Comparison of nodule endophyte composition, diversity, and gene content between *Medicago truncatula* genotypes

Mannix Burns[1], Brendan Epstein[1], Liana T. Burghardt[1,2*]

[1]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108

[2]current affiliation: Department of Plant Science, The Pennsylvania State University, University Park, PA, 16802, USA

*Author for correspondence: Liana T. Burghardt; E-mail: liana.burghardt@gmail.com

Funding: This work was supported by the National Science Foundation under Grant Numbers, IOS-1724993, 1856744. This work was supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under MIN-71-030 and Project #PEN04760 and Accession #1025611.

23

ABSTRACT

Leguminous plants form symbiotic relationships with rhizobia. These nitrogen-fixing bacteria live in specialized root organs called nodules. While rhizobia form the most notable host relationship within root nodules, other bacterial endophytes also inhabit nodules and can influence host-rhizobia interactions as well as exert effects of their own. In this study, we elucidate the effect of intraspecific host genetic variation and between generational legacies on root nodule endophyte communities in *Medicago truncatula*. While the diversity of endophytes in nodules was similar across hosts, both nodule endophyte composition and gene functional groups differed. In contrast, neither the presence nor identity of a host in the previous generation (either A17 or R108) had a significant effect on the nodule endophyte diversity or composition. However, whether or not a host was present significantly altered gene functional groups. We conclude that genetic variation within a legume host species can play both direct and indirect effects on establishment of nodule microbiomes. Further studies, including GWAS and functional assays, can open the door for engineering and optimizing nodule endophyte communities that promote growth or have other beneficial qualities.

Keywords (4-6): *Medicago-Ensifer*; root nodule microbiome, genotype-dependent effects, host-soil feedbacks, phytobiomes, symbiosis

INTRODUCTION

Endophytic bacteria and fungi are common inhabitants of plant tissues, such as roots and leaves, and often live in the soil when not infecting a host. In the case of legumes, nitrogen-fixing bacterial species (collectively called rhizobia) form symbiotic relationships with the host in specialized root organs called nodules (Long, 1989). Although rhizobia are the primary bacterial inhabitants, other bacteria also inhabit nodules and have measurable phenotypic effects. For instance, co-inoculation of common bean (*Phaseolus vulgaris*) with endophytic strains of *Bacillus*, *Pseudomonas*, or *Burkholderia* reduces damping-off (Ferreira et al., 2020) while co-inoculation with *Agrobacterium* reduces nodulation (Mhamdi et al., 2005; Mrabet et al., 2006). Endophytic bacteria in Chickpeas (*Cicer arietinum L.*) have been shown to increase plant growth and suppress root rot (Egamberdieva et al., 2017). While a few examples have been studied, overall, little is known about the identity and phenotypic effects of nodule endophytes despite their potential value as biocontrol agents (Martínez-Hidalgo & Hirsch, 2017).

The composition of nodule microbiome communities differs between plant species, suggesting that genetic differences can have substantial effects on which bacteria inhabit nodules (Xiao et al., 2017). There is reason to suspect, based on data from other plant tissues, that genetic variation within a species could also play a role in determining nodule endophyte composition. For example, tomato cultivars differ in leaf microbiomes (Morella et al., 2020), and host genotype plays a role in the composition of the endophyte communities in natural populations of some plant species such as *Boechera stricta* (Wagner et al., 2016) and *Populus balsamifera* (Bálint et al., 2013) and

agricultural cultivars of *Zea mays* (Ikeda et al., 2013). Furthermore, reverse genetics experiments indicate that genes involved in symbiosis may play a role in determining the composition of the microbial community in the roots, rhizosphere, and nodules (Zgadzaj et al., 2016). Previous studies in the legumes *Medicago truncatula*, *Glycine max* (soybean) and *Vigna unguiculata* (cowpea), respectively, have shown mixed results: plant host genotype influences nodule endophyte composition (Brown et al. 2020; Sharaf et al. 2019), but not nodule endophyte diversity (Leite et al. 2016). To our knowledge, no studies have tested for intergenerational effects of prior hosts on nodule endophyte communities—although both inter- and intra-specific plant genetic variation can clearly influence soil microbiome composition and function (Fischer et al., 2013, Cloutier et al., 2020). Due to the scarcity of studies and the broad phylogenetic range of host species tested, it is difficult to formulate *a priori* predictions. Here we examine nodule endophyte communities in two accessions of the model legume *Medicago truncatula* and test for effect of exposure to a previous generation of each host genotype.

M. truncatula harbors genetic variation for the diversity and composition of nitrogen-fixing *Ensifer*, also referred to as *Sinorhizobium,* strains with which they form nodules (Heath & Stinchcombe 2013, Heath & Grillo 2016, Burghardt et al., 2018, Burghardt et al., 2019). Burghardt et. al (2019) used a “select and resequence” experiment —whole genome sequencing of rhizobia in pools of nodules—to measure the relative frequency of 101 co-inoculated strains in two Medicago genotypes. The rhizobial community in the model genotype A17 was less diverse than and compositionally divergent from the rhizobial community in the model genotype R108

(Burghardt et al., 2019). In the same experiment, they found evidence that selective legacies of prior host genotypes exert effects on the diversity and composition of rhizobia in nodules—although the magnitude of the effect was far weaker than the direct effect of host genotype. We were curious if these diversity and compositional differences extend to non-rhizobial bacteria in the nodules.

Here, we use the data from Burghardt et al. (2019) to elucidate the effect of intraspecific host genetic variation and the between-generational legacies of that genetic variation on root nodule endophyte communities in *M. truncatula* hosts. Specifically, we ask 1) are there significant differences in overall diversity (the number and abundance of different endophytes), composition (the specific types of endophytes present), and gene content of the endophyte communities between the two hosts; and 2) does the presence or identity of a prior host influence the overall diversity, composition, and gene content of the endophyte community in subsequent generations?

METHODS

Experimental Design

We assessed nodule endophyte communities in two *Medicago truncatula* genotypes (A17 and R108) growing in the greenhouse (Fig. 1). A full description of the experimental design can be found in (Burghardt et al., 2019). Soil and water were sterilized via autoclave (120-minute liquid cycle) while seeds were surface sterilized for 30-45 seconds in 10% bleach then rinsed with sterilized water. To confirm sterility, both

water and a dilute soil slurry were plated on rich media to check for microbial growth. Briefly, we filled thirty 1L pots with sterilized low-nitrogen peat soil (SunGro LP5) and inoculated each with a mixture of 101 previously-sequenced *Ensifer meliloti* strains (~10^9 cells per pot). Sixteen of these pots were in the greenhouse for twelve weeks with no plants while seven pots were planted with 10-12 seeds of A17 and seven with 10-12 seeds of R108. Six weeks after planting, the above-ground vegetation was cut off and the roots and nodules were allowed to decompose for 6 weeks. Next, 10-12 sterilized and pre-germinated seeds of either A17 or R108 were planted in the soil. Eight of the pots that had not been planted previously received A17 seeds, and the other eight received R108 seeds. All pots that previously had host plants were planted with R108 seeds so the effect of prior selection could be measured in a common genotype. R108→A17 and A17→A17 treatments were not investigated due to sequencing and space limitations. The plants were allowed to grow and form nodules for six weeks before they were harvested, at which point all the nodules were removed. These nodules were pooled, surface sterilized for 30-45 seconds in 10% bleach, rinsed with sterilized water four times, and crushed using a sterile pestle. The bacterial community (*Ensifer* strains + other endophytes) from the crushed nodules was separated through slow centrifugation to pellet plant debris and bacteroids (400 x g for 10 minutes), and fast centrifugation of the supernatant from the previous step (10,000 x g for 8 minutes) to pellet the bacteria. Full details can be found in Burghardt et al. (2019).

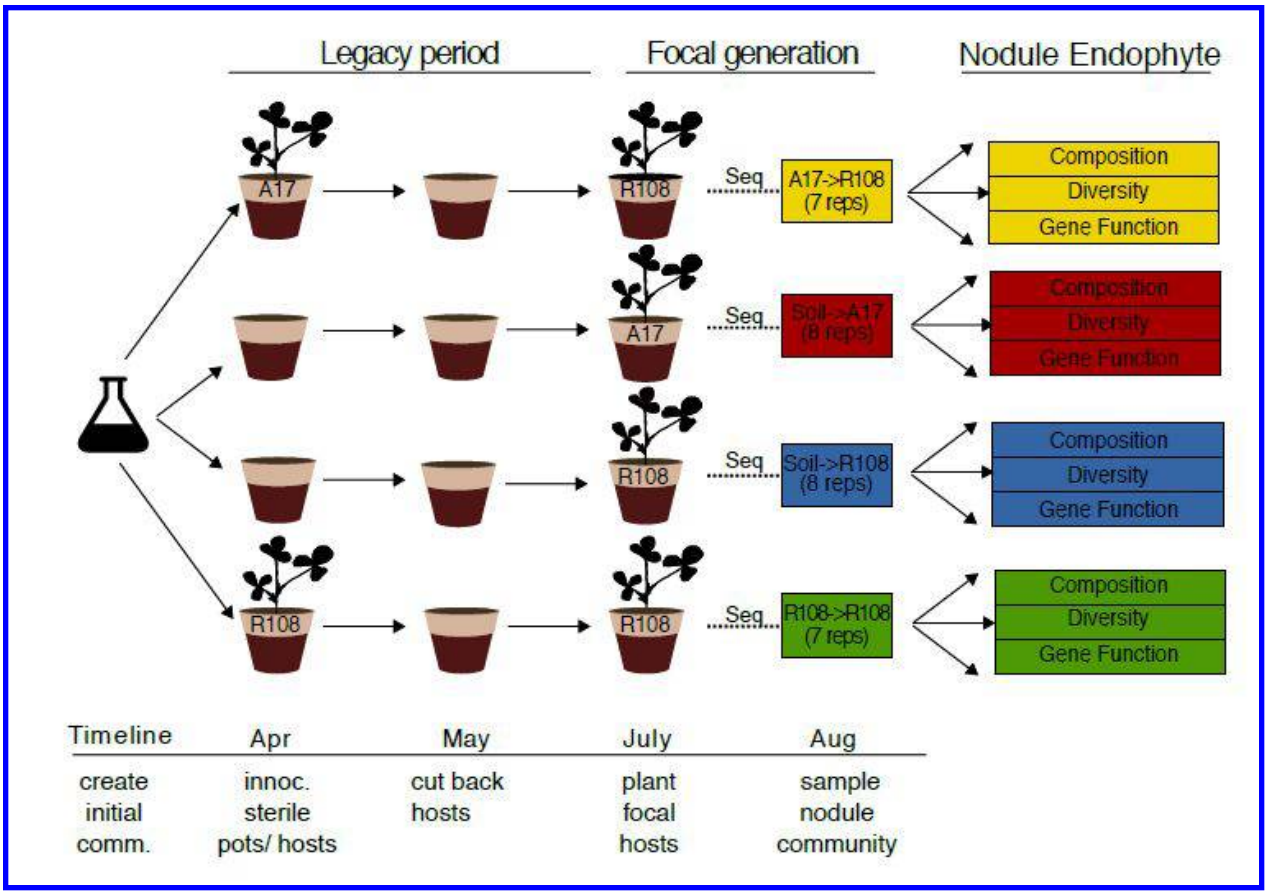


Figure 1: Experimental Design used to test for the direct effect of host genetic variation (Soil→R108 vs. Soil→A17) and host genetic legacies (A17→R108 vs. R108→R108) on nodule endophyte communities. Figure modified from Burghardt et. al. 2019.

As previously described, DNA was extracted from the crushed nodules, sequenced, and aligned reads were mapped to *E. meliloti* USDA1106, and used to assess strain frequencies of *E. meliloti* in nodules (Burghardt et al 2019). In this paper, we used two methods, BLAST and MEGAN6 (described below) to map the unaligned reads (i.e., sequencing bycatch) to non-rhizobial endophytic bacteria.

BLASTing Reads

We used BLAST to search the NCBI database for sequences with high similarity to the non-*Ensifer* and non-*Medicago* nodule reads, then processed the results using custom R and shell scripts (see supplementary materials). We randomly sampled 500,000 reads (out of ~1-4 million total) from each pot replicate that did not align to either the *E. meliloti* or *M.* truncatula reference genomes and used blastn to search bacterial accessions in the NCBI nt database; the maximum number of target sequences was set to 50, and the maximum e-value was set to 10^{-5}. For each read, we retained the highest scoring hit. We filtered the reads further by disposing of all reads for which the best hit was to *Sinorhizobium* (a synonym of *Ensifer*) or unclassified bacteria. which as expected, eliminated approximately 90-97% of reads. After filtering, between 2.0% and 9.5% of the initial 500,000 reads came back with known classified hits to endophytic bacteria other than *Sinorhizobium.* The unmatched reads could be due to the presence of fungi or other microorganisms as well as bacteria without sequences in the NCBI database. We chose the best hit for every read and counted the number of reads with best matches by species, genus, and family (genus names were converted to family using the “taxize” R package). One sample (Soil→R108 replicate 12) was removed as an outlier due to containing several orders of magnitude higher proportion of a single bacterial genus, *Burkholderia,* indicating an apparent infection.

Data Subsets

Before testing for significant differences in endophyte composition, diversity, and function between treatments, we created four datasets. The first two subsets, used for most of the analysis, included all results from the BLAST search. The Soil→A17 and Soil→R108 subset included replicates was used to assess for direct effects of host

genotype. The second subset included all R108 replicates (Soil→R108, A17→R108, and R108→R108) and was used to assess the indirect legacy effects. The soil→A17 and Soil→R108 subset was further narrowed to focus on highly represented genera whose mean abundance was higher than 2% in at least one of the two host genotypes. This subset contained 15 bacterial families that represented 77-88% of the total community of each sample. This subset was used to investigate whether significant differences in composition between treatments were present in high abundance bacteria. An additional data subset of the top 15 bacterial families from all four treatments was used to visualize intra and inter-host endophyte composition.

Community Composition

To visualize differences in endophyte composition across hosts, we used a) non-metric multidimensional scaling (metaMDS function in R with k=2), which measures changes in rank order of genera between different replicates and b) Bray-Curtis dissimilarity distance plots using the “vegdist” and “pco” functions (vegan package in R). We tested for significant differences among host genotypes using a permutational multivariate analysis of variance (PERMANOVA) on the Bray-Curtis distance matrix (“anova” and “adonis” functions in the vegan R package) as well as analyzing multivariate homogeneity of group dispersions (betadisper and anova in the vegan R package) to assess the microbiome consistency across replicates.

Community Diversity Metrics

To determine if the diversity of endophyte communities differed between host genotypes, we quantified community diversity in each replicate using the Renyi diversity index ("renyi" function in the vegan package in R). We used the "anova" function in R to test for differences among host plant treatments at various scale parameters, including species richness (α=0), which only counts the number of species present and Berger-Parker (α=∞), which is maximized only when all species are present at equivalent abundances. We also tested two commonly used intermediate values where both species richness and evenness contribute to the index: exponent of Shannon's diversity (α=1) and Inverse Simpsons indices (α=2).

Automated Taxonomic Pipeline

To confirm the results gathered from the NCBI BLAST data we ran our reads through an independent pipeline. We used DIAMOND (Buchfink et al., 2015), a high-throughput DNA read alignment program, to identify matching sequences in the NCBI nr protein database, and MEGAN6 (Huson et al., 2016), which maps the aligned reads to a taxonomic ID. As with the BLAST analysis, we removed any reads for which the hit was *Sinorhizobium*. After removing *Sinorhizobium* hits, 88-97% of the remaining reads had a match. We also used this pipeline to categorize genes into functional groups (GO biological processes) for comparative analysis between genotypes. Using the "Interpro" function within MEGAN6 we mapped the aligned reads to IPR (Interpro) IDs which we then converted to GO biological processes using an R script (mapping file: ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro2go). Only GO biological processes

with at least 40 hits within a replicate were included. Both the taxa and GO term mapping files can be found on the MEGAN6 website (The July 2019 mapping files were used in this analysis, https://software-ab.informatik.uni-tuebingen.de/download/megan6/welcome.html). In addition, we tested for significant gene function enrichment differences between treatments using a permutational multivariate analysis of variance (PERMANOVA) on a linear model for each functional gene category, then corrected p-values for multiple tests using a Bonferroni correction.

RESULTS

Effect of Host Genotype on Endophyte Community Diversity

Our analysis of the root nodule microbial community identified > 160 bacterial families present in nodules. We found no evidence that endophyte diversity differed between the two host genotypes (Fig. 2a). This was the case both for metrics that emphasize the number of families such as species richness ($F_{df=1}$=0.020, p=0.89) and Shannon diversity ($F_{df=1}$=0.71, p=0.41), as well as those that emphasize community evenness such as Inverse Simpson's ($F_{df=1}$=0.78, p=0.39) and Berger-Parker ($F_{df=1}$=1.21, p=0.29). Thus, there are no significant differences in the diversity of the endophyte communities of the A17 and R108 *M. truncatula* genotypes. This conclusion held at the genus level as well (Fig. S1). We also found no effect of host legacy treatment on endophyte diversity in nodules; species richness (p=0.83), Shannon diversity (p=0.43), Inverse Simpson's (p=0.84), and Berger-Parker (p=0.95) were all non-significant (Fig. 2b).

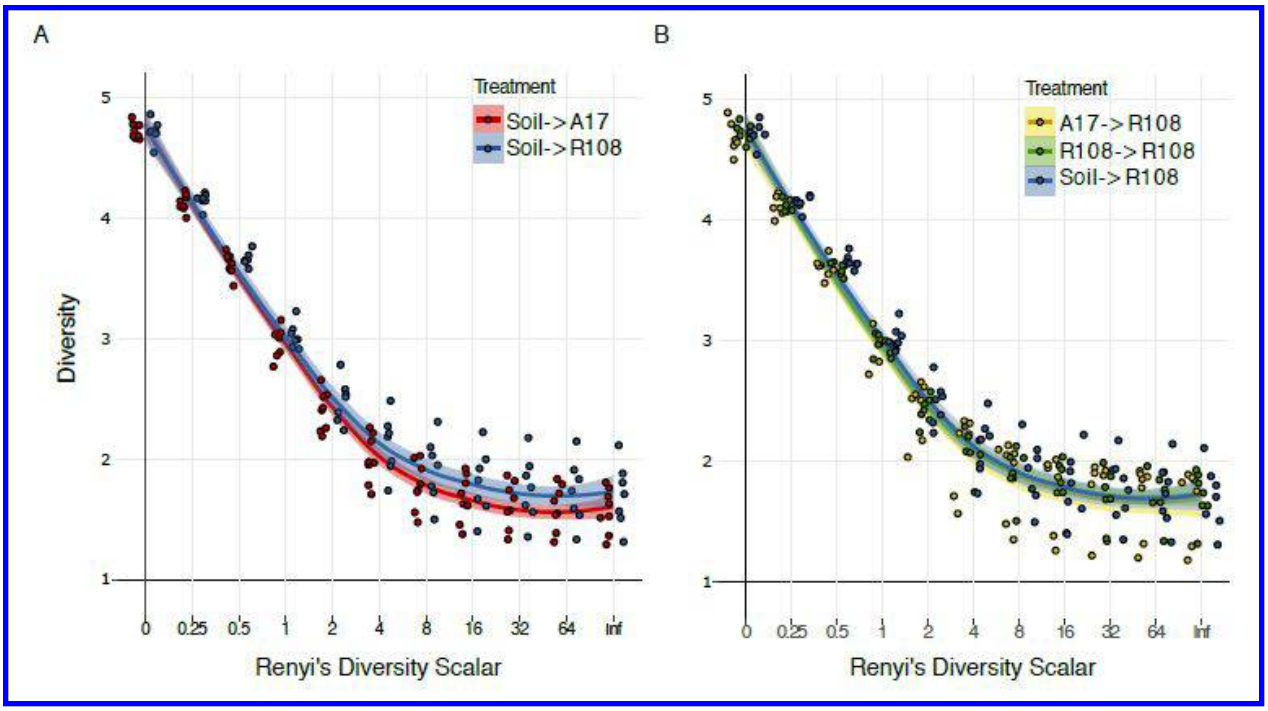


Figure 2: Endophyte community diversity does not differ between (A) hosts or (B) host legacies across a range of scale parameters. Each point represents a replicate pot and a best-fit line is used to summarize trends across scale α parameters. The shaded areas represent a 95% confidence interval. Diversity metrics were calculated on the full family dataset from the blastn analysis.

Compositional Comparison of Endophyte Taxa between Host Genotypes

While at the endophyte family level there are no significant differences in nodule community diversity, there are significant differences in composition between host genotypes (PERMANOVA, $F_{df=1}$=3.1, R^2=0.19, p=0.003, Fig. 3). These differences can be visualized by both non-metric multidimensional scaling (Fig. S2) and principal

coordinate analysis (Fig. 4a) of Bray Curtis dissimilarities. In addition to the compositional differences, there is weak evidence that the A17 replicates (red) are more compositionally consistent than R108 replicates (blue) (analysis of multivariate homogeneity of group dispersions, $F_{df=1}$=3.8, p=0.073). By contrast, at the genus level, there were significant differences between host genotypes (Fig. S3) for both endophyte community composition (p=0.001) as well as compositional consistency (p= 0.033).

Out of more than 160 families (approximately 15% of known bacterial families), the fifteen most abundant (mean frequency > 2% in at least one treatment) in each host genotype collectively comprised more than 80% of the entire dataset. We tested for differences in community composition using just these 15 taxa to determine whether the differences between hosts were reflective of the behavior of abundant families. Consistent with the full dataset, there was a significant, direct effect of host genetic variation on community composition (PERMANOVA on the Bray-Curtis matrix $F_{df=1}$=3.1, R^2=0.19, p=0.005). The bacterial families with the most substantial shifts in abundance between treatments included Xanthomonadaceae, Pseudomonadaceae, and Enterobacteriaceae, all of which were enriched in the Soil→R108 treatment. A larger sample size would be required to make claims about the significance of these shifts, some of which could be attributed to outliers. There was no evidence of a difference between A17 and R108 replicate dispersion ($F_{df=1}$=3.9, p=0.69) (Fig. 3). At the genus level, there were significant differences between host genotypes for both endophyte community composition as well as compositional consistency amongst replicates in the abundant family dataset (Fig. S4). While the most common families detected by the MEGAN6 pipeline differed from those detected by the BLAST pipeline, both analyses

revealed significant differences in community composition and no differences in replicate consistency between host genotypes (Fig. S5 & S6), and the proportion of variance explained by host genotype was similar (R^2_{BLAST}=0.19, R^2_{MEGAN}=0.23).

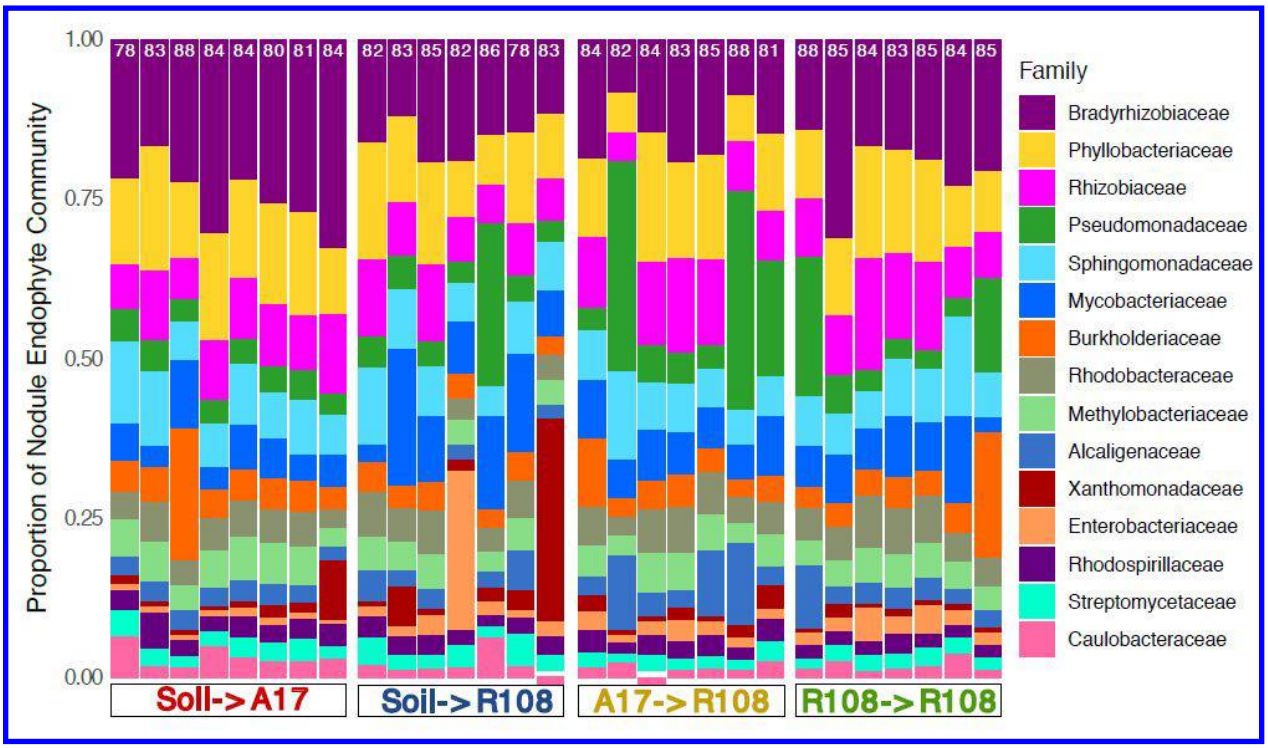
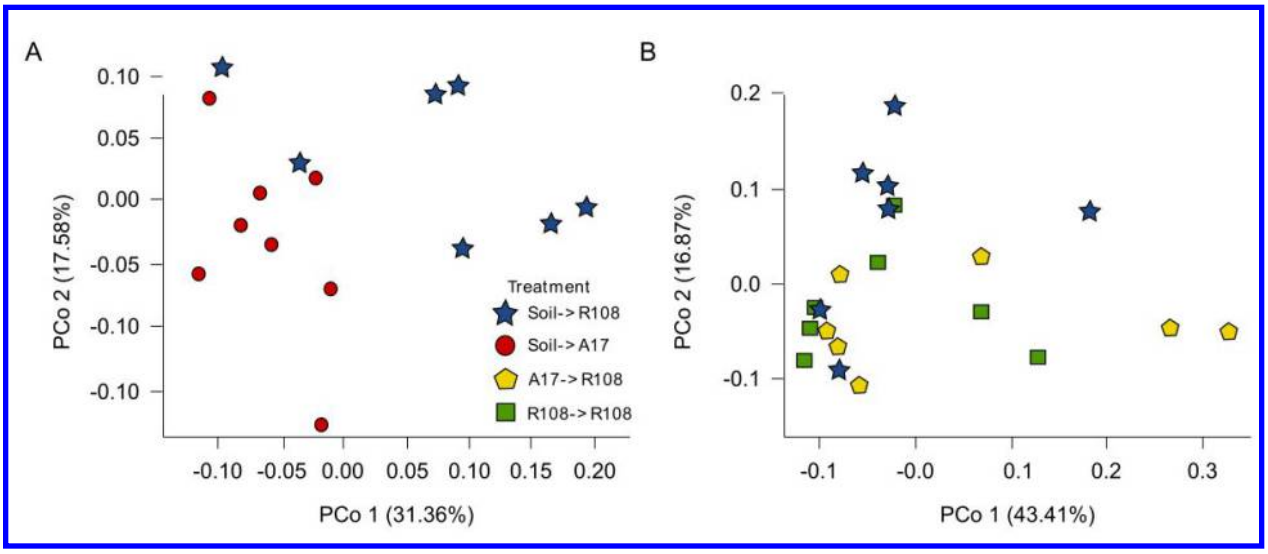


Figure 3: Shifts in community composition of abundant bacterial families in nodules. Stacked bar plot of endophyte composition in nodules from each replicate pot. Families were considered abundant if they occurred with a mean frequency greater than 0.02 in at least one host genotype in the dataset generated from BLAST. Numbers above each bar denote the percentage of all endophytes present represented by these fifteen families.

Because host genotype had a significant effect on endophyte community composition, we also tested whether the identity (A17, R108) or presence of a previous host influenced non-rhizobial endophyte composition. Previously, we found that prior

283 exposure to a host and the genetic identify of that host shifted *Ensifer* strain composition
284 in nodules (Burghardt et al. 2019). Here we find that neither of these legacies effect the
285 non-rhizobial endophyte community at the family level (Fig. 4b; PERMANOVA
286 $F_{df=1}$=1.27, R^2=0.12, *p*=0.23) or the genus level (Fig. S7, $F_{df=1}$=1.31, R^2=0.123,
287 *p*=0.207).



288

289 *Figure 4: Endophyte compositional comparison by Principal Coordinate Analysis*
290 *of Bray-Curtis dissimilarity between (A) host treatments and (B) legacy treatments.*
291 Treatments include Soil→R108 (blue), Soil→A17 (red), A17→R108 (yellow),
292 R108→R108 (green). The entire community dataset of more than 400 genera identified
293 by BLAST was used for this figure Outcomes of statistical tests can be found in Table
294 S1.

295

296 Gene Function Comparison Between Host Endophyte Communities

While compositional differences between host endophyte communities were sometimes evident at the taxonomic level, we also wanted to investigate if there were differences between gene functional categories present in the endophyte communities. Using the relative abundance of GO biological process terms identified with MEGAN6, we found evidence of significant differences in gene functional groups between the Soil→A17 and Soil→R108 endophyte communities (PERMANOVA, $F_{df=1}$=8.48, R^2=0.39, *p*= 0.001) (Fig. 5a). Thus, in addition to differing taxa composition, host genotypic differences have a direct effect on the abundance of functional gene categories in nodules. Interestingly, in contrast to the community composition analysis, host presence/absence in the previous generation had a significant effect on the abundance of functional gene groups (Fig. 5b) Soil→R108 vs. A17→R108 combined with R108→R108 (PERMANOVA, $F_{df=1}$=14.87, R^2=0.44, *p*= 0.001). However, the identity of the prior host genotype had no effect (A17→R108 vs. R108→R108, $F_{df=1}$=1.71, R^2=0.12, *p*= 0.068).

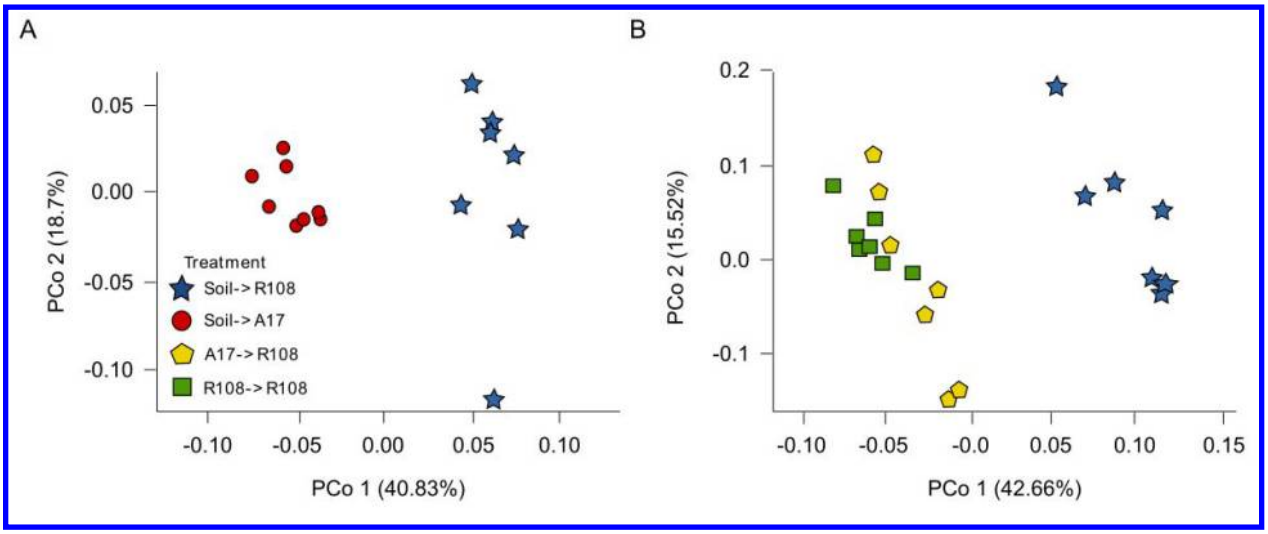


Figure 5: *Functional gene category abundance comparison by Principal Coordinate Analysis of Bray-Curtis dissimilarity between (A) host treatments and (B) legacy treatments.* Treatments include Soil→R108 (blue), Soil→A17 (red), A17→R108 (yellow),

R108→R108 (green). Functional gene categories were identified with MEGAN. Outcomes of statistical tests can be found in Table S1.

In addition to testing for significant differences between the gene functional groups in the two host endophyte communities, we also investigated which of these functional groups were most significantly enriched within a treatment using a permutational multivariate analysis of variance (PERMANOVA) on a linear model for each functional gene category (Table 1). Enriched functions that differ across host genotypes include core energy metabolism, catabolism of aromatic compounds, carbohydrates and pentose sugars and sodium transport. Gene functional enrichment imposed by Soil vs. R108 legacies included DNA repair mechanisms, branched chain amino acid and sulfate transport, and fatty acid and formate biosynthesis. These functional differences between treatments likely depend on a variety of factors, including plant-microbe interactions, microbe-microbe interactions, and nutrient availability in the soil. Because this experiment started from sterile soil in a greenhouse rather than field soil, specific functional differences are difficult to contextualize.

DISCUSSION

We found negligible differences in the diversity of the root nodule endophyte community between two genotypes of *M. truncatula*, even though those two genotypes differ in rhizobial strain diversity (Burghardt et al., 2019). However, the two host genotypes differed significantly in the composition of their root nodule microbiome and in abundance of gene functional groups. By contrast, neither the presence of a prior

host nor the genotypic identity of the prior host effected endophyte diversity, composition, or compositional variation among replicates. However, the presence of a host, regardless of genotype, influenced gene functional abundance. Our results suggest that the direct effect of host genetic differences is a stronger driver of non-rhizobial nodule endophyte communities than enrichment legacies from prior plants.

Comparisons of nodule endophyte communities in different legume species, such as soybeans and alfalfa, have shown that host species does influence root nodule microbiomes (Xiao et al., 2017). However, not all tissues and plant species exhibit intraspecific microbiome variation. For example, *Boechera stricta* genotypes have significantly different microbiome communities in the leaves, but not the roots (Wagner et al., 2016). Using sequencing bycatch from Burghardt et al. 2019, we extend findings from previous studies to show that genetic variation within a legume host species (*M. truncatula*) can also determine root nodule endophyte communities. The reuse of previously generated sequencing by-catch to ask new questions, as done here, has some precedent (e.g. Garcia et al. 2018), but is not common. However, going forward, we believe this method will become increasingly commonplace as the amount of community level, whole-genome sequence data increases.

Within the scope of our experiment, we observed that a few bacterial families drove the compositional differences between the A17 and R108 hosts. Of note, *Pseudomonadaceae* appeared in higher frequency in all three R108 host treatments than in A17. Thus, the R108 genotype may be more susceptible to *Pseudomonadaceae* colonization than A17. Different *Pseudomonadaceae* species have different niches and can be both mutualists and pathogens (Melnyk et al., 2019). Several species have

359 growth promoting properties as well as provide protection from certain parasitic fungi in
360 plants (Haas & Défago, 2005) and *C. elegans* (Zhang et al., 2017). Species level data
361 indicated that *P. aeruginosa*, *P. fluorescens*, *Pseudomonas sp. 31-12*, and *P. putida*
362 were the most prevalent species in our samples, with each being present across
363 treatments. *P. fluorescens* (Hol et al. 2013)*, P. putida* (Costa-Gutierrez et al. 2020*),* and
364 *Pseudomonas sp. 31-12* (Russel et al. 2018) are known to be growth promoting,
365 whereas *P. aeruginosa* is an opportunistic pathogen (Walker et al. 2004). Follow up
366 studies could investigate the prevalence of these species as well as whether their
367 beneficial or detrimental properties in *M. truncatula* align with the same effects in other
368 organisms.

369 Prior hosts can have significant legacy effects on *Ensifer* strain frequencies in
370 root nodules (Burghardt et al., 2019), and on the entire microbial community of later
371 generations (Kardol et al., 2007, Meisner et al., 2013, Li et al. 2019). However, this
372 “legacy” selection does not extend to the nodule endophyte community composition
373 more broadly. In this experiment, we found that neither the presence nor identity of a
374 host in the previous generation significantly affected nodule endophyte community
375 composition in a second generation. While a small sample size, this result indicates
376 that only the present host has a major effect on endophyte community composition in
377 nodules. These compositional differences also extend to differences in the abundance
378 of gene functional groups among plant genotypes. Interestingly, nearly half of the gene
379 functional groups were shown to be expressed significantly differently between hosts,
380 with most being enriched in the Soil→R108 treatment (as compared to the Soil→A17
381 treatment). Previous studies in the rhizosphere of other legumes (Mendes et al., 2014)

382 and non-leguminous plants (Ofek-Lalzar et al., 2014) have shown the importance of
383 membrane transport and nitrogen metabolism within the plant rhizosphere. The
384 exclusion of rhizobia allowed us to focus on functions associated with non-rhizobial
385 nodule endophytes, which could potentially reveal functions important to non-rhizobial
386 endophytes within the rhizosphere and nodules. For instance, the non-oxidative branch
387 of the pentose phosphate shunt is essential for the establishment of some mutualistic
388 plant-microbe relationships (Hawkins et al., 2018) and degradation aromatic acid via the
389 beta-ketoadipate pathway underlies tomato pathogenicity of the tomato wilt pathogen
390 *Fusarium oxysporum* (Michielse et al., 2012). Thus, the enrichment of these functional
391 groups may have consequences for host fitness.

392 Consistent with the findings of Burghardt et al. (2019) for rhizobial community
393 composition, we found that the presence or absence of a host is a stronger driver of
394 functional groups than the genotype of the previous host; it also suggests that functional
395 group differences are not driven solely by differences in family-level composition. Our
396 finding is interesting from an agricultural perspective because of the increasing use of
397 legumes as winter and summer cover crops (Baraibar et al. 2020). Our results suggest
398 that the addition of these legume cover crops to cropping rotation systems could
399 influence the endophyte composition of the subsequent focal legume crops (e.g.,
400 soybeans) (Bakker et al., 2018, Yuan et al., 2018, Lang et al., 2019).

401 Determining which plant host genes allow nodule inhabitation of beneficial
402 mutualistic endophytes has wide ranging implications across a variety of plant-related
403 and agricultural fields, such as increasing crop yield, growth, providing disease
404 resistance, and many other benefits (Cordovez et al., 2019; Dini-Andreote, 2020;

Hubbard et al., 2019; Schlaeppi & Bulgarelli, 2015). Some endeavors are already underway, such as genome-wide association studies (GWAS), which can be used to identify how genetic variation among plant hosts maps to differences in the microbial community (Beilsmith et al., 2019). Analyzing the direct effects on host fitness of specific bacterial species inhabiting the root nodules would allow us to determine which species are plant growth promoting and should be selected for by genetically modified crops and other host plants. For example, co-inoculation of *Rhizobium* with various endophytes such as *Bacillus* under certain conditions results in increased rhizobia proliferation as well as promoted plant growth in leguminous plants (Korir et al., 2017; Petersen et al., 1996). Searching for other similar synergies between endophyte species through co-inoculation experiments could help create optimal inoculation communities for maximal crop production. Along with beneficial strains, legumes have been shown to form nodules with non-mutualistic rhizobia strains (Checcucci et al., 2016), and the same behavior may apply to the nodule endophyte community in general. Once host genes underlying nodule microbiome are identified, optimization via breeding programs or gene editing could be used to engineer plants to resist colonization by pathogens.

ACKNOWLEDGEMENTS

We would like to thank Peter Tiffin for providing his lab resources and for consultation on our manuscript, as well as Maggie R. Wagner and Regina Bledsoe for providing valuable input on early versions of our manuscript. This work was supported by the

National Science Foundation under Grant Numbers, IOS-1724993 and IOS-1856744. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This work was supported by the USDA National Institute of Food and Agriculture and Hatch Appropriations under MIN-71-030 and Project PEN-04-760 Accession #1025611.

REFERENCES

- Bakker, P. A. H. M., Pieterse, C. M. J., de Jonge, R., & Berendsen, R. L. (2018). The soil-borne legacy. Cell, 172(6), 1178–1180.
- Bálint, M., Tiffin, P., Hallström, B., O'Hara, R. B., Olson, M. S., Fankhauser, J. D., Piepenbring, M., & Schmitt, I. (2013). Host genotype shapes the foliar fungal microbiome of balsam poplar (*Populus balsamifera*). *PloS One*, *8*(1), e53987.
- Baraibar, Barbara, Ebony G. Murrell, Brosi A. Bradley, Mary E. Barbercheck, David A. Mortensen, Jason P. Kaye, and Charles M. White. 2020. Cover crop mixture expression is influenced by nitrogen availability and growing degree days. *PloS One* 15 (7): e0235868.
- Beilsmith, K., Thoen, M. P. M., Brachi, B., Gloss, A. D., Khan, M. H., & Bergelson, J. (2019). Genome-wide association studies on the phyllosphere microbiome: Embracing complexity in host-microbe interactions. *The Plant Journal* Vol. 97, Issue 1, pp. 164–181. https://doi.org/10.1111/tpj.14170

- Brown, S. P., Grillo, M. A., Podowski, J. C., and Heath, K. D. 2020. Soil origin and plant genotype structure distinct microbiome compartments in the model legume *Medicago truncatula. Microbiome*. 8:139.
- Bucciarelli, B., Hanan, J., Palmquist, D., & Vance, C. P. (2006). A standardized method for analysis of *Medicago truncatula* phenotypic development. *Plant Physiology*, *142*(1), 207–219.
- Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60.
- Burghardt, L. T., Epstein, B., Guhlin, J., Nelson, M. S., Taylor, M. R., Young, N. D., Sadowsky, M. J., & Tiffin, P. (2018). Select and resequence reveals relative fitness of bacteria in symbiotic and free-living environments. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(10), 2425–2430.
- Burghardt, L. T., Epstein, B., & Tiffin, P. (2019). Legacy of prior host and soil selection on rhizobial fitness in planta. *Evolution; International Journal of Organic Evolution*, *73*(9), 2013–2023.
- Checcucci, A., Azzarello, E., Bazzicalupo, M., Galardini, M., Lagomarsino, A., Mancuso, S., Marti, L., Marzano, M. C., Mocali, S., Squartini, A., Zanardo, M., & Mengoni, A. (2016). Mixed nodule infection in *Sinorhizobium meliloti-Medicago sativa* symbiosis suggest the presence of cheating behavior. *Frontiers in Plant Science*, *7*, 835.

468 • Cloutier, M. L., Murrell, E., Barbercheck, M., Kaye, J., Finney, D., García-
469 González, I., & Bruns, M. A. (2020). Fungal community shifts in soils with varied
470 cover crop treatments and edaphic properties. *Scientific Reports*, 10(1), 6198.
471 • Cordovez, V., Dini-Andreote, F., Carrión, V. J., & Raaijmakers, J. M. (2019).
472 Ecology and Evolution of Plant Microbiomes. *Annual Review of Microbiology*, *73*,
473 69–88.
474 • Costa-Gutierrez, Stefanie B., María Jesús Lami, María Carolina Caram-Di Santo,
475 Ana M. Zenoff, Paula A. Vincent, María Antonia Molina-Henares, Manuel
476 Espinosa-Urgel, and Ricardo E. de Cristóbal. (2020). Plant growth promotion by
477 *Pseudomonas Putida* KT2440 under saline stress: role of eptA. *Applied*
478 *Microbiology and Biotechnology* 104 (10): 4577–92.
479 • Dini-Andreote, F. (2020). Endophytes: the second layer of plant defense. *Trends*
480 *in Plant Science*, *25*(4), 319–322.
481 • Egamberdieva, D., Wirth, S. J., Shurigin, V. V., Hashem, A., & Abd_Allah, E. F.
482 (2017). Endophytic bacteria improve plant growth, symbiotic performance of
483 Chickpea (*Cicer arietinum L.*) and induce suppression of root rot caused by
484 *Fusarium solani* under salt stress. *Frontiers in Microbiology*, Vol. 8.
485 https://doi.org/10.3389/fmicb.2017.01887
486 • Ferreira, L. D. E. V. M., de Vasconcelos Martins Ferreira, L., de Carvalho, F.,
487 Andrade, J. F. C., Oliveira, D. P., de Medeiros, F. H. V., & de Souza Moreira, F.
488 M. (2020). Co-inoculation of selected nodule endophytic rhizobacterial strains
489 with *Rhizobium tropici* promotes plant growth and controls damping off in

490 common bean. In *Pedosphere* Vol. 30, Issue 1, pp. 98–108.
491 https://doi.org/10.1016/s1002-0160(19)60825-8

- 492 Fischer, D.G., Chapman, S.K., Classen, A.T., Gehring, C. A., Grady, K. C.,
493 Schweitzer, J. A., Whitham, T. G. (2014). Plant genetic effects on soils under
494 climate change. *Plant Soil* 379, 1–19.
- 495 Garcia, B. J., Labbé, J. L., Jones, P., Abraham, P. E., Hodge, I., Climer, S.,
496 Jawdy, S., Gunter, L., Tuskan, G. A., Yang, X., Tschaplinski, T. J., & Jacobson,
497 D. A. (2018). Phytobiome and transcriptional adaptation of Populus deltoides to
498 acute progressive drought and cyclic drought. In *Phytobiomes Journal* Vol. 2,
499 Issue 4, pp. 249–260.
- 500 Hawkins, J. P., Ordonez, P. A., & Oresnik, I. J. (2018). Characterization of
501 mutations that affect the nonoxidative Pentose Phosphate pathway in
502 *Sinorhizobium meliloti. Journal of Bacteriology*, 200(2).
- 503 Haas, D., & Défago, G. (2005). Biological control of soil-borne pathogens by
504 *fluorescent pseudomonads*. *Nature Reviews. Microbiology*, *3*(4), 307–319.
- 505 Heath KD, Grillo MA. (2016). Rhizobia: tractable models for bacterial evolutionary
506 ecology. *Environmental Microbiology* 18: 4307–4311.
- 507 Heath KD, Stinchcombe J. (2013). Explaining mutualism variation: a new
508 evolutionary paradox? *Evolution* 68: 309–317.
- 509 Hol, W H Gera, Bezemer, T Martijn, & Biere, Arjen. (2013). Getting the ecology
510 into interactions between plants and the plant growth-promoting bacterium
511 *Pseudomonas fluorescens*. *Frontiers in Plant Science*, 4, 81–81.

512 • Hubbard, C. J., Li, B., McMinn, R., Brock, M. T., Maignien, L., Ewers, B. E.,
513 Kliebenstein, D., & Weinig, C. (2019). The effect of rhizosphere microbes
514 outweighs host plant genetics in reducing insect herbivory. *Molecular Ecology*,
515 *28*(7), 1801–1811.
516 • Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S.,
517 Ruscheweyh, H.-J., & Tappu, R. (2016). MEGAN Community Edition - interactive
518 xxploration and analysis of large-scale microbiome sequencing data. In *PLOS*
519 *Computational Biology*, Vol. 12, Issue 6, p. e1004957.
520 https://doi.org/10.1371/journal.pcbi.1004957
521 • Hynes, Russell K., Tim J. Dumonceaux, Jakkrapong Kangsopa, and Jennifer R.
522 Town. (2018). Genome sequence of a plant growth-promoting *Rhizobacterium*,
523 *Pseudomonas Sp. Strain 31-12*. *Microbiology Resource Announcements*, 7 (6).
524 • Ikeda, A. C., Bassani, L. L., Adamoski, D., Stringari, D., Cordeiro, V. K., Glienke,
525 C., Steffens, M. B. R., Hungria, M., & Galli-Terasawa, L. V. (2013). Morphological
526 and genetic characterization of endophytic bacteria isolated from roots of
527 different maize genotypes. *Microbial Ecology*, *65*(1), 154–160.
528 • Kardol, P., Cornips, N. J., van Kempen, M. M. L., Tanja Bakx-Schotman, J. M., &
529 van der Putten, W. H. (2007). Microbe-mediated plant-soil feedback causes
530 historical contingency effects in plant community assembly. In *Ecological*
531 *Monographs* (Vol. 77, Issue 2, pp. 147–162).
532 • Korir, H., Mungai, N. W., Thuita, M., Hamba, Y., & Masso, C. (2017). Co-
533 inoculation effect of rhizobia and plant growth promoting Rhizobacteria on

534 common bean growth in a low Phosphorus soil. In *Frontiers in Plant Science*
535 (Vol. 08). https://doi.org/10.3389/fpls.2017.00141
536 • Lang, M., Bei, S., Li, X., Kuyper, T. W., & Zhang, J. (2019). Rhizoplane bacteria
537 and plant species co-determine Phosphorus-mediated microbial legacy effect.
538 *Frontiers in Microbiology*, 10, 2856.
539 • Leite, Jakson, Doreen Fischer, Luc F. M. Rouws, Paulo I. Fernandes-Júnior,
540 Andreas Hofmann, Susanne Kublik, Michael Schloter, Gustavo R. Xavier, and
541 Viviane Radl. (2016). Cowpea nodules harbor non-rhizobial bacterial
542 communities that are shaped by soil type rather than plant genotype. *Frontiers in*
543 *Plant Science* 7: 2064.
544 • Li, X., Jousset, A., de Boer, W., Carrión, V. J., Zhang, T., Wang, X., & Kuramae,
545 E. E. (2019). Legacy of land use history determines reprogramming of plant
546 physiology by soil microbiome. *The ISME Journal*, 13(3), 738–751.
547 • Long, S. R. (1989). Rhizobium-legume nodulation: life together in the
548 underground. *Cell*, *56*(2), 203–214.
549 • Martínez-Hidalgo, P., & Hirsch, A. M. (2017). The nodule microbiome: N2-fixing
550 rhizobia do not live alone. In *Phytobiomes Journal*, Vol. 1, Issue 2, pp. 70–82.
551 • Meisner, A., De Deyn, G. B., de Boer, W., & van der Putten, W. H. (2013). Soil
552 biotic legacy effects of extreme weather events influence plant invasiveness.
553 *Proceedings of the National Academy of Sciences of the United States of*
554 *America*, 110(24), 9835–9838.

555 • Melnyk, R. A., Hossain, S. S., & Haney, C. H. (2019). Convergent gain and loss
556 of genomic islands drive lifestyle changes in plant-associated *Pseudomonas*. *The*
557 *ISME Journal*, 13(6), 1575–1588.
558 • Mendes, L. W., Kuramae, E. E., Navarrete, A. A., van Veen, J. A., & Tsai, S. M.
559 (2014). Taxonomical and functional microbial community selection in soybean
560 rhizosphere. *The ISME Journal*, 8(8), 1577–1587.
561 • Mhamdi, R., Mrabet, M., Laguerre, G., Tiwari, R., & Aouani, M. E. (2005).
562 Colonization of *Phaseolus vulgaris* nodules by Agrobacterium-like strains.
563 *Canadian Journal of Microbiology*, *51*(2), 105–111.
564 • Michielse, C. B., Reijnen, L., Olivain, C., Alabouvette, C., & Rep, M. (2012).
565 Degradation of aromatic compounds through the β-ketoadipate pathway is
566 required for pathogenicity of the tomato wilt pathogen *Fusarium oxysporum f. sp.*
567 *lycopersici*. *Molecular Plant Pathology*, 13(9), 1089–1100.
568 • Morella, Norma M., Francis Cheng-Hsuan Weng, Pierre M. Joubert, C. Jessica
569 E. Metcalf, Steven Lindow, and Britt Koskella. 2020. “Successive Passaging of a
570 Plant-Associated Microbiome Reveals Robust Habitat and Host Genotype-
571 Dependent Selection.” *Proceedings of the National Academy of Sciences of the*
572 *United States of America* 117 (2): 1148–59.
573 • Mrabet, M., Mnasri, B., Romdhane, S. B., Laguerre, G., Aouani, M. E., &
574 Mhamdi, R. (2006). Agrobacterium strains isolated from root nodules of common
575 bean specifically reduce nodulation by *Rhizobium gallicum*. *FEMS Microbiology*
576 *Ecology*, *56*(2), 304–309.

- Ofek-Lalzar, M., Sela, N., Goldman-Voronov, M., Green, S. J., Hadar, Y., & Minz, D. (2014). Niche and host-associated functional signatures of the root surface microbiome. *Nature Communications*, 5, 4950.
- Petersen, D. J., Srinivasan, M., & Chanway, C. P. (1996). Bacillus polymyxa stimulates increased *Rhizobium etli* populations and nodulation when co-resident in the rhizosphere of *Phaseolus vulgaris*. *FEMS Microbiology Letters*, *142*(2-3), 271–276.
- Schlaeppi, K., & Bulgarelli, D. (2015). The plant microbiome at work. *Molecular Plant-Microbe Interactions: MPMI*, *28*(3), 212–217.
- Sharaf, Hazem, Richard R. Rodrigues, Jinyoung Moon, Bo Zhang, Kerri Mills, and Mark A. Williams. (2019). Unprecedented Bacterial Community Richness in Soybean Nodules Vary with Cultivar and Water Status. *Microbiome* 7 (1): 63.
- Wagner, M. R., Lundberg, D. S., Del Rio, T. G., Tringe, S. G., Dangl, J. L., & Mitchell-Olds, T. (2016). Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications*, *7*, 12151.
- Walker, T. S., Bais, H. P., Déziel, E., Schweizer, H. P., Rahme, L. G., Fall, R., & Vivanco, J. M. (2004). *Pseudomonas aeruginosa*-plant root interactions. Pathogenicity, biofilm formation, and root exudation. *Plant physiology*, 134(1)
- Xiao, Xiao, X., Chen, W., Zong, L., Yang, J., Jiao, S., Lin, Y., Wang, E., & Wei, G. (2017). Two cultivated legume plants reveal the enrichment process of the microbiome in the rhizocompartments. In *Molecular Ecology* (Vol. 26, Issue 6, pp. 1641–1651).

599 • Yuan, J., Zhao, J., Wen, T., Zhao, M., Li, R., Goossens, P., Huang, Q., Bai, Y.,
600 Vivanco, J. M., Kowalchuk, G. A., Berendsen, R. L., & Shen, Q. (2018). Root
601 exudates drive the soil-borne legacy of aboveground pathogen infection.
602 Microbiome, 6(1), 156.
603 • Zgadzaj, R., Garrido-Oter, R., Jensen, D. B., Koprivova, A., Schulze-Lefert, P., &
604 Radutoiu, S. (2016). Root nodule symbiosis in *Lotus japonicus* drives the
605 establishment of distinctive rhizosphere, root, and nodule bacterial communities.
606 *Proceedings of the National Academy of Sciences of the United States of*
607 *America*, *113*(49), E7996–E8005.
608 • Zhang, F., Berg, M., Dierking, K., Félix, M.-A., Shapira, M., Samuel, B. S., &
609 Schulenburg, H. (2017). C*aenorhabditis elegans* as a Model for Microbiome
610 Research. In *Frontiers in Microbiology* (Vol. 8).

611

612

613

614

615

616

617

618 Table1: *Functional gene category enrichment from Megan6 taxon identification pipeline*

A17 vs R108 Enrichment			Soil vs Host Legacy Enrichment		
Gene Function	p-value	Fold Change	Gene Function	p-value	Fold Change
beta-ketoadipate pathway	0.000084	0.629	ATP synthesis coupled proton transport	0.000053	2.18
translation	0.00015	1.48	fatty acid biosynthetic process	0.000117	1.41
carbohydrate catabolic process	0.000162	1.8	DNA repair	0.000143	1.41
ATP synthesis coupled proton transport	0.000646	1.88	sulfate transport	0.000187	1.94
mitochondrial electron transport, ubiquinol to cytochrome c	0.00624	4.61	oxidation-reduction process	0.000253	1.31
5-phosphoribose 1-diphosphate biosynthetic process	0.00749	2.08	amino acid transmembrane transport	0.000347	1.33
pentose-phosphate shunt, non-oxidative branch	0.0102	0.667	base-excision repair	0.000413	1.5
ATP synthesis coupled electron transport	0.012	4.17	branched chain amino acid biosynthetic process	0.000436	3.67
phosphorelay signal transduction system	0.0139	1.41	translation	0.00166	1.63
sodium ion transport	0.0146	0.641	formate metabolic process	0.00357	1.41

619 Notes: The ten most enriched functional gene groups between treatments. Soil →A17
620 vs Soil → R108 is shown on the left and Host → R108 vs Soil → R108 is shown on the
621 right. All fold changes display the enrichment in the Soil → R108 treatment. The p-
622 values displayed have been adjusted using a Bonferroni correction.

623

624

625

626

627

Figure 1: Experimental Design used to test for the direct effect of host genetic variation (Soil → R108 vs. Soil→A17) and host genetic legacies (A17 →R108 vs. R108→R108) on nodule endophyte communities. Figure modified from Burghardt et. al. 2019.

Figure 2: Endophyte community diversity does not differ between (A) hosts or (B) host legacies across various scale parameters. Each point represents a pot replicate, and we used a best-fit line to summarize trends across scale α parameters. The shaded areas represent a 95% confidence interval. Diversity metrics were calculated on the full family dataset from the blastn analysis. Diversity metrics were calculated on the full family dataset from the blastn analysis.

Figure 3: Shifts in community composition of abundant bacterial families in nodules. Stacked bar plot of endophyte composition in nodules from each replicate pot. Families were considered abundant if they occurred with a mean frequency greater than 0.02 in at least one host genotype in the dataset generated from BLAST. Numbers above each bar denote the percentage of all endophytes present represented by these fifteen families.

Figure 4: Endophyte compositional comparison by Principal Coordinate Analysis of Bray-Curtis dissimilarity between (A) host treatments and (B) legacy treatments. Treatments include Soil→R108 (blue), Soil→A17 (red), A17→R108 (yellow),

649 R108→R108 (green). The entire community dataset of more than 400 genera identified
650 by BLAST was used for this figure Outcomes of statistical tests can be found in Table
651 S1.

652

653 Figure 5: *Functional gene category abundance comparison by Principal Coordinate*
654 *Analysis of Bray-Curtis dissimilarity between (A) host treatments and (B) legacy*
655 *treatments.* Treatments include Soil→R108 (blue), Soil→A17 (red), A17→R108 (yellow),
656 R108→R108 (green). Functional gene categories were identified with MEGAN.
657 Outcomes of statistical tests can be found in Table S1.

658

659

660

661

662

663

664

Supplemental Material for:

Title: Comparison of nodule endophyte composition, diversity, and gene content between *Medicago truncatula* genotypes

Authors: Mannix Burns[1], Brendan Epstein[1], Liana T. Burghardt[1,2]*

Affiliations:

[1]Department of Plant and Microbial Biology, University of Minnesota, St. Paul, MN 55108

[2]current affiliation: Department of Plant Science, The Pennsylvania State University, University Park, PA

Supplemental Figures:

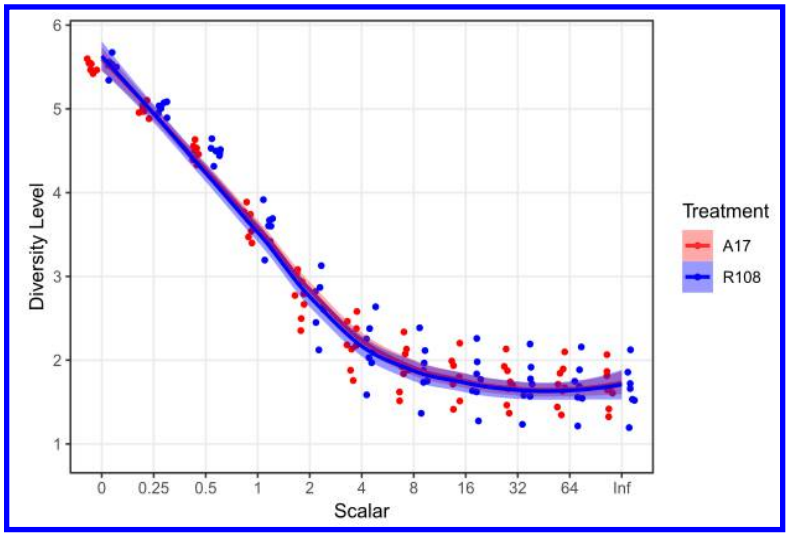


Figure S1: Endophyte community diversity does not differ between hosts. Genera were identified by using BLAST. Each point represents a replicate pot and a best-fit line is used to summarize trends across scale (α) parameters. The shaded areas represent a 95% confidence interval. The Soil→A17 and Soil→R108 communities follow similar trends with little evidence of differences in diversity with metrics including species richness (α= 0, p=0.59), Shannon diversity (α=1, p=0.24), Inverse Simpson's (α= 2, p=0.23), and Berger-Parker (α= inf., p=0.41).

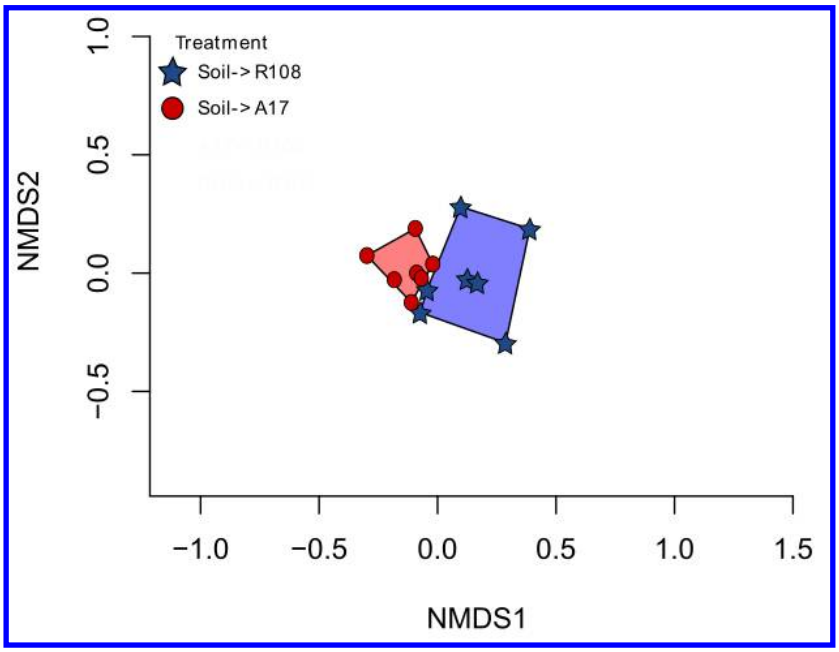


Figure S2: Non-metric multidimensional scaling of host endophyte communities at the family level. Graphical representation of the changes in rank order between host replicates using data identified with BLAST. Treatments include Soil→A17 (red) and Soil→R108 (blue).

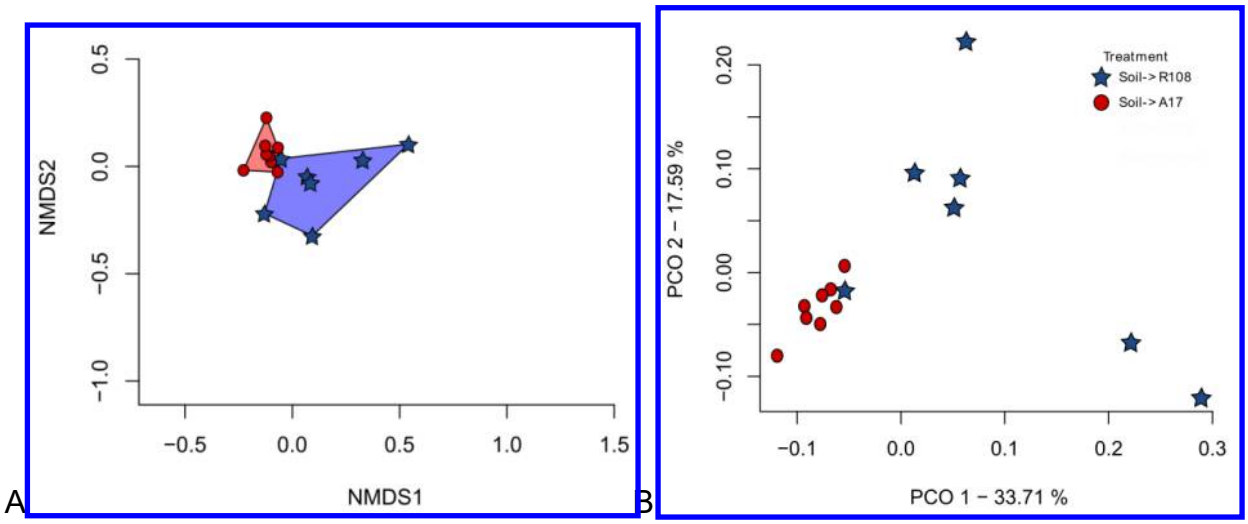


Figure S3: Endophyte composition differs between genotypes at the genus level

(A) Non-metric multidimensional scaling of host endophyte communities provides a graphical representation of the changes in rank order between host replicates. (B) Principal Coordinate Analysis of Bray-Curtis dissimilarity between replicates. The entire community dataset of more than 400 genera identified by BLAST was used for this figure, with Soil→A17 shown in red and Soil→R108 in blue. Compositional differences were quantified with PERMANOVA ($F_{df=1}$=3.9, R^2=0.23, p=0.001) and compositional clustering with analysis of multivariate homogeneity ($F_{df=1}$=5.7, p= 0.033). Both results were significant at the genus level.

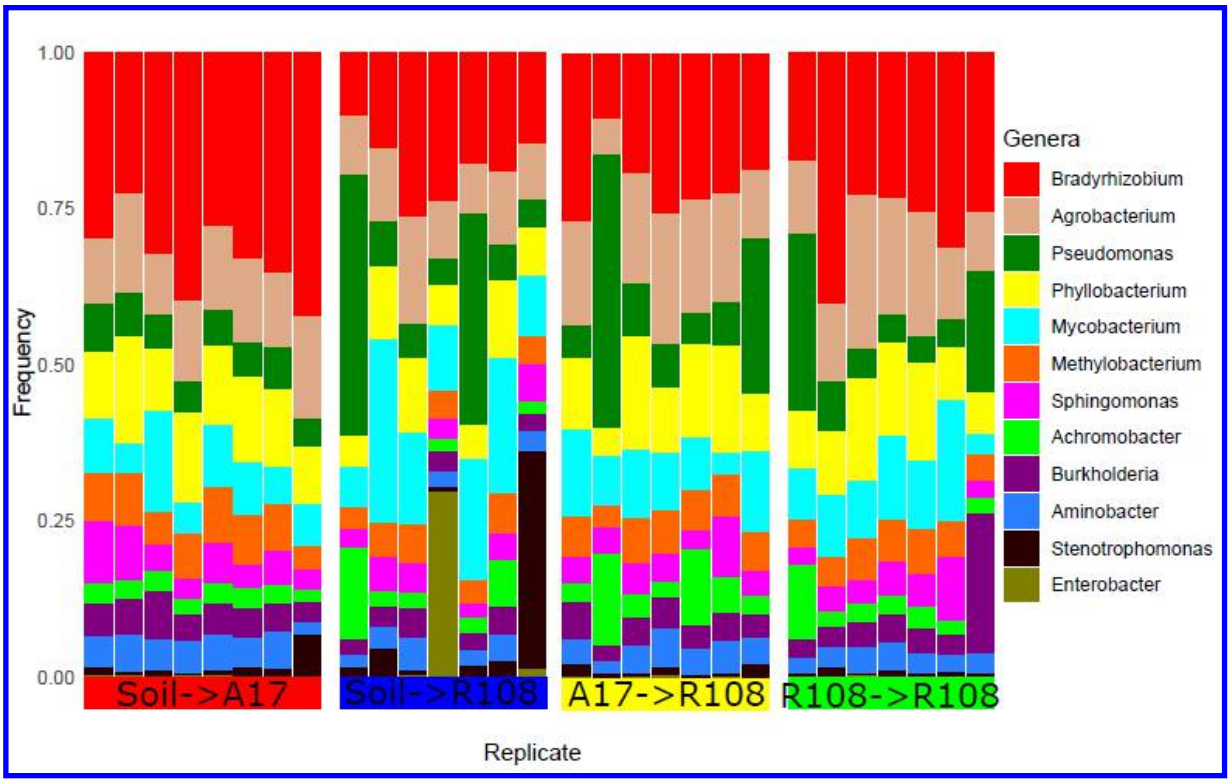


Figure S4: Shifts in community composition of common genera at the genus level from BLAST dataset. Stacked bar plot of the twelve most dominant genera between treatments. Comparison between the Soil→A17 and Soil→R108 twelve genera dataset indicated significant differences in composition ($F_{df=1}$=4.5, R^2=0.26, p=0.001) and compositional clustering ($F_{df=1}$=8.8, p= 0.011).

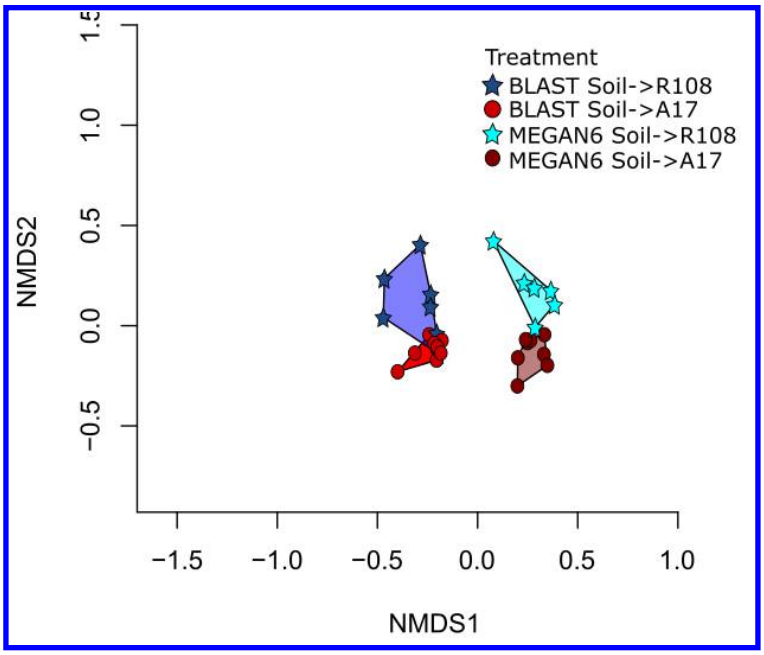


Figure S5: Comparison of non-metric multidimensional scaling of host endophyte communities. Comparison between BLAST and the DIAMOND/MEGAN6 taxonomic pipeline at the family level. Shown is a graphical representation of the changes in rank order between host replicates. Treatments include BLAST Soil→A17 (red), BLAST Soil→R108 (blue), and DIAMOND/MEGAN6 Soil→A17 (maroon), and DIAMOND/MEGAN6 Soil→R108 (sky blue). The abundant family data set was used to create this figure. Similarities in the relationship between the treatments in both pipelines is visualized.

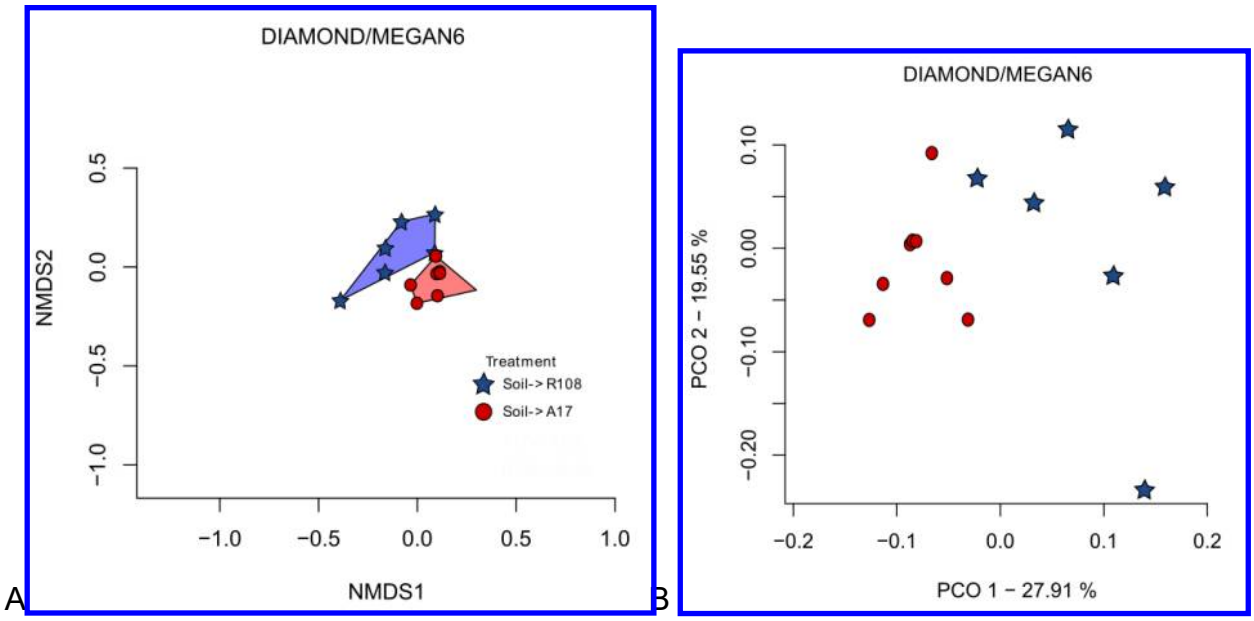


Figure S6: DIAMOND/MEGAN6 endophyte composition at the family level differs between genotypes (A) Non-metric multidimensional scaling of host endophyte communities provides a graphical representation of the changes in rank order between host replicates. (B) Principal Coordinate Analysis of Bray-Curtis dissimilarity between replicates. The entire community dataset was used for this figure. Treatments include Soil→A17 (red) and Soil→R108 (blue). Analysis was done using the automated taxonomic DIAMOND/MEGAN6 pipeline described in the methods. Compositional differences were quantified with PERMANOVA ($F_{df=1}$=3.8, R2=0.23, p=0.001) and compositional clustering with analysis of multivariate homogeneity ($F_{df=1}$=1.4, p= 0.26). These results were similar to the family level results using BLAST (Fig. 3).

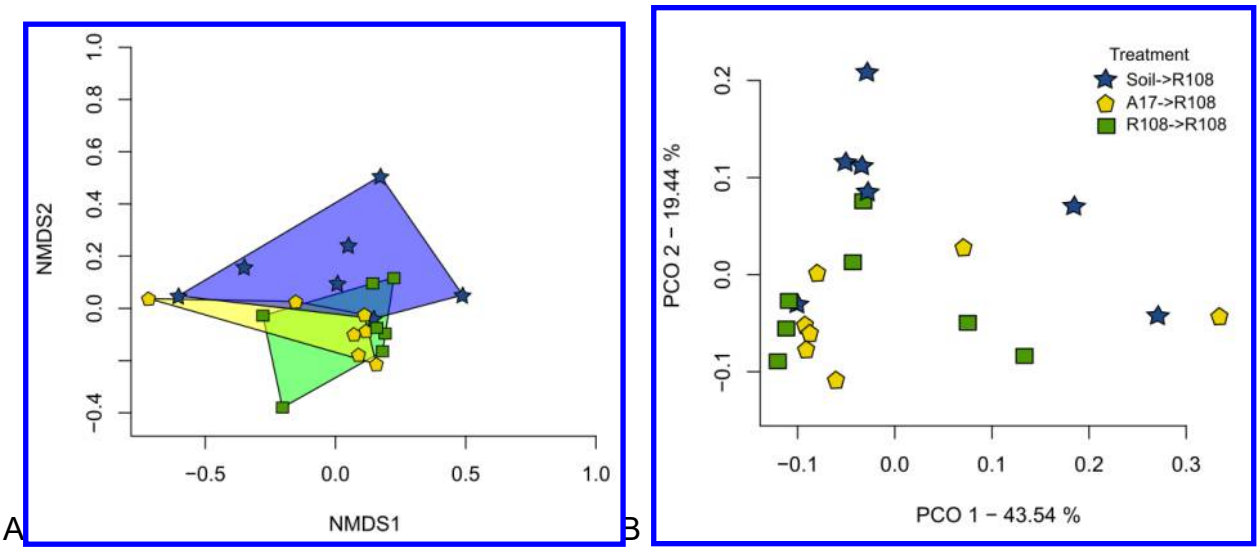


Figure S7: Endophyte composition does not differ between genotypes in legacy setting at the genus level (A) Non-metric multidimensional scaling of host endophyte communities. Shown is a graphical representation of the changes in rank order between host replicates. (B) Principal Coordinate Analysis of Bray-Curtis dissimilarity between second generation R108 replicates. The complete genera data set (from BLAST) was used to create this figure. Treatments include Soil→R108 (blue), A17→R108 (yellow), and R108→R108 (green). Compositional differences were quantified with PERMANOVA ($F_{df=1}$=1.31, R^2=0.123, p=0.207) and compositional clustering with analysis of multivariate homogeneity ($F_{df=1}$=0.695, p= 0.512). Neither result was significant at the genus level.

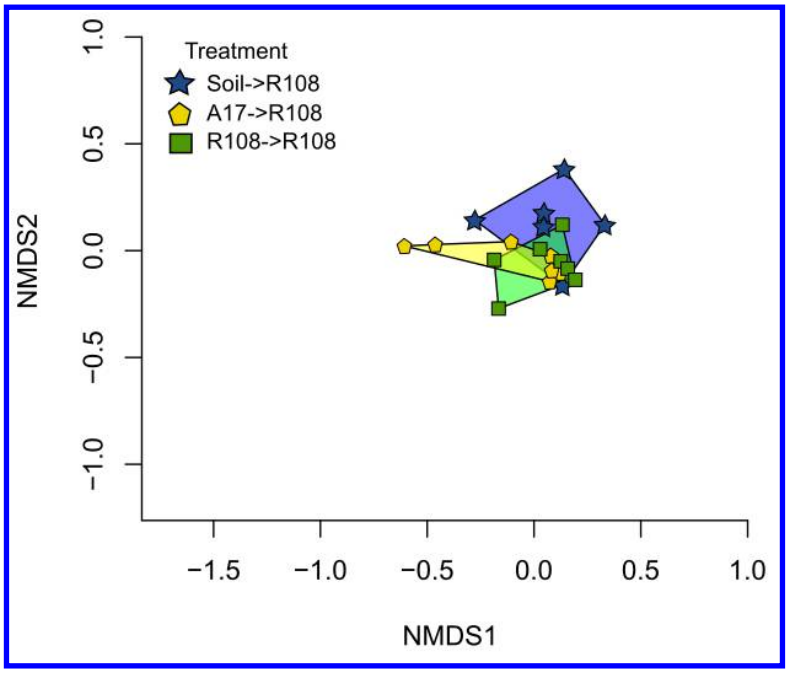


Figure S8: Non-metric multidimensional scaling of legacy host endophyte communities at the family level. Treatments include Soil Soil→ R108 (blue), A17→R108 (yellow), and R108→R108 (green). The complete family BLAST data set was used to create this figure.

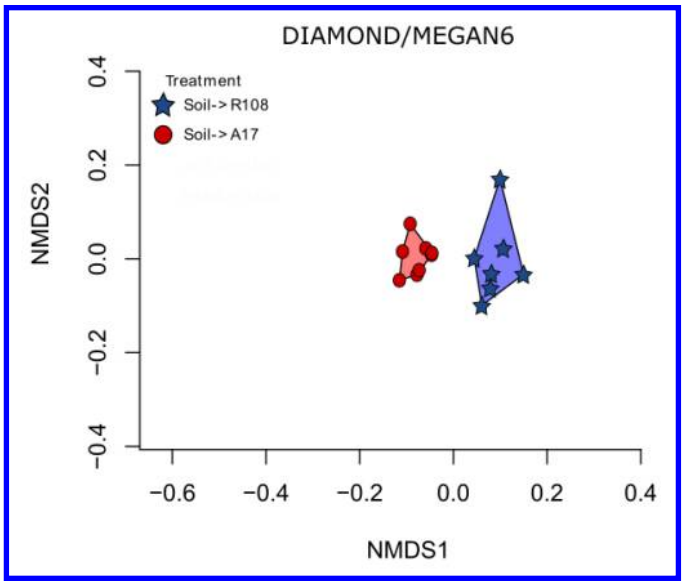


Figure S9: Functional gene category abundance differs between host genotypes. Non-metric multidimensional scaling of host endophyte gene functions. Shown is a graphical representation of the changes in rank order of functional gene categories between host replicates. Treatments include Soil→ A17 (red) and Soil→ R108 (blue).

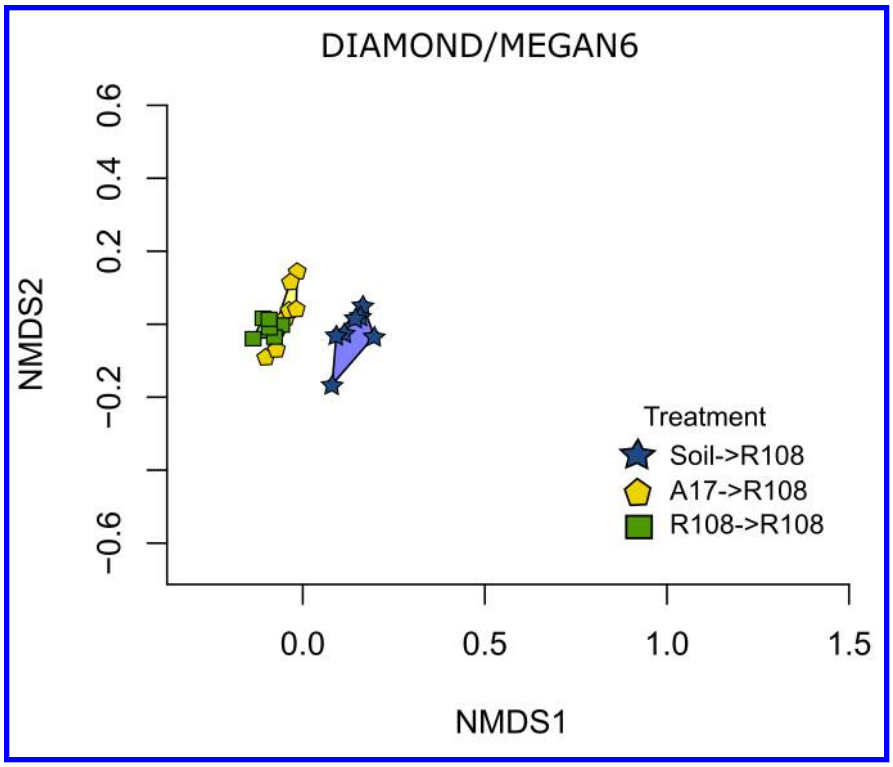


Figure S10: Functional gene category abundance differs between host genotypes in a legacy setting. Non-metric multidimensional scaling of host endophyte gene functions. Shown is a graphical representation of the changes in rank order of functional gene categories between host replicates. Treatments include Soil→ R108 (blue), A17→ R108 (yellow), and R108→ R108 (green). The complete family data set was used to create this figure.

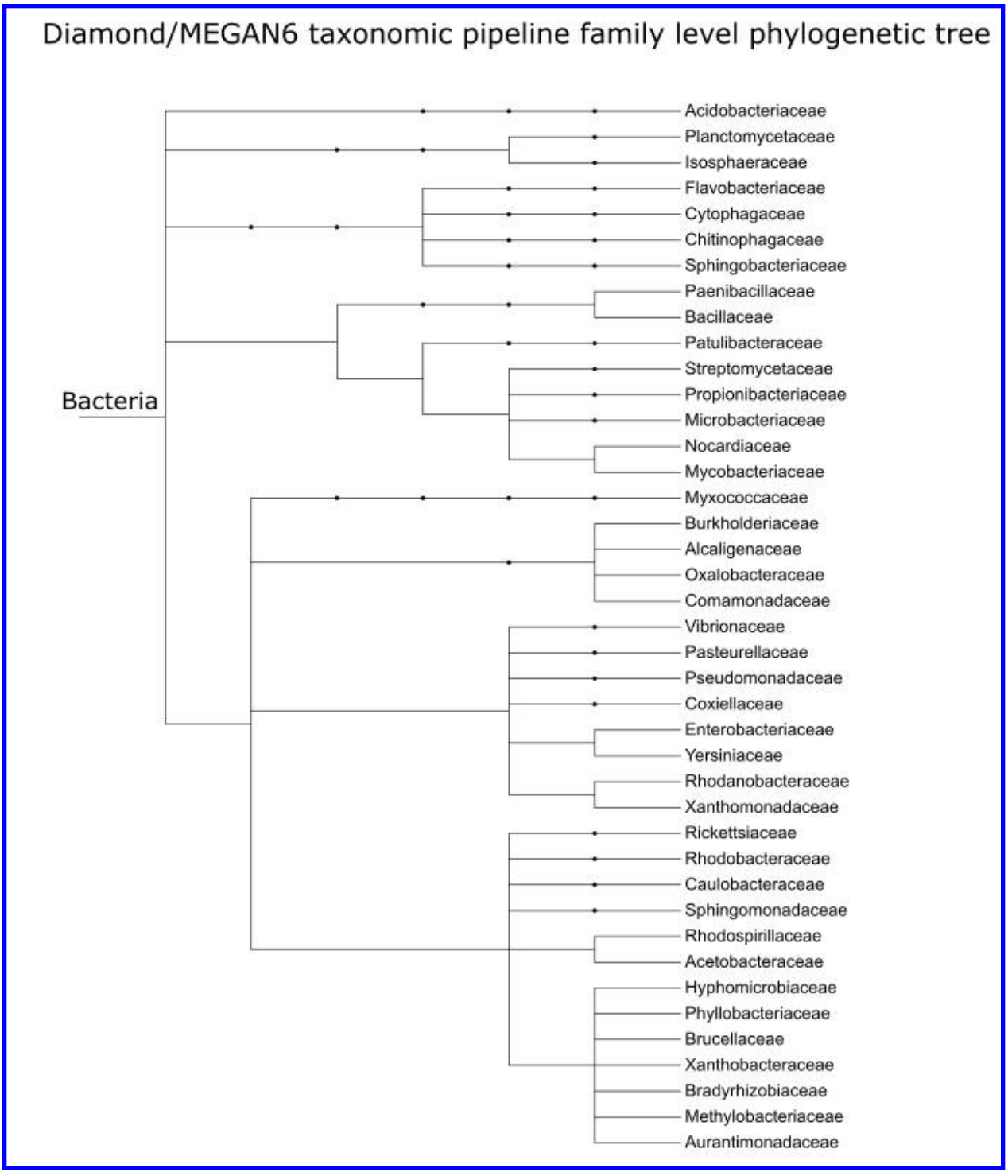


Figure S11: Phylogenetic tree for the taxa present in the MEGAN taxonomic pipeline at the family level. When not displayed as a branching point, nodes are denoted by a point.

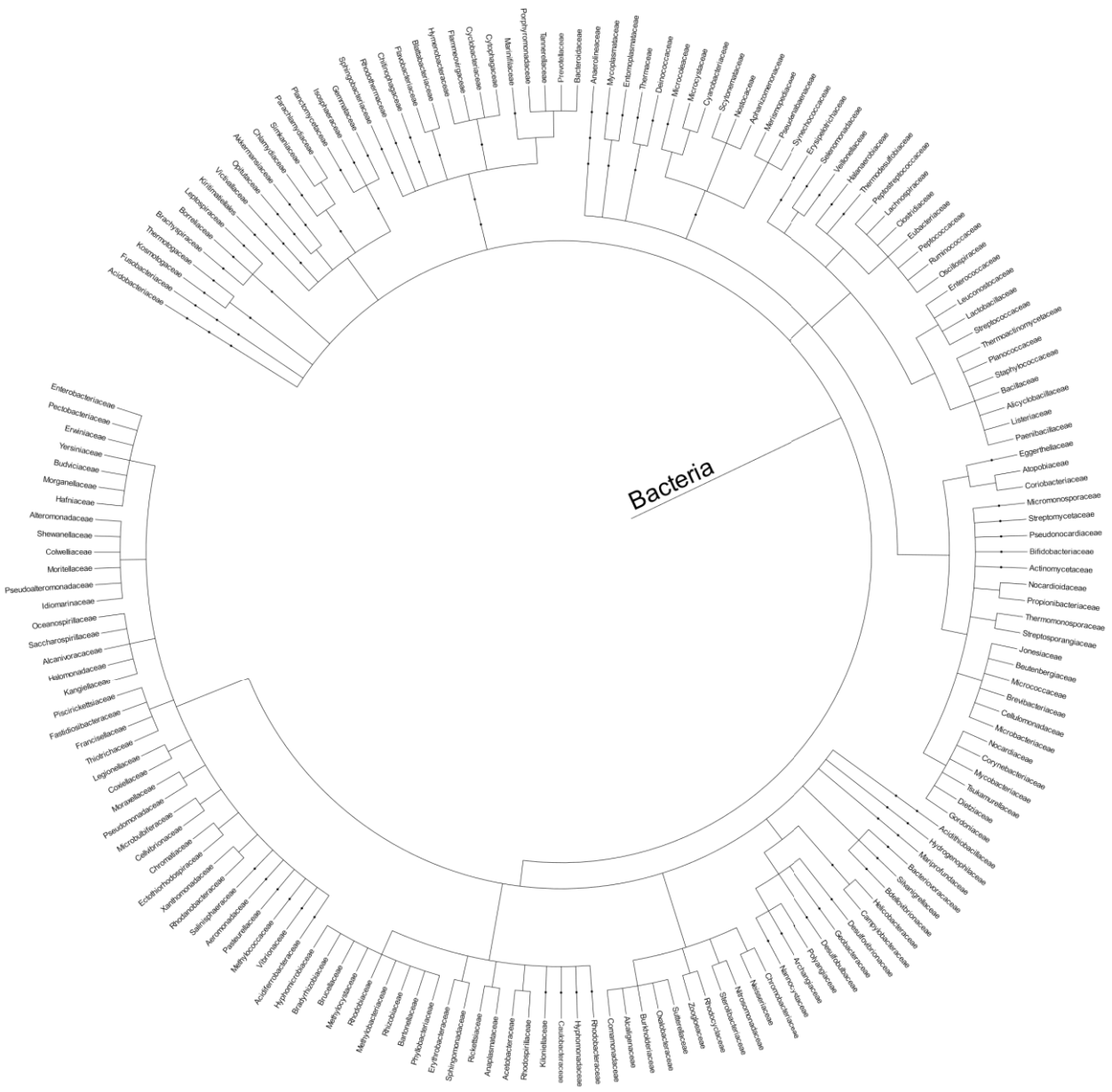


Figure S12: Phylogenetic tree for the taxa present in the BLAST analysis at the family level. When not displayed as a branching point, nodes are denoted by a point.

Table S1: *Summary of statistical analysis at the family level*

Treatment comparison	Permanova	Multivariate homogeneity of group dispersions
A17 vs R108 family composition	**0.003**	0.073
A17 vs R108 genus composition	**0.001**	**0.033**
A17 vs R108 family subset composition	**0.005**	0.69
Legacy family composition	0.23	0.74
A17 vs R108 gene function composition*	**0.001**	-
Legacy gene function composition*	**0.001**	-

Notes: The listed values are p-values. Legacy comparisons were global, not pairwise. Family and genus summary statistics were done using the BLAST method. Gene function summary statistics were done using the MEGAN method (denoted by *).

###BLAST shell script contained on pages 1-5, BLAST top hits R script on pages 6-7, BLAST taxa names R script on page 8###

```
#!/bin/bash

###BASH shell, BLASTS reads against bacteria and fungi databases (only bacteria was used for analysis) and runs them through the blast loop script and blast names script###

WALLTIME="72:00:00"

QUEUE="small"

NTHREADS=20

#

declare -A N_READS_SAMPLED      # Number of unaligned reads to randomly sample

N_READS_SAMPLED[bacteria]=500000

N_READS_SAMPLED[fungi]=100000

MAX_TARGET_SEQS=50            # -max_target_seqs option to blastn

EVALUE=1e-5                   # Evalue cutoff for BLAST

OUTFMT="6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore sscinames stitle"

TAXA=("bacteria" "fungi")


PROJDIR="${HOME}/Path_to_project_directoryproject"

INDIR="${PROJDIR}/Path_to_input_directoryresults"

OUTDIR="${PROJDIR}/Path_to_output_directoryresults"

GIDIR="${PROJDIR}/data/blast"


WORKDIR="${OUTDIR}/working"

LOGDIR="${OUTDIR}/log"

ARRAYDIR="${OUTDIR}/arrayjobdata"

SCRIPTDIR="${OUTDIR}/script_copies"

SCRATCHDIR="/scratch.global/${USER}/Path_to_scratch_directory"
```

```
tophit="${PROJDIR}/script/mannix/blast_script_loop_MB.R" # script to extract top hits
tabulate="${PROJDIR}/script/mannix/blast_names_loopMB.R"

if [[ "$PBS_ENVIRONMENT" == "PBS_BATCH" ]]; then

   newgrp "tiffinp"
   source "${HOME}/.bashrc"
   source "${HOME}/bin/init-modules.sh"

   module load ncbi_blast+/2.7.1.CentOS7
   module load R/3.5.0
   module load samtools/1.9
   module load seqtk/1.3

   set -euo pipefail

   POOL="$(cut -f 1 "${ARRAYDIR}/${PBS_ARRAYID}")"
   TAXON="$(cut -f 2 "${ARRAYDIR}/${PBS_ARRAYID}")"

   SECONDS=0

   mkdir -p "${OUTDIR}/${POOL}/${TAXON}"
   mkdir -p "${SCRATCHDIR}/${POOL}/${TAXON}" && cd "${SCRATCHDIR}/${POOL}/${TAXON}"

   ## samtools view f4 filters out unaligned reads
   ## samtools fasta converts reads into fasta format
   ## seqtk sample randomly samples a certain number of reads
   samtools view -f4 -b  "${INDIR}/${POOL}/alignment.cram" > "unal.bam" \
      || { echo "getting unaligned reads for ${POOL}, ${TAXON} failed"; exit 1; }
```

```
samtools fasta "unal.bam" > "unal.fasta"  \
   || { echo "converting unaligned reads to fasta for ${POOL}, ${TAXON} failed"; exit 1; }
seqtk sample "unal.fasta" "${N_READS_SAMPLED[$TAXON]}" > "reads.fasta" \
   || { echo "sampling unaligned reads for ${POOL}, ${TAXON} failed"; exit 1; }

gfile="${GIDIR}/${TAXON}_gilist.gi"

cp "$gfile" . \
   || { echo "copying GI list for ${POOL}, ${TAXON} failed"; exit 1; }

## BLAST unaligned reads
blastn -query "reads.fasta" -db "${BLASTDB}/nt" \
   -gilist "$(basename "$gfile")" \
   -evalue "$EVALUE" -num_threads "$NTHREADS" \
   -max_target_seqs "$MAX_TARGET_SEQS" -outfmt "$OUTFMT" \
   > "blastn.tsv" \
   || { echo "blast for ${TAXON}, ${POOL} failed"; exit 1; }

## R script to extract top hit for each read
cp "$tophit" .
"./$(basename "$tophit")" "blastn.tsv" "blast_top_hits.tsv" \
   || { echo "extracting top hits for ${TAXON}, ${POOL} failed"; exit 1; }

## R script to make a table of frequencies
cp "$tabulate" .
"./$(basename "$tabulate")" "blast_top_hits.tsv" "blast_names.tsv"  \
   || { echo "Tabulating taxon frequencies for ${TAXON}, ${POOL} failed"; exit 1; }

## Copy output
```

```
    cp "blastn.tsv" "blast_top_hits.tsv" "blast_names.tsv" "${OUTDIR}/${POOL}/${TAXON}" \
       || { echo "copying files for ${TAXON}, ${POOL} failed"; exit 1; }


    echo "RUN TIME $SECONDS ($(($SECONDS/60)) minutes) for ${POOL}, ${TAXON}"

    rm "${ARRAYDIR}/${PBS_ARRAYID}"

else

    mkdir -p "$OUTDIR"
    mkdir -p "$LOGDIR"
    mkdir -p "$SCRIPTDIR"
    mkdir -p "$WORKDIR"
    mkdir -p "$ARRAYDIR"
    mkdir -p "$SCRATCHDIR"

    sfile="${SCRIPTDIR}/$(date '+%Y%m%d-%H%M')-$(basename $0)"
    cp "$0" "$sfile"

    i=0
    for taxon in "${TAXA[@]}"; do
       for d in "$INDIR/"*; do
          pool="$(basename "$d")"
          if [[ -f "${INDIR}/${pool}/alignment.cram" ]]; then
             echo "$pool"$'\t'"$taxon" > "${ARRAYDIR}/${i}"
             i=$(($i+1))
          fi
       done
```

```
    done
    if [[ "$i" == 1 ]]; then
       array="0"
    else
       array="0-$(($i-1))"
    fi

    cd "$WORKDIR"
    qsub2 -A "tiffinp" -W group_list="tiffinp" \
       -q "$QUEUE" -l "walltime=${WALLTIME}" -t "$array" "$sfile"
    echo "$WORKDIR"
    echo "Submitted ${i} jobs"

fi
```

```
#!/usr/bin/env Rscript

###blast_top_hits.tsv###

cargs = commandArgs(trailingOnly=TRUE)
input_file_name = cargs[1]
output_file_name = cargs[2]

blast_data=read.csv(input_file_name,header=FALSE,sep="\t",as.is=TRUE)

blast_names=blast_data[,1]

norepeat_blast_names=unique(blast_names)

x=length(norepeat_blast_names)
m1=vector('list', length=x)
names(m1)= norepeat_blast_names


for(i in norepeat_blast_names[1:x]) {
  j=which(i==blast_data$V1)
  k=blast_data[j,]
  l=k[which.max(k$V10),]
  m1[[i]]=l
}


m2=do.call(rbind, m1)
```

```
m3=m2[,11]

write.table(m3,file=output_file_name,sep="\t",col.names=FALSE, row.names = FALSE, quote=FALSE)
```

```
#!/usr/bin/env Rscript

###blast_names.tsv###

cargs = commandArgs(trailingOnly=TRUE)
input_file_name = cargs[1]
output_file_name = cargs[2]

library(data.table)

blast_names=read.csv(input_file_name,header=FALSE,sep="\t",as.is=TRUE)

blast_names=blast_names[blast_names !="Sinorhizobium meliloti"]

blast_names=blast_names[blast_names !="Medicago truncatula"]

species_freq=as.data.frame(table(unlist(blast_names)))

species_freq=setorderv(species_freq,species_freq$V2,order=-1)

write.table(species_freq,file=output_file_name,sep="\t",col.names=FALSE, row.names = FALSE,
quote=FALSE)
```