051048. DOI: 10.1093/nar/gks1118.

Morton, James T. et al. (2019). "Establishing microbial composition measurement standards with reference frames". In: *Nature Communications* 10.1. ISSN: 20411723. DOI: 10.1038/s41467-019-10656-5. URL: http://dx.doi.org/10.1038/s41467-019-10656-5.

Nyquist, Wyman E. (1991). "Estimation of Heritability and Prediction of Selection Response in Plant Populations". In: *Critical Reviews in Plant Sciences* 10.3, pp. 235–322. ISSN: 15497836. DOI: 10.1080/07352689109382313.

Pace, Norman R., Jan Sapp, and Nigel Goldenfeld (2012). "Phylogeny and beyond: Scientific, historical, and conceptual significance of the first tree of life". In: *Proceedings of the National Academy of Sciences of the United States of America* 109.4, pp. 1011–1018. ISSN: 10916490. DOI: 10.1073/pnas.1109716109.

Paradis, Emmanuel, Julien Claude, and Korbinian Strimmer (2004). "APE: Analyses of phylogenetics and evolution in R language". In: *Bioinformatics* 20.2, pp. 289–290. ISSN: 13674803. DOI: 10.1093/bioinformatics/btg412.

Perisin, Matthew et al. (2016). "16Stimator: Statistical estimation of ribosomal gene copy numbers from draft genome assemblies". In: *ISME Journal* 10.4, pp. 1020–1024. ISSN: 17517370. DOI: 10.1038/ismej.2015.161.

Poirier, Simon et al. (2018). "Deciphering intra-species bacterial diversity of meat and seafood spoilage microbiota using gyrB amplicon sequencing: A comparative analysis with 16S

rDNA V3-V4 amplicon sequencing". In: *PLoS ONE* 13.9, pp. 1–26. ISSN: 19326203. DOI: 10.1371/journal.pone.0204629.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin (2010). "FastTree 2 - Approximately maximum-likelihood trees for large alignments". In: *PLoS ONE* 5.3. ISSN: 19326203. DOI: 10.1371/journal.pone.0009490.

Quast, Christian et al. (2013). "The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools". In: *Nucleic Acids Research* 41.D1, pp. 590–596. ISSN: 03051048. DOI: 10.1093/nar/gks1219.

Quinn, Thomas P. et al. (2019). "A field guide for the compositional analysis of any-omics data". In: *GigaScience* 8.9, pp. 1–14. ISSN: 2047217X. DOI: 10.1093/gigascience/giz107.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramírez-Sánchez, Daniela et al. (2022). "Investigating genetic diversity within the most abundant and prevalent non-pathogenic leaf-associated bacteria interacting with Arabidopsis thaliana in natural habitats". In: *Frontiers in Microbiology* 13.September. ISSN: 1664302X. DOI: 10.3389/fmicb.2022.984832.

Reimer, Lorenz Christian et al. (2022). "BacDive in 2022: The knowledge base for standardized bacterial and archaeal data". In: *Nucleic Acids Research* 50.D1, pp. D741–D746. ISSN: 13624962. DOI: 10.1093/nar/gkab961.

Rohland, Nadin and David Reich (2012). "Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture". In: *Genome Research* 22.5, pp. 939–946. ISSN: 10889051. DOI: 10.1101/gr.128124.111.

Rojas-Gätjens, Diego et al. (2022). "Antibiotic-producing Micrococcales govern the microbiome that inhabits the fur of two- and three-toed sloths". In: *Environmental Microbiology* 24.7, pp. 3148–3163. ISSN: 14622920. DOI: 10.1111/1462-2920.16082.

Rothschild, Daphna et al. (2018). "Environment dominates over host genetics in shaping human gut microbiota". In: *Nature* 555.7695, pp. 210–215. ISSN: 14764687. DOI: 10.1038/nature25973.

Sandoval-Motta, Santiago et al. (2017). "The human microbiome and the missing heritability problem". In: *Frontiers in Genetics* 8.JUN, pp. 1–12. ISSN: 16648021. DOI: 10.3389/fgene.2017.00080.

Schielzeth, Holger et al. (2020). "Robustness of linear mixed-effects models to violations of distributional assumptions". In: *Methods in Ecology and Evolution* 11.9, pp. 1141–1152. ISSN: 2041210X. DOI: 10.1111/2041-210X.13434.

Schliep, Klaus Peter (2011). "phangorn: Phylogenetic analysis in R". In: *Bioinformatics* 27.4, pp. 592–593. ISSN: 13674803. DOI: 10.1093/bioinformatics/btq706.

Schmidt, Thomas S.B., Jeroen Raes, and Peer Bork (2018). "The Human Gut Microbiome: From Association to Modulation". In: *Cell* 172.6, pp. 1198–1215. ISSN: 10974172. DOI: 10.1016/j.cell.2018.02.044. URL: https://doi.org/10.1016/j.cell.2018.02.044.

Schneider, Caroline A., Wayne S. Rasband, and Kevin W. Eliceiri (2012). "NIH Image to ImageJ: 25 years of image analysis". In: *Nature Methods* 9.7, pp. 671–675. ISSN: 15487091. DOI: 10.1038/nmeth.2089.

Seemann, Torsten (June 10, 2023). *ABRicate*. Version 0.8.13. URL: https://github.com/tseemann/abricate.

Shahzad, Raheem et al. (2018). "What is there in seeds? Vertically transmitted endophytic resources for sustainable improvement in plant growth". In: *Frontiers in Plant Science* 9.January, pp. 1–10. ISSN: 1664462X. DOI: 10.3389/fpls.2018.00024.

Shen, Wei and Hong Ren (2021). "TaxonKit: A practical and efficient NCBI taxonomy toolkit". In: *Journal of Genetics and Genomics* 48.9, pp. 844–850. ISSN: 18735533. DOI: 10.1016/j.jgg.2021.03.006. URL: https://doi.org/10.1016/j.jgg.2021.03.006.

Shen, Wei et al. (2016). "SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation". In: *PLoS ONE* 11.10, pp. 1–10. ISSN: 19326203. DOI: 10.1371/journal.pone.0163962.

Stevens, Russell B. (1960). *Cultural Practices in Disease Control*. Academic Press Inc., pp. 357–429. DOI: 10.1016/b978-0-12-395678-1.50016-3. URL: http://dx.doi.org/10.1016/B978-0-12-395678-1.50016-3.

Sundarraman, Deepika et al. (2020). "Higher-order interactions dampen pairwise competition in the zebrafish gut microbiome". In: *mBio* 11.5, pp. 1–15. ISSN: 21507511. DOI: 10.1128/mBio.01667-20.

The 1001 Genomes Consortium (2016). "1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana". In: *Cell* 166.2, pp. 481–491. ISSN: 10974172. DOI: 10.1016/j.cell.2016.05.063.

The Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the ¯owering plant Arabidopsis thaliana". In: *Nature* 408, pp. 796–815. ISSN: 10227954. DOI: 10.1134/S1022795411020074.

Thomas, Paul D. et al. (2022). "PANTHER: Making genome-scale phylogenetics accessible to all". In: *Protein Science* 31.1, pp. 8–22. ISSN: 1469896X. DOI: 10.1002/pro.4218.

Trivedi, Pankaj et al. (2020). "Plant–microbiome interactions: from community assembly to plant health". In: *Nature Reviews Microbiology* 18.11, pp. 607–621. ISSN: 17401534. DOI: 10.1038/s41579-020-0412-1. URL: http://dx.doi.org/10.1038/s41579-020-0412-1.

Van Wees, Saskia CM, Sjoerd Van der Ent, and Corné MJ Pieterse (2008). "Plant immune responses triggered by beneficial microbes". In: *Current Opinion in Plant Biology* 11.4, pp. 443–448. ISSN: 13695266. DOI: 10.1016/j.pbi.2008.05.005.

Větrovský, Tomáš and Petr Baldrian (2013). "The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses". In: *PLoS ONE* 8.2, pp. 1–10. ISSN: 19326203. DOI: 10.1371/journal.pone.0057923.

Vogel, Christine M. et al. (2021). "Protective role of the Arabidopsis leaf microbiota against a bacterial pathogen". In: *Nature Microbiology* 6.12, pp. 1537–1548. ISSN: 20585276. DOI: 10.1038/s41564-021-00997-7.

Vorholt, Julia A. (2012). "Microbial life in the phyllosphere". In: *Nature Reviews Microbiology* 10.12, pp. 828–840. ISSN: 17401526. DOI: 10.1038/nrmicro2910.

Wagner, Maggie R. (2021). "Prioritizing host phenotype to understand microbiome heritability in plants". In: *New Phytologist* 232.2, pp. 502–509. ISSN: 14698137. DOI: 10.1111/nph.17622.

Walters, William A. et al. (2018). "Large-scale replicated field study of maize rhizosphere identifies heritable microbes". In: *Proceedings of the National Academy of Sciences of the United States of America* 115.28, pp. 7368–7373. ISSN: 10916490. DOI: 10.1073/pnas.1800918115.

Wang, Qiong et al. (2007). "Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy". In: *Applied and Environmental Microbiology* 73.16, pp. 5261–5267. ISSN: 00992240. DOI: 10.1128/AEM.00062-07.

Watanabe, Kanako et al. (2001). "ICB database: The gyrB database for identification and classification of bacteria". In: *Nucleic Acids Research* 29.1, pp. 344–345. ISSN: 03051048. DOI: 10.1093/nar/29.1.344.

Weinstein, Michael M. et al. (2019). "FIGARO: An efficient and objective tool for optimizing microbiome rRNA gene trimming parameters". In: *bioRxiv*. DOI: https://doi.org/10.1101/610394. URL: https://doi.org/10.1101/610394.

Woese, C. R. and G. E. Fox (1977). "Phylogenetic structure of the prokaryotic domain: The primary kingdoms". In: *Proceedings of the National Academy of Sciences of the United*

States of America 74.11, pp. 5088–5090. ISSN: 00278424. DOI: 10.1073/pnas.74.11.5088.

Wu, Jiaqiang et al. (2021). "Honey bee genetics shape the strain-level structure of gut microbiota in social transmission". In: *Microbiome* 9.1, pp. 1–19. ISSN: 20492618. DOI: 10.1186/s40168-021-01174-y.

Yamamoto, S. and S. Harayama (1995). "PCR amplification and direct sequencing of gyrB genes with universal primers and their application to the detection and taxonomic analysis of Pseudomonas putida strains". In: *Applied and Environmental Microbiology* 61.3, pp. 1104–1109. ISSN: 00992240. DOI: 10.1128/aem.61.3.1104-1109.1995.

Zhou, Xiang and Matthew Stephens (2012). "Genome-wide efficient mixed-model analysis for association studies". In: *Nature Genetics* 44.7, pp. 821–824. ISSN: 10614036. DOI: 10.1038/ng.2310.