ROOT-ASSOCIATED BACTERIAL COMMUNITIES AS AN EXTENDED PHENOTYPE OF THE PLANT

Derek S. Lundberg

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

Chapel Hill
2014

Approved by:

Jeffrey L. Dangl

Corbin D. Jones

John F. Rawls

Matthew C. Wolfgang

Andreas P. Teske

ABSTRACT

DEREK S. LUNDBERG: Root-associated bacterial communities as an extended phenotype of the plant
(Under the direction of Jeff Dangl)

Land plants associate with a root microbiota distinct from the complex microbial community present in surrounding soil. The microbiota colonizing the rhizosphere (immediately surrounding the root) and the endophytic compartment (within the root) contribute to plant growth, productivity, carbon sequestration and phytoremediation. In my research I primarily wanted to test the hypothesis that plants that evolved to live in environments with different challenges also evolved the ability to associate with a unique microbiota as one means of overcoming these challenges. Despite great agronomic interest, genetic principles governing the derivation of host-specific endophyte communities from soil communities are poorly-understood.

I first used extensive sequencing of ribosomal 16S rRNA genes to characterize bacterial populations from hundreds of roots of different genotypes of the model plant *Arabidopsis thaliana* grown in two wild (non-native) soils from North Carolina. These results demonstrated that soil type is the major determinant of the membership of the bacterial community in the rhizosphere and the community living inside roots, and that the developmental stage of the plant as well and the plant genotype actually have relatively minor effects on the colonization behavior of major bacterial taxonomies.

Because in wild microbial communities bacteria with different genomic content may share a similar 16S rRNA gene, and because of limitations in the 16S rRNA sequencing technology, we were limited to statements about bacterial families, and could not say with

confidence to which Arabidopsis genotypes individual strains of bacteria associated. This was a major limitation, because the presence or absence of specific bacterial genes may be a strong determinant of potential host genotypes for a given symbiont. Therefore, I developed technological improvements to increase the accuracy and depth of sequencing, while meanwhile culturing individual strains of bacteria from roots and creating a gnotobiotic system for growing plants in direct association with mixtures of cultured strains. Initial results from this system demonstrate that we can culture a wide diversity of root-associated bacteria and can successfully recolonize plants with complex but defined cocktails of bacteria. Experiments to explore microbe-by genotype association in this gnotobiotic system are underway.

PREFACE

The part of the material in Chapter 1 that follows the heading “Developing systems to study plant microbiota” is part of a review that was commissioned by, and submitted to, the journal *Trends in Cell Biology*. It was ultimately rejected, largely because the field was already saturated with reviews. The review was written after the publication of the work in Chapters 2 and 3, and therefore discusses that work in its current context. For the submitted review, I share authorship with Sarah L. Lebeis, Sur Herrera Paredes, and Jeff Dangl. I contributed to and edited all sections but wrote only the final review section entitled “The search for novel host factors shaping the plant microbiome“, which involves recent conclusions and therefore is truncated in Chapter 1 and continued in Chapter 4.

Chapter 2 was published in *Nature*. I share co-first authorship with Sarah L. Lebeis, Sur Herrera Paredes, and Scott Yourstone. My intellectual contributions were helping design the experiments for the paper, the majority of new protocol development and execution, custom scripting in R for analysis of the data tables resulting from categorization and quantification of the raw sequences (including the organization of count data by bacterial taxonomy and some statistical analysis). I designed and produced graphics for all figures and tables except for figures 2.2, 2.3, 2.13, 2.14, and 2.16. I contributed little to design and implementation of the General Linear Mixed Model (Sur), CARD-FISH (Sarah), and processing of raw sequences

into clean count tables (Scott). I thank Alice Smithlund, Maciej Gonek, Victoria Madden, H. Schmidt, Matthias Rott and N. Zvenigorodsky for technical assistance, Amyé Spor, Jason Peiffer and John F Rawls for discussions, and Cathy Herring for access to Clayton field soil.

Chapter 3 was published in *Nature Methods*. I share co-first authorship with Scott Yourstone. My intellectual contributions included assistance with template tagging and PCR primer design, all wet-bench protocol design and execution of protocols, Peptide Nucleic Acid design and statistical validation, all bioinformatics downstream of MTToolbox processing, creating all figures with the exception of some parts of Figure 3.6, and writing the manuscript. Scott Yourstone designed and wrote MTToolbox, the efficient and user-friendly informatics pipeline. However, for some analysis such as that in Figure 3.8, I reproduced key aspects of this pipeline using my own inefficient code to make “consensus sequences” and “perfect consensus sequences”, so I am familiar with most computational aspects of this work as well.

For experiments mentioned in Chapter 4, I especially thank Surojit Biswas, Sur Herrera Paredes, Meredith McDonald, Natalie Breakfield, Meghan Feltcher, and Jeff Dangl. For the bacterial culture collection I thank Maciej Gonek, Alice Smithlund, Max Rose, and Meredith McDonald.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER 1

INTRODUCTION

INTRODUCTION

The plant microbiome (its collection of associated microbes) is an example of Richard Dawkins' 'extended phenotype' (Dawkins 1989). Plant genotype, both within and between species, has been correlated with differences in the associated microbiome, with consequent associations to plant growth, development, and performance. The productivity of any plant community relies in part on the respective plant-associated microbiomes, which are distinct from the microbiomes in surrounding soil. The microbiota is most simply viewed as an extension of each plant's genome; we do not know any plant genome's full functional capacity until we also know the functional capacity and assemblage drivers on its associated microbiome.

Essentially all land plants grow in intimate association with a complex microbiota. Microbes in both the phyllosphere and the rhizosphere can be either endophytic, epiphytic, or closely associated. Examples of close microbial associates include those inhabiting the fluid in pitcher plants, or those not touching roots but heavily influenced by root exudates in the nearby soil. The host plant often relies on its root-associated metagenome to provide it with critical nutrients (Bais et al. 2006), as minerals are often present in the soil in forms inaccessible to plants. In other cases plant-associated microbes, such as Pseudomonads (Mavrodi et al. 2006; Vacheron et al. 2013), can act as protectants against phytopathogens. Other microbes have been shown to improve growth through production of phytohormones

and to help plants withstand heat, salt, drought, and more (Bais et al. 2006; Rodriguez and Redman 2008). Growth-promoting mutualistic bacteria associate with the phyllosphere as well (Vorholt 2012; Whipps et al. 2008). The plant, in turn, cultivates its microbiome through means such as adjusting the pH, reducing competition for beneficial microbes, and providing an energy source, mostly in the form of carbon-rich exudates. Decaying dead roots are an important contributor of carbon to the soil, and between 5-33% of all atmospheric carbon fixed (depending on plant species and climate conditions) exits on living roots as exudates (DeDeyn et al. 2008). Thus, the global plant-associated rhizosphere microbiome is a very large carbon sequestration niche.

Studies have begun in a variety of systems to address the microbiome of various crops such as maize (Chelius and Triplett 2001), rice (Erkel et al. 2006; Oliveira et al. 2009), and other plants (Whipps et al. 2008). Understanding the causes of these differences could be transformational in plant breeding and biotechnology, because there is the potential to uncover a whole new suite of genes capable of improving plant yields in adverse conditions through exploitation and manipulation of the innate (and/or adjusted) probiotic capacity of soils and the natural environment. Additionally, rhizosphere microorganisms play a key role in long-term sequestration of the carbon secreted by plant roots (DeDeyn et al. 2008), meaning that understanding what plant gene products attract particular microbial communities could help uncover a genetic component to climate change (Ryan et al. 2009).

Microbial community influences plant health. The best-known strategy of plants to improve the uptake of phosphate, nitrogen and other minerals is to form symbioses with arbuscular mycorrhizal (AM) fungi of the phylum Glomeromycota (Schuessler et al. 2001) AM fungi associate intimately with host roots, growing inter- and intracellularly within the root cortex. Intraradical colonization enables fungal access to carbohydrates required for the

formation of extensive extraradical mycelia, which can lead to a 100-fold increase of the nutrient absorbing surface of the root, thus allowing the plant to efficiently utilize minerals and to exploit nutrient resources not available without symbiosis. AM symbiosis dates back >400 million years (Remy et al. 1994) and coincided with plant colonization of land (Heckman et al. 2001; Redecker et al. 2000), and most plant lineages associate with AM fungi. Arabidopsis, like the majority of Brassicaceae, and like several other plant lineages, has lost the association with AM fungi (Smith and Read 2010). Nevertheless, Brassicas like Arabidopsis do manage to extract phosphorous from the soil using root hairs (Bates and Lynch 2000) and perhaps associated microbes. Arabidopsis provides a valuable opportunity to focus on a plant's interaction with other important rhizosphere microorganisms in the absence of the influence of AM fungi. Additionally, Arabidopsis is the reference system for plant genetics, genomics, and molecular biology (Initiative 2000; Weigel 2012) .

The identification of plant genes associated with specific microbial community traits will lead to the search for homologs or functional equivalents in mammalian species. Multicellular eukaryotes evolved in a microbial world, hence the evolutionary conservation of host-microbe interactions can be very ancient. Indeed, host responses to microbial colonization are evolutionarily conserved between mammals and fish (Rawls et al. 2006). It is conceivable that host responses to microbial colonization, and even host modulation of surface microbial communities, are driven by processes or genes that are shared among members of different plant phlya. Once identified, host genes important in shaping microbial communities will constitute targets for intervention. Restructuring microbial communities is likely to be desirable in agriculture for promoting plant health in particular soil and climate conditions.

Developing systems to study plant microbiota

In the simplest scenarios, microbes from surrounding environments (i.e. soil, water, and air) colonize plants via a random process involving first come-first serve niche filling together with inter-microbe competition to establish idiosyncratic communities on or inside organs of each host. This scenario, though, is an unlikely mechanism for microbiota colonization of the roots, especially given recent studies of poplar (Gottel et al. 2011), *Arabidopsis thaliana* grown in multiple natural soils from two continents (Bulgarelli et al. 2012; Lundberg et al. 2012) and maize from five diverse North American fields (Peiffer et al. 2013). In these studies, whether plants are grown outdoors or in greenhouses or growth chambers, specific bacterial taxa are reproducibly enriched inside the root as compared with surrounding soils. Several generalities emerged from these studies: soils, as expected were the most diverse environment, while the intimate zone of host-microbe contact inside the root (putative endophytes) were significantly less diverse; the largest contributor to root microbiome composition was the physical proximity to the plant; the second largest contributor was the wild soil from which the community was recruited.

Recently, similar studies were carried out on above ground organs, the phyllosphere, where reproducible colonization of a subset of bacterial taxa from the surrounding aerial environment was shown in *A. thaliana* leaves (Maignien et al. 2014). However, the importance of stochastic processes and early colonization events was shown to be a major force shaping mature microbial communities in these leaves (Maignien et al. 2014) and also in apple flowers sampled over developmental time (Shade et al. 2013). Despite large genetic differences across hosts, organs sampled, surrounding environments and experimental methods, consistent taxonomies are recurrently found in the rhizosphere (Bodenhausen et al. 2013; Bulgarelli et al. 2012; Gottel et al. 2011; Lundberg et al. 2012;

Peiffer et al. 2013; Schlaeppi et al. 2013) or phyllosphere (Delmotte et al. 2009; Maignien et al. 2014; Rastogi et al. 2012) organs. Collectively, these results suggest that a core microbiota is recruited from very diverse surroundings; environment-specific enrichments of certain taxa suggest fine-tuning both on leaves and roots by the host, in combination with environmental biotic and abiotic factors (Bulgarelli et al. 2013; Vorholt 2012).

Ecological processes shaping host-associated microbial communities

Seed dispersal is an important ecological process in plants and some plant-associated microbes are known to be inherited via seeds. For instance, there is long-standing evidence for seed-based inheritance of rhizobia in legume seeds (Ash 1949), and recent studies suggest the existence of surprising microbial diversity in the seeds of maize (Johnston-Monje and Raizada 2011) and spinach (Lopez-Velasco et al. 2013). Bacterial seed coating can protect against pathogens (Hameeda et al. 2010; Wright 2005) and promote plant growth (Jetiyanon et al. 2008). Seeds can also harbor bacterial (Gitaitis and Walcott 2007), fungal (Biswas et al. 2013; Maruthachalam et al. 2013) and oomycete (Testen et al. 2013) pathogens. Thus, understanding dispersion dynamics will ultimately lead to better disease control strategies. Readily dispersed microbes might have a competitive advantage over microbes that colonize after germination. These cases might lead to “historically contingent” microbial communities where the early colonizers determine the final community, mediated by microbe-microbe interactions, or by plant mechanisms reinforcing the primacy of early colonizers. An alternative, but not exclusive, model proposes that successions of microbial communities emerge over developmental time during the host plant life cycle. Consistent with this hypothesis, different bacterial taxa preferentially colonize the apple flower at different developmental stages (Shade et al. 2013). However, in the fast

growing annual *A. thaliana*, little difference in root bacterial community was noted at two very different developmental stages, before and well after the metabolic switch in carbon allocation (Lundberg et al. 2012), and poplar trees of different ages showed similar communities (Gottel et al. 2011). To fully elucidate how the order of microbial colonization affects the plant microbiome, it would be necessary to carry out studies with time series and crossover designs; this type of design has already been used to establish the existence of such "order effects" in the context of colonization of the mammalian gut (Lee et al. 2013). Little is known about how greater differences in developmental time, or how annual versus perennial life histories influence the assembly and long term stability of plant microbiota.

While increased dispersal from sources like soil, seed, or decaying leaf litter is expected to increase diversity within individual plants, drift would counteract this effect and perhaps add further heterogeneity because species represented by very low number of individuals will have a high probability of undergoing local extinction due to stochastic fluctuation. In the context of the endophytic compartment of the *A. thaliana* root and leaf, drift might be particularly important given the relatively low estimates of a total of 10^5 endophytic bacterial cells per root system (Lundberg et al. 2012), and 10^4 cells/cm^2 on the leaf of the same species (Maignien et al. 2014) (though it should be noted that neither of these estimates were from wild-grown plants). Given that hundreds of ribotypes were detected on both organs, these results imply only tens to hundreds of individuals per ribotype. From an ecological perspective, the health of a community can be viewed as its ability to withstand and recover from perturbations, and low bacterial diversity in the mammalian gut has been associated with susceptibility to perturbation (Virgin and Todd 2011) and disease (Turnbaugh et al. 2009). It is possible that diversity plays a similar role in maintaining a healthy plant microbiome, but systematically controlling and varying diversity

in microcosm reconstitution experiments is required to fully distinguish between cause and effect.

Another process influencing microbiome composition is selection, by both abiotic (e.g. drought, salinity, nutrient availability, etc.) and biotic (e.g. influence by the host or other microbes) environmental components, which can occur due to any of these factors acting directly on microbial species in the community, or be mediated by the plant host as it reads and responds to its abiotic and/or biotic environment (Eisenhauer et al. 2013). It should be noted that selection might act either at some microbial taxonomic level, at the gene level, or both. For example legumes recruit nitrogen-fixing rhizobia to their roots only under nitrogen deprivation conditions (Coronado et al. 1995), and remarkable host-genotypic specificity has been found (Laguerre et al. 2003). Further, drought stress on the plant leads to an enrichment of bacteria that produce 1-Aminocyclopropane-1-carboxylate (ACC) deaminase (Marasco et al. 2012), which is known to reduce ethylene concentrations under stress conditions; lowering ethylene levels helps plants recover from different abiotic stresses (Cao et al. 2007). Despite this evidence, 16S rRNA gene-based experiments on samples from plants grown in natural soils have provided little insight on the effect of abiotic factors on community assembly, mainly because it has been impossible to disentangle the effect of abiotic and biotic factors in these studies. Both microbe-microbe and plant-microbe interactions might affect the ability of a specific microbe to colonize the host under different environments. An association between herbivore behavior and root microbiome has been suggested (Badri et al. 2013), and differences were found in bacterial networks constructed from rhizosphere microbiomes of different plant hosts and across variable plant host diversity (Bakker et al. 2013); however, it remains an open challenge to tie these observations to microbiome function.

Contributions of immune system surveillance receptors in sculpting microbiomes

The repertoire of innate immune receptors that recognize and respond to microbe-associated molecular patterns (MAMPs) allows for accurate sensing of changing microbial environments. Although pattern recognition receptors (PRRs) have evolved separately in plants and animals (Ausubel 2005; Ronald and Beutler 2010), the development of analogous microbial perception systems at least twice points to their importance in the lifecycle of both plants and animals. In addition, a large and divergent family of 'disease resistance proteins' gives plants the ability to specifically recognize pathogen virulence factors that overcome resistance and allow microbial colonization. The majority of plant disease resistance proteins are NB-LRR proteins (nucleotide-binding site, leucine rich repeats), which act as the primary intracellular receptors of the plant immune system (Chisholm et al. 2006; Jones and Dangl 2006). Because host-associated microbial communities are formed from the microbes present in the host's environment, it is natural to hypothesize that these resistance gene systems are particularly important in the broader process of community assembly.

The search for novel host factors shaping the plant microbiome

Much is known about plant-genotype specificity governing disease resistance or susceptibility to pathogens that threaten economically important crops; forward genetics to exploit host genotypic diversity is well advanced for dissection of these binary interactions (Dangl et al. 2013; Gururani et al. 2012). However, the influence of plant genotype on the establishment of mutualistic interactions with populations of microorganisms is less understood, probably because beneficial phenotypes are not as carefully monitored as

pathogenic phenotypes in agricultural settings. Notable exceptions, of course, are important mutualists like arbuscular mycorrhizal fungi, which have broad host ranges and provide nearly all agronomically relevant plants with phosphate in exchange for fixed carbon (Bonfante and Genre 2010) and the deeply studied, typically host-specific nitrogen-fixing rhizobia (Kondorosi et al. 2013). The extent of host specificity, if any, for individual members or functional guilds derived from the larger plant microbiota is now an area of active research. The hope is to uncover host loci affecting both general and specialized microbial colonization by either widespread or soil-specific mutualists, respectively, as well as the microbial genetic loci enabling them to communicate effectively with the host. The potential agronomic and economical value of such genetic knowledge is huge, as it might enable breeders to tailor plants to take advantage of particular types of microbial environments, or to take advantage of artificially-supplied beneficial microbial treatments as probiotics.

To test and identify the contribution of plant genotype to the composition of a specific microbiome, several groups grew different inbred plant genotypes in wild soils, under either field conditions or in controlled environmental conditions, and then defined the microbial community assembled on each host genotype by deep ribotyping (Bulgarelli et al. 2012; Bulgarelli et al. 2013; Lundberg et al. 2012; Peiffer et al. 2013; Schlaeppi et al. 2013; Turner et al. 2013). These studies revealed low variation in microbiota attributable to plant-genotype, which may nonetheless be amenable to genetic dissection (Bulgarelli et al. 2012; Lundberg et al. 2012; Peiffer et al. 2013). In cases where constructing experiments is not possible, attempts have been made to separate covariates such as geographical location, as for grape vineyards (Bokulich et al. 2013) and willow trees (Bell et al. 2013). As the phylogenetic distance between the host genotypes compared decreases, especially when comparing genotypes within a single plant species, it becomes theoretically possible to treat elements of the microbial community (such as presence/absence or the abundance of one

or more members) as a phenotype, and to map these phenotypes to plant host loci, as has been attempted in other systems (Bell et al. 2013; Benson et al. 2010; Srinivas et al. 2013).

It may ultimately be the case that, as with mycorrhizal fungi, many of the beneficial members of the plant microbiome are generalists with the ability to colonize a wide range of plant hosts (Moora et al. 2011). This is supported by the observation that the major enriched bacterial taxa are the same across different species closely related to *Arabidopsis thaliana* (~25M years of divergence) (Schlaeppi et al. 2013), and by the relatively subtle variation observed so far between the microbiota of different plant genotypes within a species. It could be that, for a plant, depending on microbial genes found only in a smaller set of specific microbes might represent too great a survival risk, since these microbes might not always be present. On the timescale of plant evolution there is a great deal of microbial turnover, so the majority of plant recruitment mechanisms may need to work over relatively wide groups of bacterial taxa, in order to associate with relatively common beneficial microbial functions. As noted several times above, however, this is definitely not always the case, as tight host specificity for some microbial associations is well documented. Whether the apparent rarity of plant host specificity in microbiota assembly is truly as uncommon as it appears to be based on 16S rRNA studies, or whether we are still merely blind to important differences in the microbiota of genetically distinct plant hosts simply because 16S rRNA studies lack the necessarily resolution, remains to be seen.

REFERENCES

Ash CGaA, O.N. (1949) A Comparison of Methods Recommended for the Surface Sterilization of Leguminous Seed. Soil Science Society of America Journal 13(C): 279-283.

Ausubel FM (2005) Are innate immune signaling pathways in plants and animals conserved? Nature immunology 6(10): 973-979.

Badri DV, Chaparro JM, Zhang R, Shen Q, Vivanco JM (2013) Application of natural blends of phytochemicals derived from the root exudates of Arabidopsis to the soil reveal that phenolic-related compounds predominantly modulate the soil microbiome. The Journal of biological chemistry 288(7): 4502-4512.

Bais HP, Weir TL, Perry LG, Gilroy S, Vivanco JM (2006) The Role of Root Exudates in Rhizosphere Interactions with Plants and Other Organisms. Annual Review of Plant Biology 57(1): 233-266.

Bakker PA, Berendsen RL, Doornbos RF, Wintermans PC, Pieterse CM (2013) The rhizosphere revisited: root microbiomics. Frontiers in plant science 4: 165.

Bates TR, Lynch JP (2000) The efficiency of *Arabidopsis thaliana* (Brassicaceae) root hairs in phosphorus acquisition. American Journal of Botany 87(7): 964-970.

Bell TH, El-Din Hassan S, Lauron-Moreau A, Al-Otaibi F, Hijri M et al. (2013) Linkage between bacterial and fungal rhizosphere communities in hydrocarbon-contaminated soils is related to plant phylogeny. ISME J.

Benson AK, Kelly SA, Legge R, Ma F, Low SJ et al. (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of Sciences 107(44): 18933-18938.

Biswas C, Dey P, Satpathy S, Sarkar SK, Bera A et al. (2013) A simple method of DNA isolation from jute (Corchorus olitorius) seed suitable for PCR-based detection of the pathogen Macrophomina phaseolina (Tassi) Goid. Letters in applied microbiology 56(2): 105-110.

Bodenhausen N, Horton MW, Bergelson J (2013) Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. PLoS One 8(2): e56329.

Bokulich NA, Thorngate JH, Richardson PM, Mills DA (2013) Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. Proc Natl Acad Sci U S A.

Bonfante P, Genre A (2010) Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis. Nature communications 1: 48.

Bulgarelli D, Rott M, Schlaeppi K, Ver Loren van Themaat E, Ahmadinejad N et al. (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature 488(7409): 91-95.

Bulgarelli D, Schlaeppi K, Spaepen S, Ver Loren van Themaat E, Schulze-Lefert P (2013) Structure and Functions of the Bacterial Microbiota of Plants. Annual Review of Plant Biology 64(1): 807-838.

Cao WH, Liu J, He XJ, Mu RL, Zhou HL et al. (2007) Modulation of ethylene responses affects plant salt-stress responses. Plant Physiol 143(2): 707-719.

Chelius MK, Triplett EW (2001) The Diversity of Archaea and Bacteria in Association with the Roots of Zea mays L. Microb Ecol 41(3): 252-263.

Chisholm ST, Coaker G, Day B, Staskawicz BJ (2006) Host-Microbe Interactions: Shaping the Evolution of the Plant Immune Response. Cell 124(4): 803-814.

Coronado C, Zuanazzi J, Sallaud C, Quirion JC, Esnault R et al. (1995) Alfalfa Root Flavonoid Production Is Nitrogen Regulated. Plant Physiol 108(2): 533-542.

Dangl JL, Horvath DM, Staskawicz BJ (2013) Pivoting the plant immune system from dissection to deployment. Science 341(6147): 746-751.

Dawkins R (1989) The extended phenotype : the long reach of the gene. Oxford ;New York: Oxford University Press, 1989.

DeDeyn GB, Cornelissen JHC, Bardgett RD (2008) Plant functional traits and soil carbon sequestration in contrasting biomes. Ecology Letters 11(5): 516-531.

Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B et al. (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. Proc Natl Acad Sci U S A 106(38): 16428-16433.

Eisenhauer N, Dobies T, Cesarz S, Hobbie SE, Meyer RJ et al. (2013) Plant diversity effects on soil food webs are stronger than those of elevated CO2 and N deposition in a long-term grassland experiment. Proc Natl Acad Sci U S A 110(17): 6889-6894.

Erkel C, Kube M, Reinhardt R, Liesack W (2006) Genome of Rice Cluster I archaea--the key methane producers in the rice rhizosphere. Science 313(5785): 370-372.

Gitaitis R, Walcott R (2007) The epidemiology and management of seedborne bacterial diseases. Annual review of phytopathology 45: 371-397.

Gottel NR, Castro HF, Kerley M, Yang Z, Pelletier DA et al. (2011) Distinct microbial communities within the endosphere and rhizosphere of Populus deltoides roots across contrasting soil types. Appl Environ Microbiol 77(17): 5934-5944.

Gururani MA, Upadhyaya CP, Strasser RJ, Woong YJ, Park SW (2012) Physiological and biochemical responses of transgenic potato plants with altered expression of PSII manganese stabilizing protein. Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale 58: 182-194.

Hameeda B, Harini G, Rupela OP, Kumar Rao JV, Reddy G (2010) Biological Control of Chickpea Collar Rot by Co-inoculation of Antagonistic Bacteria and Compatible Rhizobia. Indian journal of microbiology 50(4): 419-424.

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL et al. (2001) Molecular Evidence for the Early Colonization of Land by Fungi and Plants. Science 293(5532): 1129-1133.

Initiative AG (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408(6814): 796-815.

Jetiyanon K, Wittaya-Areekul S, Plianbangchang P (2008) Film coating of seeds with Bacillus cereus RS87 spores for early plant growth enhancement. Canadian journal of microbiology 54(10): 861-867.

Johnston-Monje D, Raizada MN (2011) Conservation and diversity of seed associated endophytes in Zea across boundaries of evolution, ethnography and ecology. PLoS One 6(6): e20396.

Jones JDG, Dangl JL (2006) The plant immune system. Nature 444(7117): 323-329.

Kondorosi E, Mergaert P, Kereszt A (2013) A paradigm for endosymbiotic life: cell differentiation of Rhizobium bacteria provoked by host plant factors. Annual review of microbiology 67: 611-628.

Laguerre G, Louvrier P, Allard MR, Amarger N (2003) Compatibility of rhizobial genotypes within natural populations of Rhizobium leguminosarum biovar viciae for nodulation of host legumes. Appl Environ Microbiol 69(4): 2276-2283.

Lee SM, Donaldson GP, Mikulski Z, Boyajian S, Ley K et al. (2013) Bacterial colonization factors control specificity and stability of the gut microbiota. Nature 501(7467): 426-429.

Lopez-Velasco G, Carder PA, Welbaum GE, Ponder MA (2013) Diversity of the spinach (Spinacia oleracea) spermosphere and phyllosphere bacterial communities. FEMS microbiology letters 346(2): 146-154.

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J et al. (2012) Defining the core *Arabidopsis thaliana* root microbiome. Nature 488(7409): 86-90.

Maignien L, Deforce EA, Chafee ME, Eren AM, Simmons SL (2014) Ecological Succession and Stochastic Variation in the Assembly of *Arabidopsis thaliana* Phyllosphere Communities. mBio 5(1).

Marasco R, Rolli E, Ettoumi B, Vigani G, Mapelli F et al. (2012) A drought resistance-promoting microbiome is selected by root system under desert farming. PLoS One 7(10): e48479.

Maruthachalam K, Klosterman SJ, Anchieta A, Mou B, Subbarao KV (2013) Colonization of spinach by Verticillium dahliae and effects of pathogen localization on the efficacy of seed treatments. Phytopathology 103(3): 268-280.

Mavrodi DV, Blankenfeldt W, Thomashow LS (2006) Phenazine Compounds in Fluorescent Pseudomonas Spp. Biosynthesis and Regulation. Annual review of phytopathology 44(1): 417-445.

Moora M, Berger S, Davison J, Öpik M, Bommarco R et al. (2011) Alien plants associate with widespread generalist arbuscular mycorrhizal fungal taxa: evidence from a continental-scale study using massively parallel 454 sequencing. Journal of Biogeography 38(7): 1305-1317.

Oliveira CA, Sá NMH, Gomes EA, Marriel IE, Scotti MR et al. (2009) Assessment of the mycorrhizal community in the rhizosphere of maize (Zea mays L.) genotypes contrasting for phosphorus efficiency in the acid savannas of Brazil using denaturing gradient gel electrophoresis (DGGE). Applied Soil Ecology 41(3): 249-258.

Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG et al. (2013) Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc Natl Acad Sci U S A 110(16): 6548-6553.

Rastogi G, Sbodio A, Tech JJ, Suslow TV, Coaker GL et al. (2012) Leaf microbiota in an agroecosystem: spatiotemporal variation in bacterial community composition on field-grown lettuce. ISME J 6(10): 1812-1822.

Rawls JF, Mahowald MA, Ley RE, Gordon JI (2006) Reciprocal Gut Microbiota Transplants from Zebrafish and Mice to Germ-free Recipients Reveal Host Habitat Selection. Cell 127(2): 423-433.

Redecker D, Kodner R, Graham LE (2000) Glomalean Fungi from the Ordovician. Science 289(5486): 1920-1921.

Remy W, Taylor TN, Hass H, Kerp H (1994) Four hundred-million-year-old vesicular arbuscular mycorrhizae. Proceedings of the National Academy of Sciences 91(25): 11841-11843.

Rodriguez R, Redman R (2008) More than 400 million years of evolution and some plants still can't make it on their own: plant stress tolerance via fungal symbiosis. J Exp Bot: 1109-1114.

Ronald PC, Beutler B (2010) Plant and animal sensors of conserved microbial signatures. Science 330(6007): 1061-1064.

Ryan P, Dessaux Y, Thomashow L, Weller D (2009) Rhizosphere engineering and management for sustainable agriculture. Plant and Soil 321(1-2): 363-383.

Schlaeppi K, Dombrowski N, Oter RG, Ver Loren van Themaat E, Schulze-Lefert P (2013) Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. Proc Natl Acad Sci U S A.

Schuessler A, Schwarzott D, Walker C (2001) A new fungal phylum, the Glomeromycota: phylogeny and evolution. Mycological Research 105(12): 1413-1421.

Shade A, McManus PS, Handelsman J (2013) Unexpected diversity during community succession in the apple flower microbiome. mBio 4(2).

Smith SE, Read DJ (2010) Mycorrhizal symbiosis: Academic press.

Srinivas G, Moller S, Wang J, Kunzel S, Zillikens D et al. (2013) Genome-wide mapping of gene-microbiota interactions in susceptibility to autoimmune skin blistering. Nature communications 4: 2462.

Testen AL, Jimenez-Gasco MD, Ochoa JB, Backman PA (2013) Molecular detection of Peronospora variabilis in quinoa seeds and phylogeny of the quinoa downy mildew pathogen in South America and the United States. Phytopathology.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A et al. (2009) A core gut microbiome in obese and lean twins. Nature 457(7228): 480-484.

Turner TR, Ramakrishnan K, Walshaw J, Heavens D, Alston M et al. (2013) Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. ISME J 7(12): 2248-2258.

Vacheron J, Desbrosses G, Bouffaud M-L, Touraine B, Moënne-Loccoz Y et al. (2013) Plant growth-promoting rhizobacteria and root system functioning. Frontiers in plant science 4.

Virgin HW, Todd JA (2011) Metagenomics and personalized medicine. Cell 147(1): 44-56.

Vorholt JA (2012) Microbial life in the phyllosphere. Nature reviews Microbiology 10(12): 828-840.

Weigel D (2012) Natural Variation in Arabidopsis: From Molecular Genetics to Ecological Genomics. Plant Physiology 158(1): 2-22.

Whipps JM, Hand P, Pink D, Bending GD (2008) Phyllosphere microbiology with special reference to diversity and plant genotype. Journal of Applied Microbiology 105(6): 1744-1755.

Wright DA, Swaminathan, J., Blaser, M., and Jackson, T.A. (2005) Carrot Seed Coating with Bacteria for Seedling Protection from Grass Grub Damage. New Zealand Plant Protection 58: 229-233.

CHAPTER 2

Defining the core *Arabidopsis thaliana* root microbiome[1]

INTRODUCTION

Land plants associate with a root microbiota distinct from the complex microbial community present in surrounding soil. The microbiota colonizing the rhizosphere (immediately surrounding the root) and the endophytic compartment (within the root) contribute to plant growth, productivity, carbon sequestration and phytoremediation (DeDeyn et al. 2008; Rodriguez and Redman 2008; Van der Lelie et al. 2009). Colonization of the root occurs despite a sophisticated plant immune system (Dodds and Rathjen 2011; Jones and Dangl 2006), suggesting finely tuned discrimination of mutualists and commensals from pathogens. Genetic principles governing the derivation of host-specific endophyte communities from soil communities are poorly understood. Here we report the pyrosequencing of the bacterial 16S ribosomal RNA gene of more than 600 *Arabidopsis thaliana* plants to test the hypotheses that the root rhizosphere and endophytic compartment microbiota of plants grown under controlled conditions in natural soils are sufficiently dependent on the host to remain consistent across different soil types and developmental stages, and sufficiently dependent on host genotype to vary between inbred

---

[1]Lundberg DS*, Lebeis SL*, Paredes SH*, Yourstone S*, Gehring J et al. (2012) Defining the core *Arabidopsis thaliana* root microbiome. Nature 488(7409): 86-90.
* = contributed equally

Arabidopsis accessions. We describe different bacterial communities in two geochemically distinct bulk soils and in rhizosphere and endophytic compartments prepared from roots grown in these soils. The communities in each compartment are strongly influenced by soil type. Endophytic compartments from both soils feature overlapping, low-complexity communities that are markedly enriched in Actinobacteria and specific families from other phyla, notably Proteobacteria. Some bacteria vary quantitatively between plants of different developmental stage and genotype. Our rigorous definition of an endophytic compartment microbiome should facilitate controlled dissection of plant–microbe interactions derived from complex soil communities.

MAIN

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

mycorrhizal fungi, implying that other microorganisms may fill this niche.

Microbial community structure differs across plant species (Hardoim et al. 2008; Redford et al. 2010), and there are reports of host-genotype-dependent differences in patterns of microbial associations (Inceoglu et al. 2011; Inceoglu et al. 2010). However, the divergent methods used in those studies relied on small sample sizes and low-resolution phylotyping techniques potentially confounded by off-target sequences and chimaeric amplicons. We developed a robust experimental system to sample repeatedly the root microbiome using high-throughput sequencing. Our results confirm many of the general conclusions from earlier studies and, because of controlled experimental design and the power of deep sequencing, provide a key step towards the definition of this microbiome's functional capacity and the host genes that potentially contribute to microbial association phenotypes. Such plant genes would constitute major agronomic targets.

We used 454 pyrosequencing to sequence 16S ribosomal RNA (rRNA) gene amplicons for DNA prepared from eight diverse, inbred *A. thaliana* accessions. Plants were grown from surface-sterile seeds in climate-controlled conditions in two diverse soils, respectively termed Mason Farm and Clayton (Table 2.1; detailed in Methods). For each soil, we assayed multiple individuals from each *A. thaliana* accession grown from sterile seeds in both soils across independent full-factorial biological replicates, in which all genotypes and bulk soils (pots without a plant) for a given soil type were grown in parallel (Table 2.2). We isolated separate rhizosphere and EC fractions from individual plant root systems (Fig. 2.1 and Table 2.2). We established 1114F and 1392R as our primer pair (Methods and Fig. 2.2). Using an otupipe-based pipeline (http://drive5.com/otupipe/), we grouped sequences into 97%-identical operational taxonomic units (OTUs), reduced noise and removed chimaeras. We determined technical reproducibility thresholds to conclude that OTUs defined by >25 reads in >5 samples (hereafter 25 × 5) are individually 'measurable OTUs' (Benson et al.

2010; Gottel et al. 2011) (Figs 2.2 and 2.4). All data reported here are from one run of our otupipe-based pipeline (Fig. 2.3).

Excluding additional control samples, we ribotyped 1,248 samples comprising 111 bulk soil, 613 rhizosphere and 524 EC samples, generating 9,787,070 high-quality reads (Figs 2.3 and 2.4a–c). After removing plant-sequence-derived OTUs, we obtained a table of usable OTU read counts per sample containing 6,387,407 reads distributed across 18,783 OTUs. We normalized this table of usable reads by rarefying to 1,000 reads per sample or, alternatively, by dividing the reads per OTU in a sample by the sum of usable reads in that sample, resulting in a table of relative abundances (frequencies). Using the 25 × 5 threshold, we defined 778 measurable OTUs representing 54% (3,463,632) of the usable reads (Fig. 2.4c). The diversity of the 778 measurable OTUs in soil, rhizosphere and EC fractions showed expected relative trends when compared with the diversity by fraction of all usable OTUs (Fig. 2.4d).

We used principal coordinate analysis on pairwise, normalized, weighted UniFrac distances between all samples, considering all usable OTUs, to identify the main factors driving community composition (Fig. 2.5a and 2.6a). The first principal coordinate (PCo1) revealed that the two bulk soils and their associated rhizospheres were differentiated from the respective EC fractions. Soil type was the main factor in the second component (PCo2). This pattern was recapitulated by hierarchical clustering of pairwise Bray–Curtis dissimilarities considering only measurable OTUs (Fig. 2.5b and 2.6b). Samples harvested at different developmental stages clustered together, indicating that this variable does not have a major effect on overall community composition (Fig. 2.5 and 2.6a, b; yng versus old, where yng refers to the time of appearance of an inflorescence meristem and old refers to fruiting plants with greater than 50% senescent leaves). Additional control samples from the reference genotype Col-0 harvested from four independent digs of Mason Farm soil

underscored the reproducibility of these bacterial community profiles (Fig. 2.7). Together, these data demonstrate that the interaction of diverse soil communities with plants determines the assembly of the rhizosphere, leading to winnowed ECs, that the ECs from at least these two diverse soils are very different from the starting soil communities and that there is little difference in communities over host developmental time. We fitted a general linear mixed model (GLMM) to samples from each set of plant fractions (rhizosphere or EC), plus the bulk soil controls, to identify measurable OTUs whose abundances differ significantly between plant and bulk soil as a result of soil type, developmental stage, fraction and genotype (Methods). This approach allowed us to quantify the contribution from each variable to the community composition (Table 2.3). Controlling for sequencing plate effects, plant fraction is the most important factor; its effect is strongest for the EC, consistent with our UniFrac and Bray–Curtis analyses. Soil type is less important, followed by experiment, developmental stage and, finally, genotype, which had a small but consistent effect.

Hierarchical clustering of sample groups considering 256 OTUs identified by the GLMM to differentiate rhizosphere and EC from soil recapitulated the separation of EC from soil and rhizosphere (Fig. 2.8A and Fig. 2.9A, left; compare with Fig. 2.5 and 2.6). Of these, 164 OTUs were enriched in EC samples (Fig. 2.8B, a; dark and light red bars), defining an *A. thaliana* 'EC microbiome'. Of these 164, 97 were enriched in EC samples from both soil types (Fig. 2.8B, a; dark red bars), potentially representing a core EC microbiome. By contrast, 67 of these 164 were enriched in EC to a greater extent in one soil than the other (Fig. 2.8B, a; light red bars; Fig. 2.8B, b)). Importantly, 32 OTUs were depleted in EC samples (Fig. 2.8B, a; blue bars). Some OTUs exhibited rhizosphere enrichment; these significantly overlapped the EC-enriched OTUs ($P < 10^{-16}$, one-sided hypergeometric test) and also sometimes had a soil-type component (Fig. 2.8B, c and d). Only a few rhizosphere-

specific enrichments were not also enriched in the EC. Hence, the *A. thaliana* EC microbiome is enriched for both a shared set of OTUs commonly assembled across two replicates from two diverse soils, and a set of OTUs that are assembled from each soil. We assessed taxonomic distributions, first those of the 778 measurable OTUs in soil, rhizosphere and EC fractions, and then those of the 256 EC-enriched and 32 EC-depleted OTUs (Fig. 2.8A, 2.9A). Measurable OTUs were distributed across seven dominant phyla (Fig. 2.8c and Fig. 2.9c) and contained ~50–70% of the usable reads in all fractions (Fig. 2.4c). Phyla distribution of the EC-enriched OTUs reflected that of the entire EC. Conversely, the phyla distribution of the EC-depleted OTUs typically resembled that of the rhizosphere fraction (Fig. 2.8C). The lower Shannon diversity of the EC fraction is consistent with enrichment for a subset of dominant phyla. Specifically, the EC microbiome was dominated by Actinobacteria, Proteobacteria and Firmicutes, and was depleted of Acidobacteria, Gemmatimonadetes and Verrucomicrobia, when soil types were considered either together or separately (Fig. 2.8C, Fig. 2.9C and Fig 2.19). Lower-order taxonomic analysis (Fig. 2.8D and Fig. 2.9D) demonstrated that enrichment of a low-diversity Actinobacteria community in the EC was driven by a subset of families, predominantly Streptomycetaceae.

Other phyla, such as Proteobacteria, were represented by both EC enrichments and EC depletions at the family level (Fig. 2.8E and Fig. 2.9E). Strikingly, two alphaproteobacterial families, Rhizobiaceae and Methylobacteriaceae, and two gammaproteobacterial families, Pseudomonadaceae and Moraxellaceae, dominated the EC population in their respective classes (Fig. 2.8F, α and γ, and Fig. 2.9F, α and γ). Equally striking was the EC redistribution of particular alpha- and gammaproteobacterial families that were common in soil and rhizosphere (Fig. 2.8F and 2.9F).

Specific OTUs, three from the family Streptomycetaceae and one from the order Sphingobacteriales, demonstrate the robustness of EC enrichments (Fig. 2.10a–d and Fig.

2.11a–d). A few OTUs were either significantly enriched in rhizosphere but not in the EC (Fig. 2.10e, f, and Fig. 2.11e, f), or were associated with one of the two developmental stages (Fig. 2.10g, h, and Fig. 2.11g, h). Data in Figs. 2.8, 2.9, 2.10, and 2.11 demonstrate that entire taxa at various levels are enriched in or depleted from the EC microbiome. Additionally, rhizosphere taxa capable of colonizing the root vicinity are nonetheless prevented from colonizing the EC. Several OTUs differentiated inbred *A. thaliana* accessions. Genotype-dependent enrichments and depletions were significant but weak. To identify accession-dependent effects specific to a soil type or a developmental stage, we fitted a partial GLMM that modelled each genotype against bulk soil for each experiment or developmental stage group, and tested the model's predictions with a non-parametric Kruskal–Wallis test corrected for multiple testing (Methods). We considered only those significant accession-dependent effects that were present in the same direction in both biological replicates. We further required that these OTUs have a consistent prediction in the full GLMM, which narrowed the field to 12 OTUs (or 27 with frequency-normalized data). In Fig. 2.10, we display relative abundances of two such OTUs, one for each soil type, both Actinobacteria (Fig. 2.10i, j and Fig. 2.11i, j). That these enrichments were detected by the full GLMM (which accounts for plate effects due to 454 sequencing), and were sequenced over several plates (Fig. 2.13) supports a true genotype effect. Thus, a small subset of the EC microbiome is likely to be quantitatively influenced by host-genotype-dependent fine-tuning in specific soil environments. This could allow compensatory contributions of the EC microbiome and host genome variation to overall metagenome function.

Because the rhizoplane is stripped during preparation of EC fractions, we confirmed the presence of live bacteria on roots using catalysed reporter deposition and fluorescence in situ hybridization (CARD–FISH) to whole Col-0 root segments (Eickhorst and Tippkötter 2008). Eubacteria were common on unsonicated roots (Fig. 2.13a). Actinobacteria detected

with probe HGC69a were visible on the surface of roots grown in Mason Farm soil, and co-localized with a subset of the eubacterial signals using double CARD–FISH (Fig. 2.13b), suggesting that their enrichment in EC fractions either comes from, or egresses through, the rhizoplane. Similarly, we confirmed the rare presence on the rhizoplane of Bradyrhizobiaceae (Fig. 2.14c), a family with members defined by the GLMM as more abundant in Mason Farm rhizosphere than Mason Farm EC (Fig. 2.10f and Fig. 2.11f). We enumerated the relative number of CARD–FISH signals on a set of filters made from equal amounts of material harvested in the same way as were the samples processed for pyrotag sequencing (Fig. 2.14a, b). We confirmed that Actinobacteria were found in higher abundance, and that Bradyrhizobiaceae were present in lower abundances, in EC samples than in the bulk soil and rhizosphere samples. We also noted that emerging lateral roots were typically heavily colonized by a variety of bacteria (Fig. 2.14d) consistent with previous observations (Chi et al. 2005). These results are PCR-independent support for our sequencing methods.

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

A subset of the soil bacterial population is typically enriched in rhizosphere samples (Dennis et al. 2010). Thus, a diverse bacterial community can surround the root surface and

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

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

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

MATERIALS AND METHODS

General strategy

Seed sterility was verified by plating and deep-sequencing of homogenates from sterile seedlings (Fig. 2.15). We established seedling growth, harvesting and DNA preparation pipelines as detailed in the specific sections below. We defined the bacterial community within each soil, and the community associated with plant roots across a number of controlled experimental variables: soil type, plant sample fraction, plant age and plant genotype. For plant age, we harvested roots from two developmental stages: at the formation of an inflorescence meristem (yng) and during fruiting when ≥50% of the rosette leaves were senescent (old). The former represents plants at the peak of photosynthetic conversion to carbon, whereas the latter represents a stage well after the source–sink shift has occurred, marking the change in carbon allocation from vegetal to reproductive utilization (Masclaux et al. 2000). We prepared two microbial sample fractions from each individual plant: a rhizosphere (bacteria contained in the layer of soil covering the outer surface of the root system that could be washed from roots in a buffer/detergent solution), and EC (bacteria from within the plant root system after sonication-based removal of the rhizoplane; Fig. 2.1). We also collected control soil samples (soil treated in parallel, but without a plant grown in it).

Soil collection and analysis

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

Clayton (CL) soils at a soil:sand ratio of 2:1. Soil micronutrient analysis was performed on pure and 2:1 mixed soils by the University of Wisconsin soil testing labs.

Seed sterilization and germination

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

To address whether there were seed-borne microbes that might survive surface sterilization, one-week-old seedlings were taken from sterile MS plates and homogenized by aseptic bead beating under non-bacteriolytic conditions (three 3-mm glass balls per 2-ml tube, with 300-µl PBS, using a FastPrep from MP Bio at speed 4.0 m s−1 for 10 s). The homogenate was streaked onto 1/10-strength LB, 1/10-strength '869' and KB media. No colonies were observed. To detect potential unculturable microbes, we pyrosequenced 16S amplicons from the same homogenates using bacteriolytic DNA preps from the genotypes Col-0, Cvi-0, Sha-0 and Tsu-0 (Fig. 2.15). Each accession was individually barcoded and sequenced with 1114F and 1392R, yielding 21,935, 20,747, 23,141 and 20,272 reads, respectively. A matching number of total reads was sampled from each accession using pooled data from the full experimental data set for comparative analysis. Thus, 86,095 high-quality reads were obtained from both non-sterile plants and sterile plants, the majority of which were chloroplast sequences. See Fig. 2.15 for results.

Seedling growth

One-week-old healthy seedlings were aseptically transplanted from MS plates to sterile (autoclaved) 2.5-inch-square pots filled with either MF or CL soil, with one seedling per pot. Seedlings were transferred by lifting from underneath the cotyledon leaves using

open tweezers; no pressure was applied to the hypocotyl. Some pots were designated 'bulk soil' and were not given a plant. All pots, including bulk soil controls, were always watered from the top with a shower of distilled water (non-sterile) as an accessible proxy for rain water that avoids chlorine and other tapwater additives. Pots were spatially randomized and placed in growth chambers providing short days of 8 h light (800–1,000 lx) at 21 °C and 16 h dark at 18 °C. The use of short days was to help synchronize flowering time between *A. thaliana* genotypes and to facilitate robust rosette and root growth. After harvesting the floral transition developmental stage, remaining plants and bulk soils were moved from the growth chamber to 16-h days in the greenhouse to promote a more synchronized flowering and senescence for the senescent developmental stage.

Harvesting

Each plant was killed and harvested at one of two developmental time points: (1) at the floral transition and (2) after fruiting when senescence is well underway. We considered the floral transition to have begun when the shoot apical meristem was first apparent in five or more plants. Cvi-0, Sha-0 and Ct-1 occasionally flowered one to two weeks earlier under our conditions than the other *A. thaliana* genotypes. The senescence harvest began when five or more plants showed 50% or more yellow and/or brown rosette leaves (S. Levey 2005); this occurred approximately four to five weeks after transfer to the greenhouse. Senescence occurred in the same order as bolting (flowering).

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

The aboveground plant organs were aseptically removed. Loose soil was manually removed from the roots by kneading and shaking with sterile gloves (sprayed with 70% EtOH) and by patting roots with a sterile (flamed) metal spatula—this 'neighboring soil' fell to the sterile (flamed) work surface. We followed the established convention of defining

rhizosphere soil as extending up to 1 mm from the root surface (van Elsas et al. 1988) and we removed loose soil on all root surfaces until remaining aggregates were within this range. Roots were placed in a clean and sterile 50-ml tube containing 25 ml phosphate buffer (per litre: 6.33 g of NaH2PO4·H2O, 16.5g of Na2HPO4·7H2O, 200 µl Silwet L-77). Tubes were vortexed at maximum speed for 15 s, which released most of the rhizosphere soil from the roots and turned the water turbid. The turbid solution was then filtered through a 100-µm nylon mesh cell strainer into a new 50-ml tube to remove broken plant parts and large sediment. The roots were transferred from the empty tube to a new sterile 50-ml tube with 25-ml sterile phosphate buffer, and the turbid filtrate was centrifuged for 15 min at 3,200g to form a pellet containing fine sediment and microorganisms.

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

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

DNA extraction

To extract DNA, the samples were resuspended in a lysis buffer and microbial cells were mechanically lysed through bead beating. For all bulk soil and rhizosphere data, bead beating and purification were performed with the MoBio PowerSoil kit (SDS/mechanical lysis) because of its unmatched ability to remove humics and other PCR inhibitors in our soil. EC DNA from Arabidopsis experiments was prepared with the MP Bio Fast DNA Spin Kit for soil (also a SDS/mechanical lysis) because the more intense bead-beating protocol and lysis matrix gave improved lysis of whole roots and higher DNA yield, and soil PCR inhibitors were less of a problem with these samples. Our procedure yielded around 1 µg of DNA per rhizosphere sample, and more total DNA for EC samples (although a significant portion of EC DNA sequenced was of host origin). Although MoBio Powersoil and MP Bio Fast DNA use highly similar bead-beating/mechanical lysis methods, we developed a custom method of sample pre-homogenization that allowed us to prepare some EC samples using the MoBio kit. A comparison of Col-0 fractions soil, rhizosphere and EC across four soil digs of MF, where EC was prepared using MoBio in two digs and MP Bio in the other two digs, shows that although we cannot rule out a slight kit effect, both kits produce highly similar clustering separating EC from rhizosphere and soil fractions (Fig. 2.7, replicates 3 and 4). DNA quantity was assessed with the Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen) and a plate fluorospectrometer.

PCR

For each 1114F-barcoded 1392R primer set, PCR reactions with ~10 ng of template were performed in triplicate along with a negative control to reveal contamination. The PCR program used was 95 °C for 3 min followed by 30 cycles each of 95 °C for 30 s, 55 °C for 45 s and 72 °C for 1 min, followed by 72 °C for 10 min and then cooling to 16 °C. We first verified that the no-template control did not contain DNA via gel electrophoresis, and then pooled the three replicate PCR products and quantified DNA from each pool with PicoGreen (Invitrogen). Pooled PCR products from 30–48 barcoded samples were then combined in equimolar ratios into a master DNA pool, which was cleaned with Mo-Bio UltraClean PCR Clean-Up kit before submission for standard JGI pyrosequencing using a half-plate of Roche 454-FLX with titanium reagents.

454 pyrotag sequencing

To identify organisms present in each sample, 454 sequencing of the SSU rRNA genes was performed. For 454 sequencing, the SSU rRNA genes present in each sample were amplified with the primers 1114F and 1392R containing the 454 adaptors (Engelbrektson et al. 2010). Each sample was assigned a reverse primer with a unique 5-bp barcode, allowing 30–48 samples to be pooled per half-plate. In preparation for sequencing, working aliquots of the master pool were immobilized on beads and amplified by emulsion PCR, the emulsion was broken with isopropanol, DNA-carrying beads were enriched and the enriched beads were loaded on the instrument for sequencing. During the emPCR protocol, we reduced the amplification primer amount from 460 µl in the standard protocol to 58 µl per emulsion cup. This is the same amount of primer used for the paired-end emPCR protocol. One-and-three-quarter million beads were loaded in each plate region (reduced from 2,000,000 beads per region in the standard protocol). A detailed standard protocol is available on request.

Primer test and technical reproducibility

We first tested three sets of broad-specificity 16S rRNA 5′ primers (Engelbrektson et al. 2010) (Fig. 2.2a,b) and established technical reproducibility metrics. We used 13 samples chosen from each of the three sample fractions (soil, rhizosphere and EC) and both soil types (MF and CL) (Fig. 2.2c). Each sample was amplified individually with each of the forward primers (804F, which broadly targets bacteria and archaea; 926F, a universal primer; and 1114F, which broadly targets bacteria), paired with the barcoded universal reverse primer (1392R) and sequenced twice to measure technical reproducibility. We identified bacteria by grouping highly similar (97% identity) sequences into OTUs (Methods). We chose 1114F for our experiments, on the basis of its broad coverage of the bacterial domain (Lane 1991) and higher usable data yield (Fig. 2.2f–i and Fig. 2.16).

We identified bacteria present by grouping highly similar (97% identity) sequences into OTUs using a standard QIIME (quantitative insights into microbial ecology)-based pipeline (Caporaso et al. 2010) with default settings; thus, this stand-alone test consists of a different set of OTUs than those described in this work. The primer test samples are included in our submitted data and are found on 454 half-plates 26b and 27a. The

progressive drop-out analysis, displaying the coefficient of determination (R^2) of the least-squares regression between the two technical replicates as low-abundance OTUs are sequentially discarded, was calculated using the software R with a custom script.

Primer, specificity, sequence

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

926F universal: 5′-actcaaaggaattgacgg-3′.

1114F bacteria: 5′-gcaacgagcgcaaccc-3′.

1392R barcoded universal: 5′-XXXXXacgggcggtgtgtrc-3′.

Sequence processing pipeline and assignment of OTUs

As each 454 plate was sequenced, raw reads from individual plates were immediately run through PYROTAGGER (Kunin and Hugenholtz 2010) to diagnose plate quality so that plates could be re-queued if necessary. Plates with a reasonable number of long, high-quality raw reads with matching barcodes were used in the final analysis of OTU picking and taxonomy assignment. Using QIIME-1.4.029, short reads were removed and the remaining reads were trimmed to 220 bp, and low-quality reads were removed from the analysis using default quality settings (http://qiime.org/scripts/split_libraries.html). These high-quality sequences were clustered into OTUs using a custom script derived from otupipe (http://drive5.com/otupipe). The three main steps used from otupipe include (1) de-replicating sequences to reduce the size of the data set and the run time of clustering analysis, (2) de-noising sequences by forming clusters of 97% identity and representing these with the consensus sequence, and (3) forming OTUs by clustering de-noised consensus sequences at 97% identity.

The consensus sequence of sequences in each OTU was used as a representative sequence. Each representative sequence was assigned a taxonomy by two methods: (1) using the RDP classifier (Sul et al. 2011) trained on the 4 February 2011 Greengenes reference sequences and (2) by assigning the Greengenes (DeSantis et al. 2006) taxonomy of the best BLAST hit within a combined database including the complete Greengenes 16S

database and 18S *A. thaliana* sequences from NCBI. By the BLAST-based method, sequences without a hit below the E-value threshold of 0.001 are considered unclassified.

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

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

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

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

We also created a rarefied table; because some samples, particularly samples from the EC, had fewer than 1,000 usable reads in the unnormalized usable OTU table, counts from independent samples sharing the same soil type, genotype, fraction, age and experiment were pooled to make groups of at least 1,000 reads, and the sample names were changed to reflect the pooling that had taken place. Then all samples were rarefied to 1,000 counts using the rrarefy() function in the vegan package of R (Oksanen et al. 2011).

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

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

Because the majority of OTUs were represented by a very small number of reads and these OTUs were not technically reproducible (Fig. 2.2d, e), both the rarefaction-normalized and the frequency-normalized OTU tables were thresholded to generate measurable OTUs for the majority of analyses (the major exception being the UniFrac analysis in Fig. 2.5: weighted UniFrac distance is robust to rare OTUs). An OTU was deemed measurable if and only if there were ≥25 reads in ≥5 samples in the unnormalized usable OTU table. As described in the text and Fig. 2.2, this threshold was derived from the fact that the correlation between abundance in the same OTU in technical replicates improved greatly as OTUs approached an abundance of 25 reads, and from the fact that although contamination might create an OTU at this abundance once, the probability of an OTU being spurious decreases greatly if it occurs at a measurable level in several (we chose ≥5) independent samples.

Detection of differentially enriched OTUs by the GLMM

The OTU abundances were analyzed with a GLMM to estimate the effect of the different variables on each measurable OTU. The lme4 R package (Bates et al. 2011) was used to fit the model. The abundance of each OTU on each sample (y_{ij}) was $\log_2$-transformed and modelled as a function of the abundance of the same OTU in bulk soil samples (std_check) as a fixed effect, and plant genotype (b_1), sample type (plant or bulk soil, b_2), plant developmental stage (b_3), soil type (b_4), sequencing half-plate (b_5) and biological replicate (b_6) were modelled as random effects. The full model is specified by

$$y_{ij} = \beta \times \text{std\_check} + b_{1ij} + b_{2ij} + b_{3ij} + b_{4ij} + b_{5ij} + b_{6ij} + e_{ij}$$

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

There were not enough paired samples of rhizosphere and EC from the same individual plant to model the effect of both fractions directly. Instead, the abundance table was split into EC and rhizosphere samples, and the effect of each fraction with respect to bulk soil controls was estimated. The same model specification was used independently on both fractions, and for both the frequency and the rarefied tables (see Methods on sequence processing pipeline).

For each level of the random effects, the conditional mode and 95% prediction interval were estimated by Markov chain Monte Carlo sampling from the fitted model. A specific level is considered to have an effect on an OTU if the prediction interval of its conditional mode does not include zero.

Partial GLMM

There were not enough samples to estimate all the interaction effect between all variables without drastically reducing the size of the data set and our statistical power (Table 2.2). To assess specific interactions of the genotype effect with other variables, a constrained version of the previously defined GLMM was used that employed only the fixed effect (std_check) and the random effects for plant genotype (b_1) and sample type (b_2). Samples were split into groups of the same experiment, developmental stage and fraction (thus, all the other variables from the full model are tested within each group), and the model was fitted and analysed in the same way as the full GLMM. A non-parametric Kruskal–Wallis test was used to verify independently the predictions of the partial GLMM for significance, where P values were corrected to Q values using the Benjimani–Hochberg FDR method; predictions from each partial GLMM with a Q value >0.05 were discarded as insignificant. The intersection of the significant genotype predictions between both biological replicates of each condition was calculated.

Scanning electron microscopy sample preparation

Arabidopsis roots were fixed in 2% paraformaldehyde, 2.5% glutaraldehyde and 0.15 M sodium phosphate buffer, pH 7.4. The samples were dehydrated using a gradual ethanol series (30%, 50%, 75%, 100%, 100%) and dried in a Samdri-795 supercritical dryer

using carbon dioxide as the transitional solvent (Tousimis Research Corporation). Roots were mounted on aluminium planchets with double-sided carbon adhesive and coated with 10 nm of gold–palladium alloy (60:40 Au:Pd, Hummer X Sputter Coater, Anatech USA). Images were made using a Zeiss Supra 25 FESEM operating at 5 kV and a working distance of 5 mm, and with a 10-µm aperture (Carl Zeiss SMT Inc.), at the Microscopy Services Laboratory, Pathology and Laboratory Medicine, UNC at Chapel Hill.

Log_2 transformation

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

Heat maps

Heat maps were constructed using custom scripts and the function heatmap.2 from the R package gplots (Warnes 2011). For better visualization, all data was log_2-transformed. Hierarchical clustering of rows and columns in the heat maps is based on Bray–Curtis similarities and uses group-average linkage.

Diversity

The Shannon diversity index and the non-parametric Chao1 diversity were calculated with the vegan package in R (Oksanen et al. 2011). The exponential function was applied to the Shannon diversity index to calculate the true Shannon diversity (effective number of species).

Rarefaction curves

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

Taxonomy histograms and statistics

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

Sample clustering using UniFrac

A phylogenetic tree was built with the representative sequence for each OTU and the pairwise, normalized, weighted UniFrac distance (Lozupone and Knight 2005). For UniFrac, representative sequences from all non-plant OTUs, including those that did not meet the 25×5 sample threshold, were considered. UniFrac distances between samples are based on the fraction of branch length that is unique to each sample in a shared phylogenetic tree composed of OTU representative sequences from all samples. Thus, samples containing OTUs of highly divergent sequences will be more distant from each other, because the OTUs comprising each sample will occupy different major branches on the shared phylogenetic tree of OTUs, whereas samples containing highly similar OTUs will share these major branches. In weighted UniFrac, the branch length unique to each sample is multiplied by the frequency at which that OTU occurs in the sample. Thus, weighted UniFrac can detect differences between two samples that have the same set of OTUs that differ quantitatively between the samples.

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

CARD–FISH application to roots

We applied a modified protocol described previously (Eickhorst and Tippkötter 2008). Briefly, several root systems from a bolting Col-0 grown in MF were fixed using 4% formaldehyde in PBS at 4 °C for 3 h, washed twice in PBS and stored in 1:1 PBS:molecular-grade ethanol at −20 °C. Treatments with lysozyme solution (1 h at 37 °C, 10 mg ml^{-1}; Fluka) and achromopeptidase (30 min at 37 °C, 60 U ml^{-1}; Sigma) were sequentially used for prokaryotic cell-wall permeabilization. Endogenous peroxidases were inactivated with methanol treatment amended by 0.15% H_2O_2 at room temperature for 30 min and washed again. Probes targeting either the 16S or the 23S rRNA (EUB338 (5′-GCTGCCTCCCGTAGGAGT-3′, 35% formamide), NON338 (5′-ACTCCTACGGGAGGCAGC-3′, 30% formamide), HGC69a (5′-TATAGTTACCACCGCCGT-3′, 25% formamide) and Brady4 (5′-CGTCATTATCTTCCCGCACA-3′, 30% formamide)) were defined using probeBase (Loy et al. 2007) (http://www.microbial-ecology.net/default.asp), labelled with enzyme horseradish peroxidase on the 5′ end (Invitrogen), diluted in hybridization buffer (final concentration of 0.19 ng ml^{-1}) with each probe's optimum formamide concentration, and hybridized at 35 °C for 2 h. Unbound probes were washed away from samples in wash buffer (NaCl content adjusted according to the formamide concentration in the hybridization buffer) at 37 °C for 30 min. Fluorescently labelled tyramide was used for signal amplification, and samples were washed before mounting on glass slides.

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

For application of samples onto filters, bulk MF soil, rhizosphere and EC samples from four sets of Col-0 roots were pooled and harvested in the way described above before

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

Sample naming in OTU tables

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

Soil name	Total Minerals											
	P	K	Ca	Mg	S	Zn	B	Mn	Fe	Cu	Al	Na
	%	%	%	%	%	ppm	ppm	ppm	ppm	ppm	ppm	ppm
CL 2:1	**0.01**	**0.01**	**0.01**	**0.01**	**0.002**	**6.40**	**<2**	**15.57**	**1384.0**	**<0.5**	**3310.4**	**21.8**
CL	0.02	0.02	0.02	0.01	0.004	16.69	<2	22.84	1774.0	<0.5	4637.5	18.5
MF 2:1	**0.03**	**0.05**	**0.14**	**0.07**	**0.02**	**34.23**	**<2**	**280.70**	**5799.6**	**5.84**	**11543.7**	**35.1**
MF	0.05	0.10	0.24	0.13	0.03	58.66	<2	493.55	9546.1	11.66	20439.0	56.5

Soil name	Carbon and Nitrogen					Physical Analysis				
	NH_4	NO_3	Total C	Total N	C/N Ratio	Sand	Silt	Clay	Texture	pH
	ppm	ppm	%	%		%	%	%		
CL 2:1	**1.82**	**0.71**	**0.38**	**0.02**	**16.1**	**91**	**3**	**6**	**Sand**	**6.3**
CL	1.80	1.36	0.47	0.03	17.4	87	6	7	Loamy Sand	6.4
MF 2:1	**1.61**	**18.10**	**1.67**	**0.11**	**14.6**	**69**	**22**	**9**	**Sandy Loam**	**6.0**
MF	2.03	34.10	2.66	0.21	12.9	45	42	13	Loam	5.9

CL 2:1	Clayton 2:1 (2 parts Clayton soil : 1 part sand)	GPS Location
CL	100% Clayton soil	+35° 39' 59.45", -78° 29' 35.91"
MF 2:1	Mason Farm 2:1 (2 parts Mason Farm soil : 1 part sand)	GPS Location
MF	100% Mason Farm soil	Amount or size of sample collected

Table 2.1: Mason Farm and Clayton soil micronutrient analysis and GPS location.

a

Arabidopsis Genotypes and Seed Stocks				
Accession	**Region**	**Latitude**	**Longitude**	**Stock Center**
Col-0	USA (Germany?)	38.3	-92.3	CS22625
Ct-1	Italy	37.3	15	CS22639
Cvi-0	Cape Verde Isl.	16	-24	CS22614
Ler-1	Poland	52-53	15-16	CS22618
Mt-0	Libya	33	23	CS22642
Oy-0	Norway	60.23	6.13	CS22658
Shahdara	Tajikistan	38.35	68.48	CS22652
Tsu-0	Tsushima	34-35	136-137	CS28780

b

Table of Samples (after pooling smaller samples for normalization by rarefaction)

	Col-0		Ct-1		Cvi-0		Ler-1		Mt-0		Oy-0		Sha-0		Tsu-0		Soil	Dig date	Experiment Start
	R	EC	R	EC	R	EC	R	EC	R	EC	R	EC	R	EC	R	EC	S		
CL1 yng	11		11		9		10	1	10	4	10	4	10	3	8	5	10		
CL1 old	9	6	10	6	9	8	10	8	9	7	10	6	10	5	9	7	10	Dec-09	Feb-10
CL2 yng	5	7	10	4	5	1	3	4	4	2	7	2	7	3	7	1	8		
CL2 old	9	4	9	5	7	7	10	3	8	5	9	7	11	9	7	7	10	Dec-09	Apr-10
MF1 yng	8	3	8	4	8	4	8	8	7	6	8	2	9	8	10	6	10		
MF1 old	9	7	6	6	8	8	9	8	8	5	9	7	10	9	9	5	10	Feb-10	Mar-10
MF2 yng	10	10	7	7	10	10	10	10	8	8	10	10	10	10	8	8	10		
MF2 old	9	9	7	7	9	9	5	5	9	9	10	10	9	9	9	9	10	Apr-10	Apr-10
MF3 yng	6	6															11	Jun-10	Aug-10
MF4 yng	7	7															10	Nov-10	Jan-11

Table of Samples (Frequency)

	Col-0		Ct-1		Cvi-0		Ler-1		Mt-0		Oy-0		Sha-0		Tsu-0		Soil	Dig date	Experiment Start
	R	EC	R	EC	R	EC	R	EC	R	EC	R	EC	R	EC	R	EC	S		
CL1 yng	11		11		9		10	1	10	7	10	7	10	7	8	7	10		
CL1 old	9	8	10	8	9	9	10	8	9	7	10	7	10	8	9	8	10	Dec-09	Feb-10
CL2 yng	5	9	10	7	8	4	3	6	4	5	8	8	7	9	7	3	8		
CL2 old	9	9	9	8	7	8	10	3	8	7	9	9	11	12	7	9	10	Dec-09	Apr-10
MF1 yng	8	10	8	8	8	11	8	9	7	8	8	5	9	11	10	10	10		
MF1 old	9	8	6	8	8	10	9	8	8	5	9	8	10	9	10	5	10	Feb-10	Mar-10
MF2 yng	11	11	10	10	10	10	10	10	8	8	10	10	10	10	8	8	10		
MF2 old	9	9	7	7	9	9	5	5	9	9	10	10	9	9	9	9	10	Apr-10	Apr-10
MF3 yng	6	10															12	Jun-10	Aug-10
MF4 yng	8	8															10	Nov-10	Jan-11

Table 2.2: Genotypes, seed stocks, and sample numbers.

(a) *Arabidopsis thaliana* genotypes and seed stocks used.

(b) Number of high quality samples for the frequency-normalized table (top) and the rarefaction normalized table (bottom), in which some replicate samples were pooled to make the rarefaction threshold. Does not include the four sterile seedling samples (Figure 2.15).

	Rarefied (% variance)		Frequency (% variance)	
Variable	**R with S**	**EC with S**	**R with S**	**EC with S**
454 plate	7.53	20.76	13.31	23.72
Accession	0.40	0.42	0.88	0.68
Experiment	2.66	4.94	4.03	7.27
Age	1.66	1.11	2.93	1.55
Soil type	5.53	8.34	7.69	8.24
Fraction	7.66	25.33	12.01	16.86
Residual	74.57	39.10	59.14	41.67

Table 2.3. Percent variance explained by each variable in the Full GLMM.

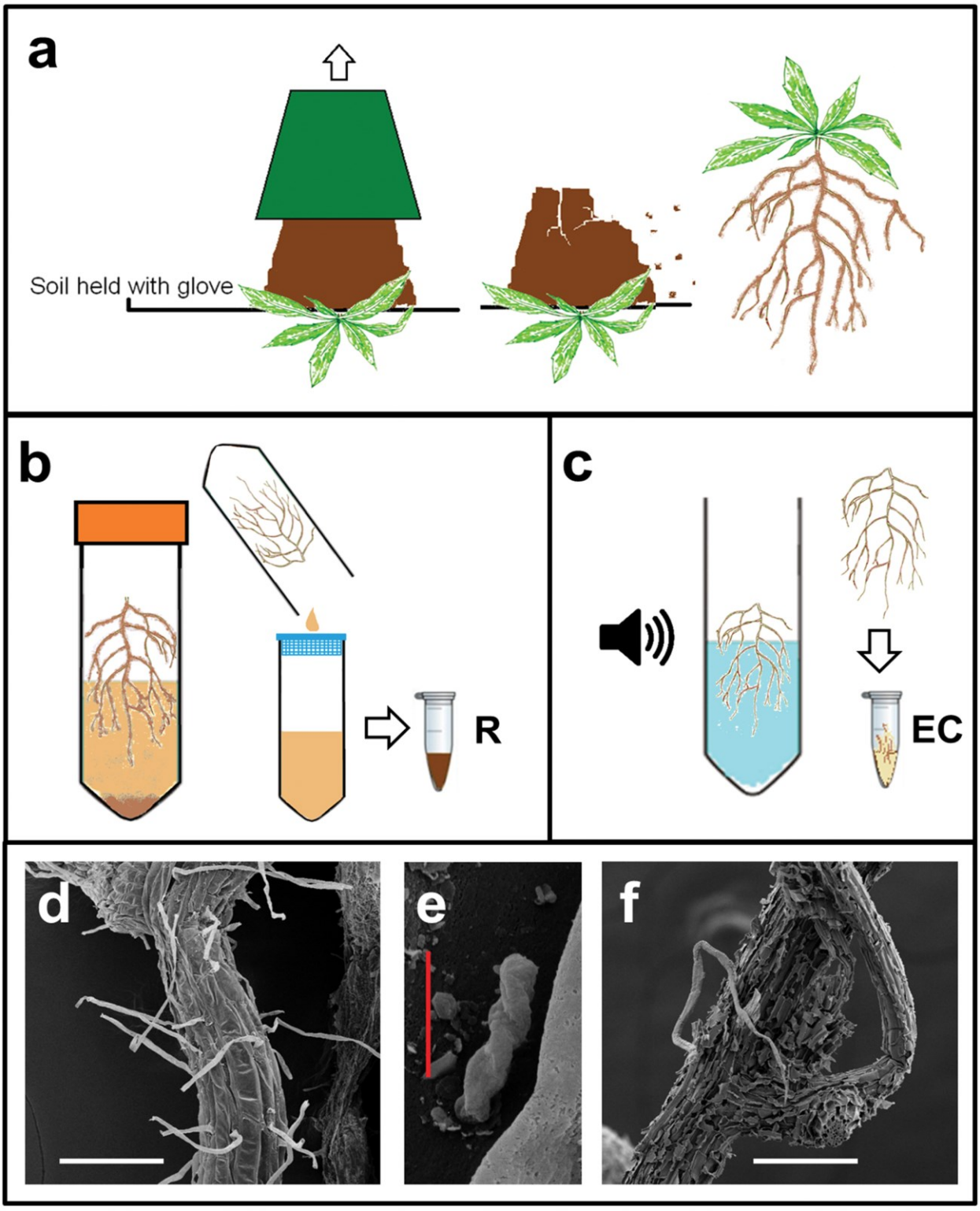

a
Soil held with glove
b
R
c
EC
d
e
f

Figure 2.1. Harvesting scheme

(a) Using gloves and a flame-sterilized work surface, plants are overturned, pots are removed, and soil is crumbled/brushed away leaving ≤1 mm rhizosphere soil on roots.

(b) The above-ground parts are cut away and rhizosphere soil is harvested from roots by shaking them in sterile phosphate buffer with Silwet L-77; the rinse is pelleted and becomes the rhizosphere R fraction.

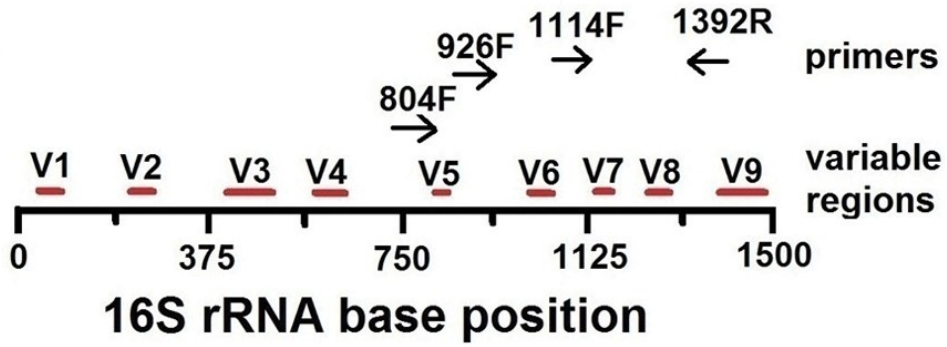
(c) Roots are placed in a new tube with sterile phosphate buffer and sonicated for five 30 second bursts at low intensity (see Methods). The surface-cleaned roots are then snap frozen and lyophilized to become the EC fraction.

(d) SEM showing intact root surface after rhizosphere soil has been removed, but prior to sonication. Scale = 100 microns.

(e) SEM showing a root-surface bacterium on root shown in **d**. Scale = 1 micron.

(f) SEM showing the disruptive clearing of nearly the entire root surface after sonication. Scale = 100 microns.

a

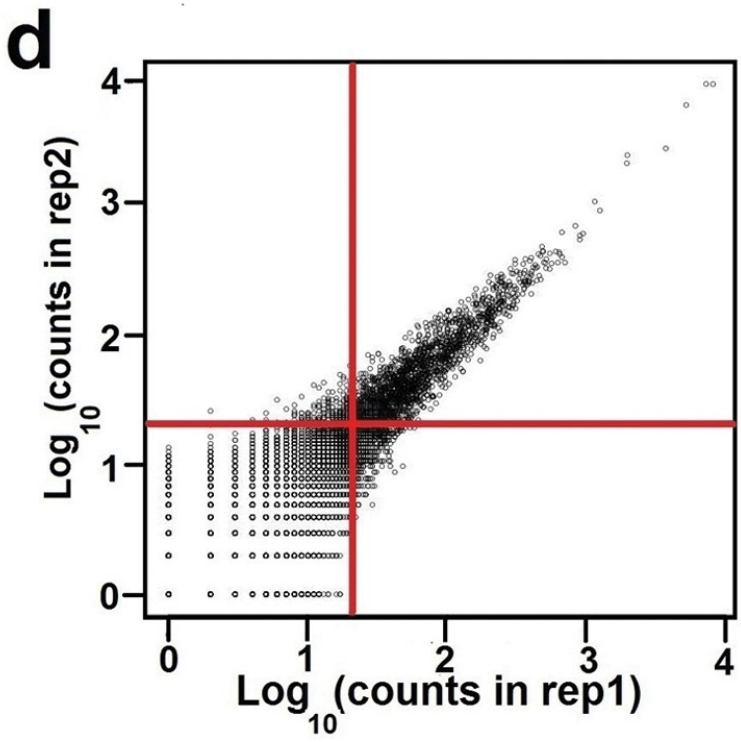


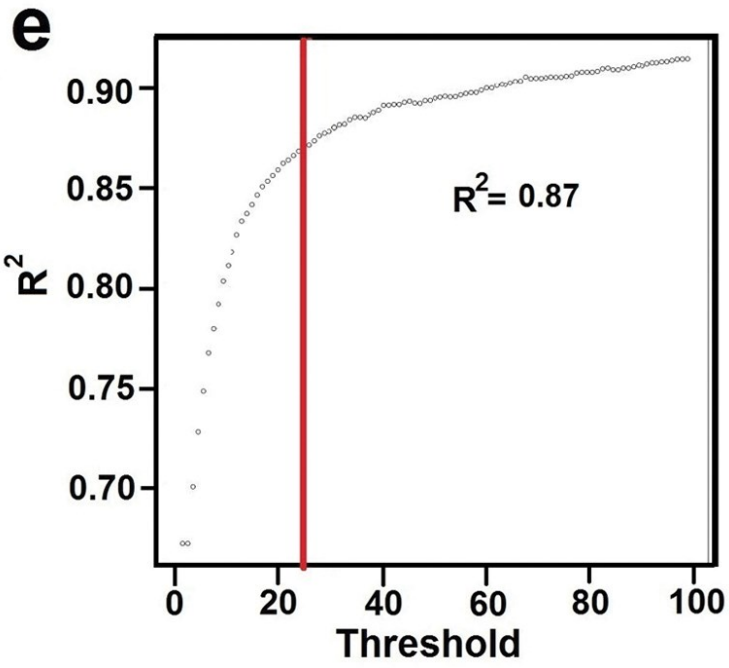
b

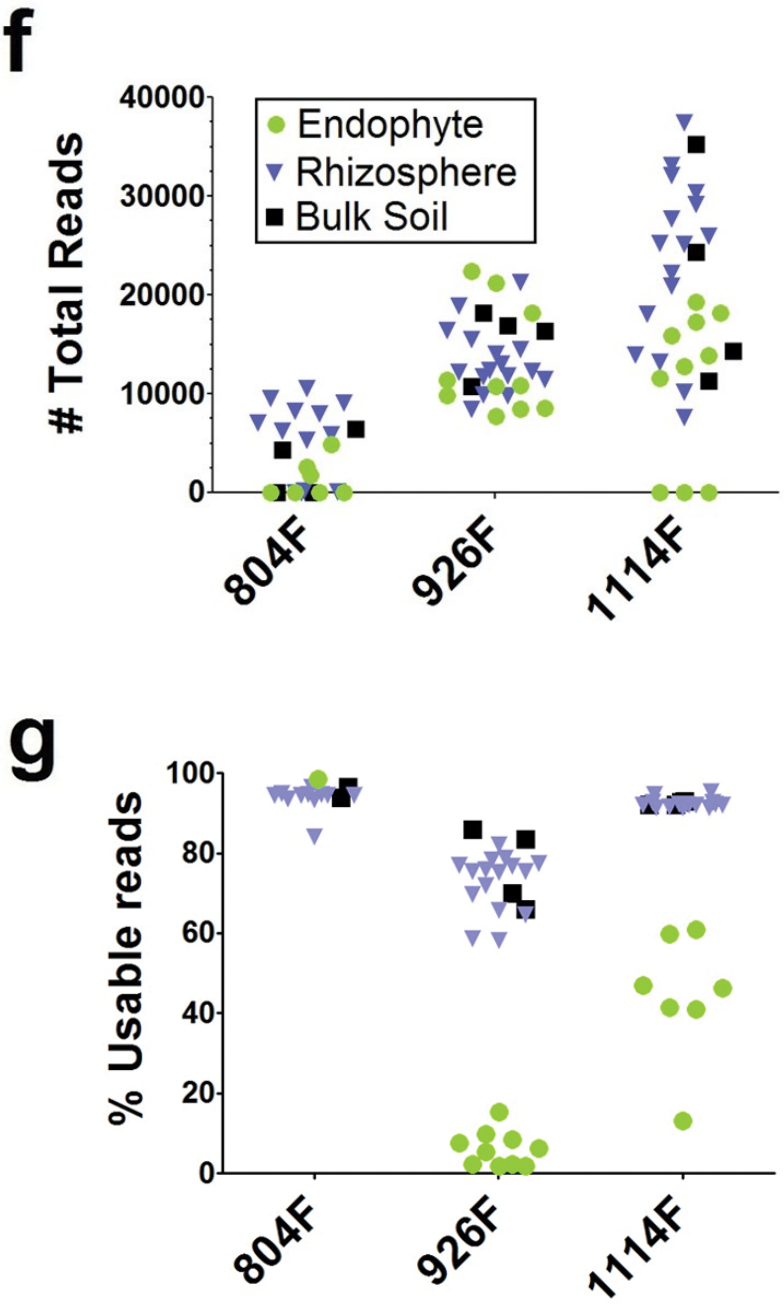
Primer Name	Sequence (L to R 454 titanium primer, linker, XXXXX=barcode, and 16S primer)			
804F	**5'**-CCTATCCCCTGTGTGCCTTGGCAGTCTC	ag	**attagataccc**DRgtagt- **3'**	
926F	**5'**-CCTATCCCCTGTGTGCCTTGGCAGTCTC	ag	**aaactYaaaKgaattgacgg-3'**	
1114F	**5'**-CCTATCCCCTGTGTGCCTTGGCAGTCTC	ag	**gcaacgagcgcaaccc-3'**	
1392R	**5'**-CCATCTCATCCCTGCGTGTCTCCGACTC	ag	XXXXX	**acgggcggtgtgtRc** - 3'
	D= A, G, or T, **R**= A or G, **Y**= C or T, and **K**= G or T			

c

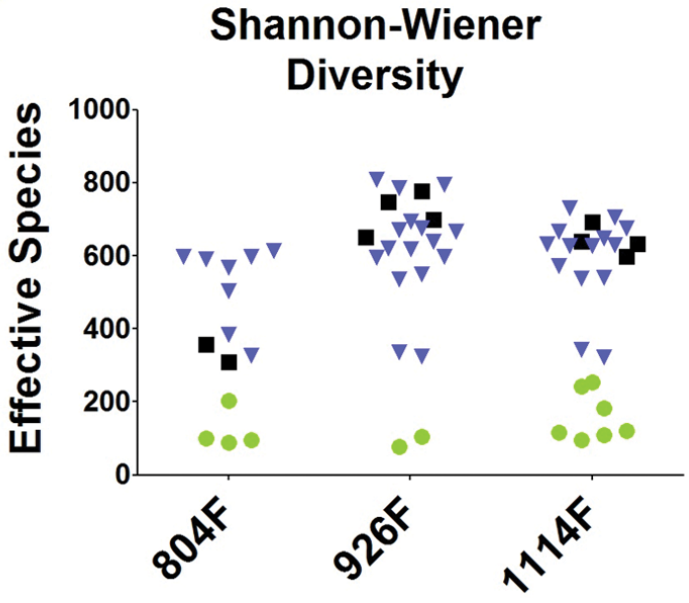
Primer Test Set		
Sample Fraction	**Soil Type**	**Number of Samples**
Bulk Soil	Mason Farm	1
Rhizosphere	Mason Farm	5
Endophyte	Mason Farm	2
Bulk Soil	Clayton	1
Rhizosphere	Clayton	2
Endophyte	Clayton	2
Total		**13**


d
Log10 (counts in rep2)
Log10 (counts in rep1)
0
1
2
3
4


e
R2
R2= 0.87
0.90
0.85
0.80
0.75
0.70
0
20
40
60
80
100
Threshold


f
Total Reads
40000
30000
20000
10000
0
Endophyte
Rhizosphere
Bulk Soil
804F
926F
1114F
g
% Usable reads
100
80
60
40
20
0
804F
926F
1114F

h


Shannon-Wiener Diversity
Effective Species
1000
800
600
400
200
0
804F
926F
1114F

i

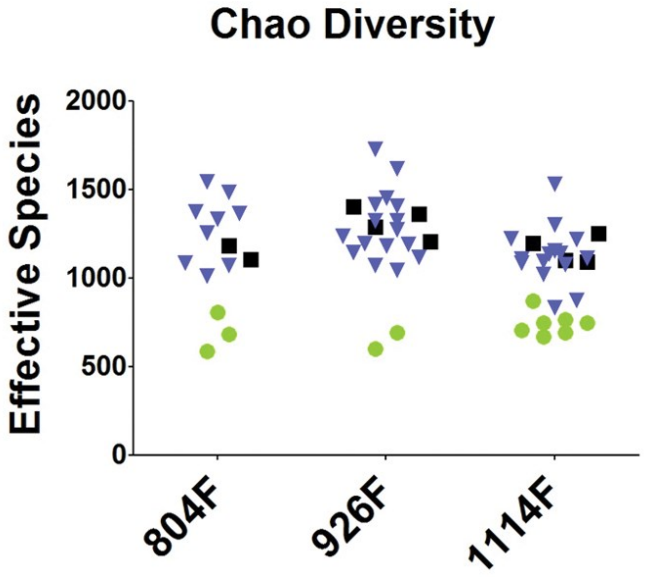

Chao Diversity
Effective Species
2000
1500
1000
500
0
804F
926F
1114F

Figure 2.2. Primer test and technical reproducibility

(a) Position on the 16S gene of each of the primers tested.

(b) Sequence of each primer used.

(c) Composition of the 13 samples tested.

(d) Log10 transformation of raw reads per OTU for one independent replicate (x-axis) vs. the other (y-axis), where both replicates were PCR-amplified and sequenced from the same sample (axes labels are transformed and cover a range of 0-10,000 reads). The intersection of the red lines shows where an OTU with 25 reads in both replicates would lie.

(e) Progressive drop-out analysis displaying the R^2 correlation of the data in **d** as OTUs with low read numbers are discarded. When only OTUs with ≥25 reads are considered (red line) the R^2 is acceptable at 0.87, a balance between reproducibility and data loss for low-abundance OTUs. In **f-i,** green circles are EC samples, blue triangles are R samples, and black squares are bulk soil samples.

(f) Total reads obtained from amplicons made with 804F, 926F, or 1114F paired with bar-coded 1392R.

(g) Percent of the 'usable' reads from **f** which are not identified as plant or chimeric OTUs.

(h) Shannon-Weiner species diversity of 1000 usable reads (for each sample with ≥1000 reads).

(i) Chao1 diversity of 1000 usable reads from each sample (for each sample with ≥1000).

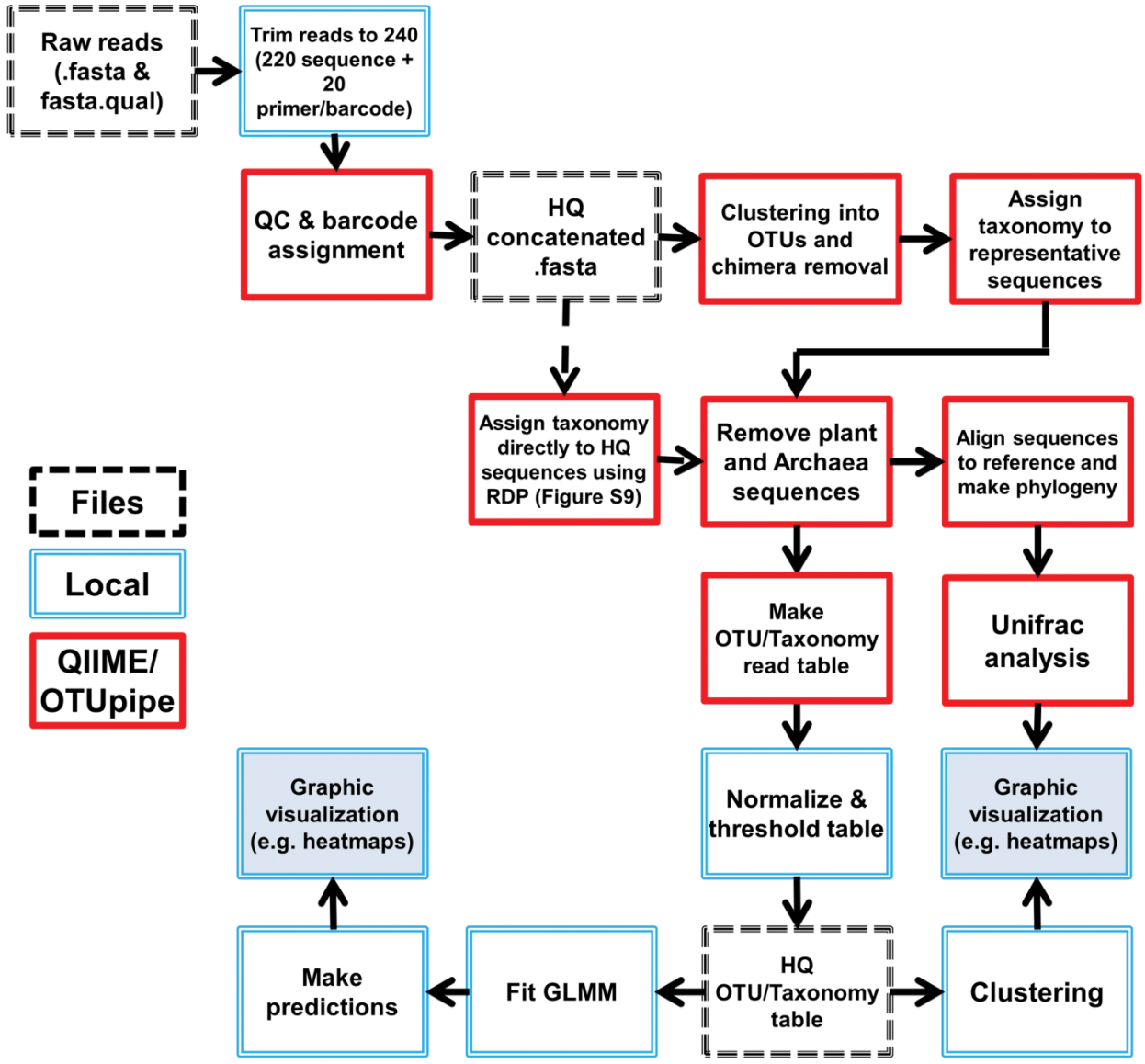
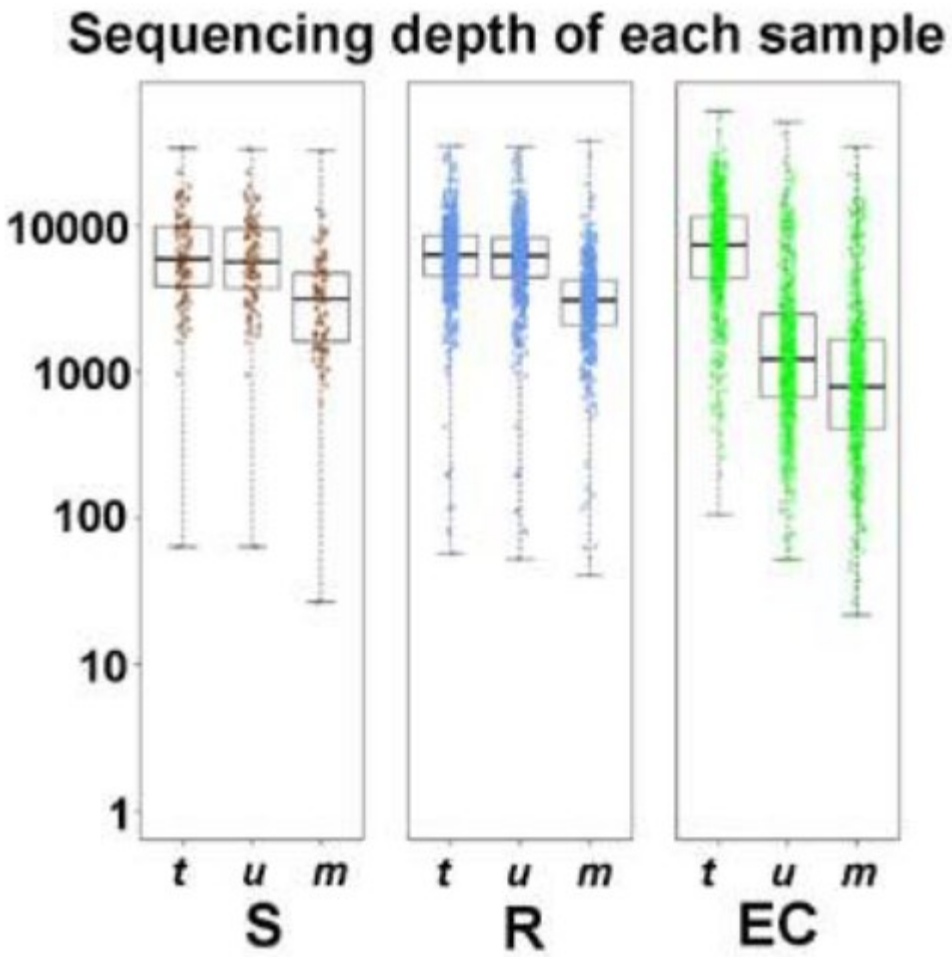


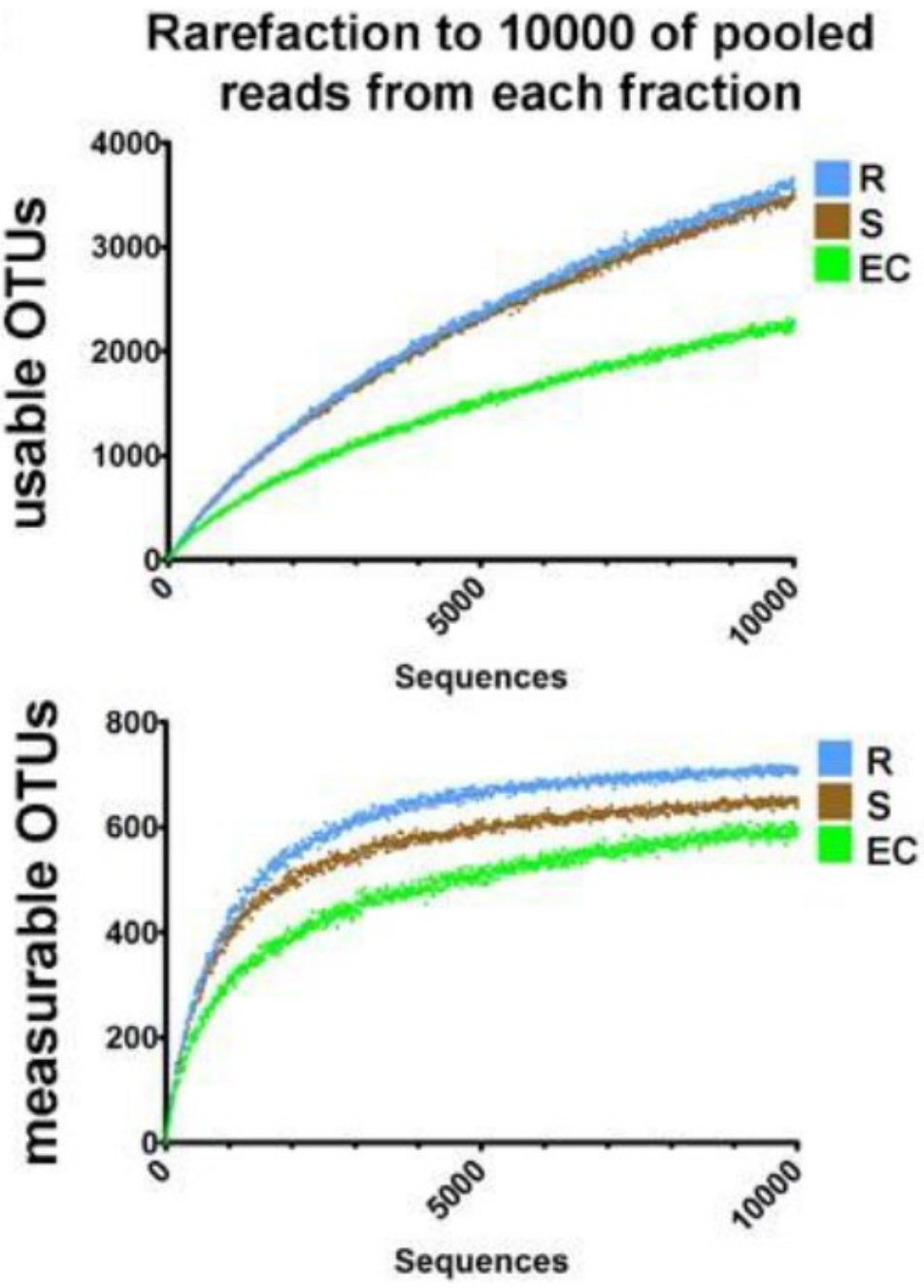
Figure 2.3. Informatics pipeline

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

a


Sequencing depth of each sample
10000
1000
100
10
1
t u m
S
t u m
R
t u m
EC

b


Rarefaction to 10000 of pooled reads from each fraction
usable OTUs
4000
3000
2000
1000
0
0
5000
10000
Sequences
R
S
EC
measurable OTUs
800
600
400
200
0
0
5000
10000
Sequences
R
S
EC

c

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

d

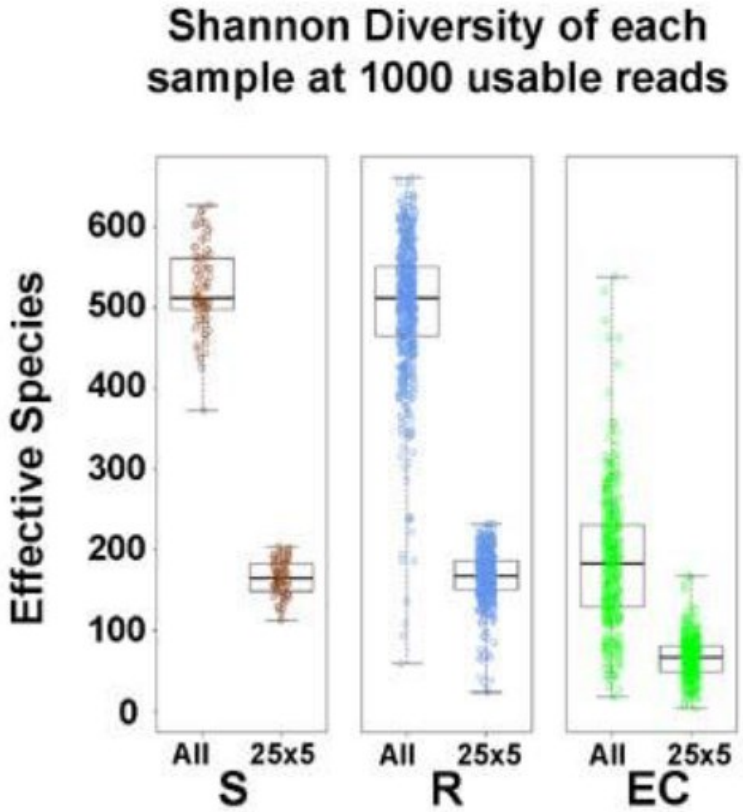


Figure 2.4. Sequencing statistics and quality.

(a) Sequencing depth per sample in reads for the three sample fractions S, R, and EC. Each dot represents a single plant or soil sample. Within each fraction, the total (*t),* usable *(u)*, and measurable *(m)* read counts are shown for all samples. The box plots contain the 1st and 3rd quartiles, split by the median; whiskers extend to include the farthest outliers.

(b) Rarefaction curves to 10,000 sequences for cumulative reads from S, R, and EC fractions considering all usable OTUs (top) and only measurable OTUs (bottom)

(c) Table, split by sample fraction, summarizing: cumulative numbers of total high quality reads, 'usable' (non-plant & non-chimera) reads, number of OTUs after the technical reproducibility '25x5' threshold is applied, 'measurable' reads (reads contained in OTUs that pass the 25x5 threshold).

(d) Shannon diversity of individual samples from each fraction, calculated from the rarefaction-normalized table, before (left) and after (right) applying the 25x5 measurable OTU threshold.

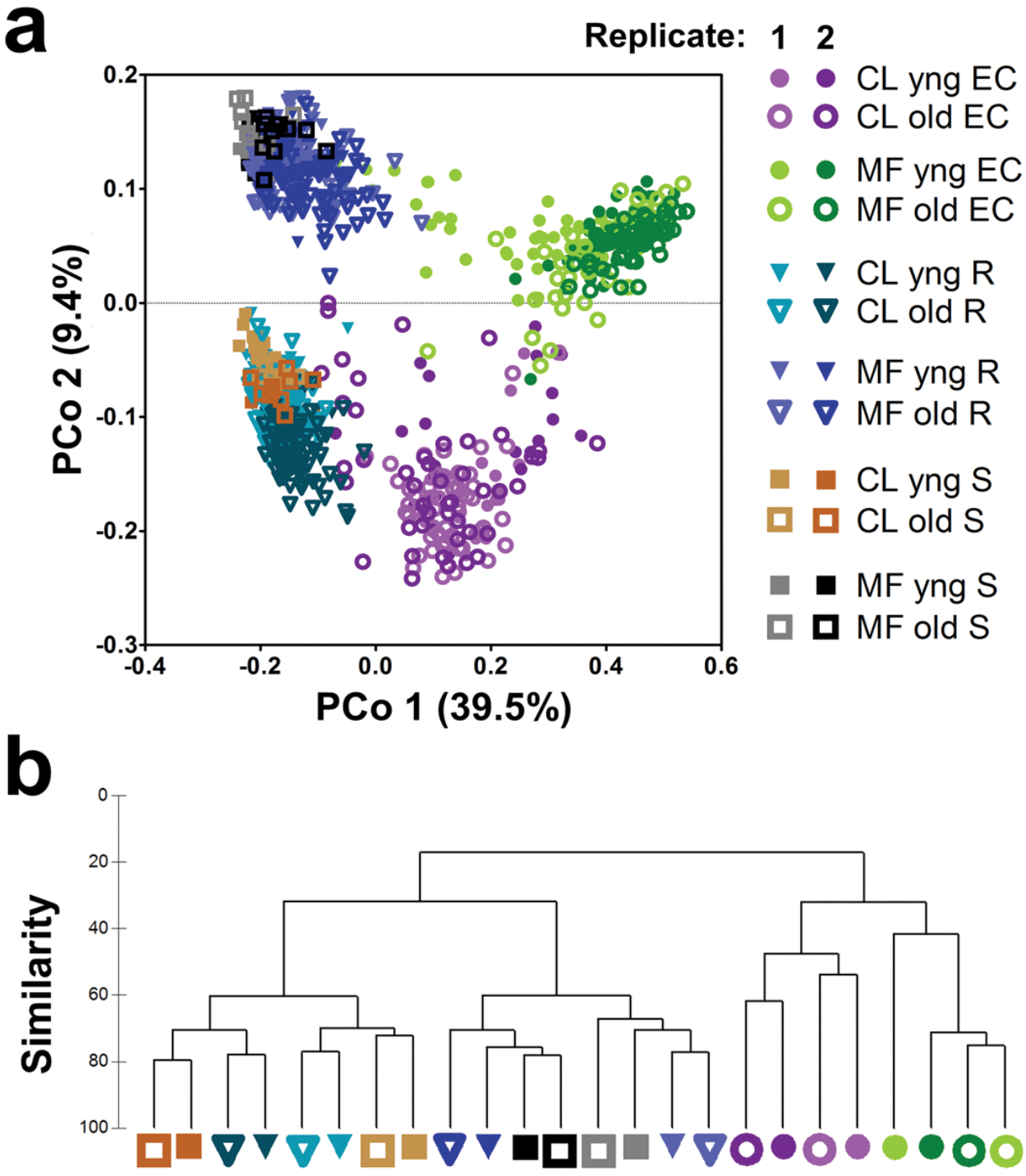

a
Replicate: 1 2
CL yng EC
CL old EC
MF yng EC
MF old EC
CL yng R
CL old R
MF yng R
MF old R
CL yng S
CL old S
MF yng S
MF old S
PCo 2 (9.4%)
PCo 1 (39.5%)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.2
0.0
0.2
0.4
0.6
b
Similarity
0
20
40
60
80
100

Figure 2.5. Sample fraction and soil type drive the microbial composition of root-associated endophyte communities.

(a) Principal Coordinate Analysis (PCoA) of pairwise normalized weighted Unifrac distances between samples based on rarefaction to 1,000 reads in unthresholded, usable OTUs. CL, Clayton; MF, Mason Farm; R, rhizosphere; S, soil

(b) Rarefied counts for the 2535 thresholded, measurable OTUs from each of 24 soil, stage or fraction groups were $\log_2$-transformed (Methods) to make 24 representative samples (branch labels), and pairwise Bray–Curtis similarity was used to cluster these representatives hierarchically (group-average linkage).

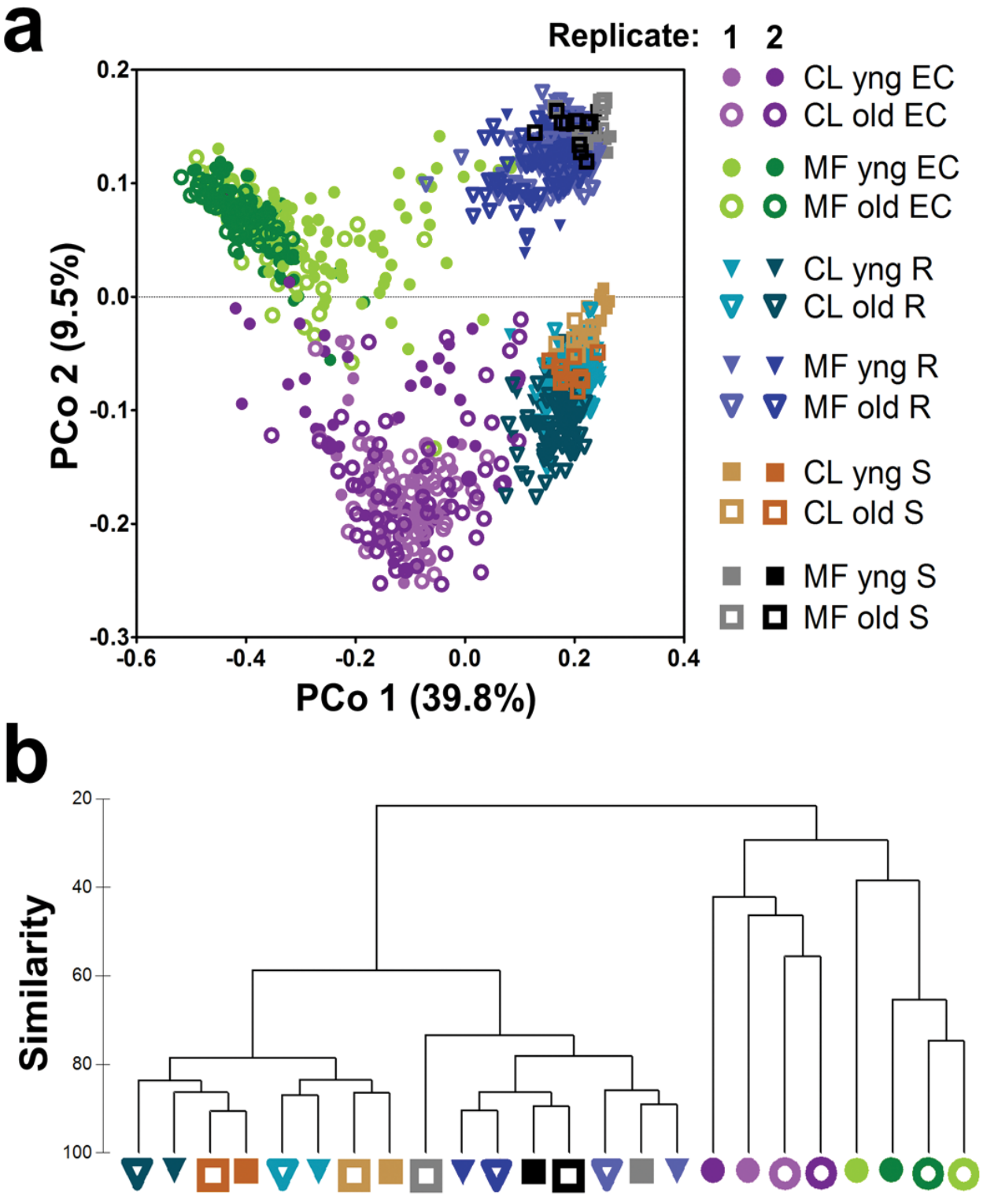

a
Replicate: 1 2
CL yng EC
CL old EC
MF yng EC
MF old EC
CL yng R
CL old R
MF yng R
MF old R
CL yng S
CL old S
MF yng S
MF old S
PCo 2 (9.5%)
PCo 1 (39.8%)
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.6
-0.4
-0.2
0.0
0.2
0.4
b
Similarity
20
40
60
80
100

Figure 2.6. Sample fraction and soil type drive the microbial composition of root-associated endophyte communities.

(a) Principal Coordinate Analysis (PCoA) of pairwise normalized weighted Unifrac distances between the samples considering relative abundance of all (*unthresholded*) OTUs.

(b) The median RAs for the *25x5 thresholded* 'measurable' OTUs from each of 24 soil/stage/fraction groups were $\log_2$ transformed (see methods) to make 24 representative samples (branch labels) and the pairwise Bray Curtis Similarity was used to hierarchically cluster these representatives (group average linkage).

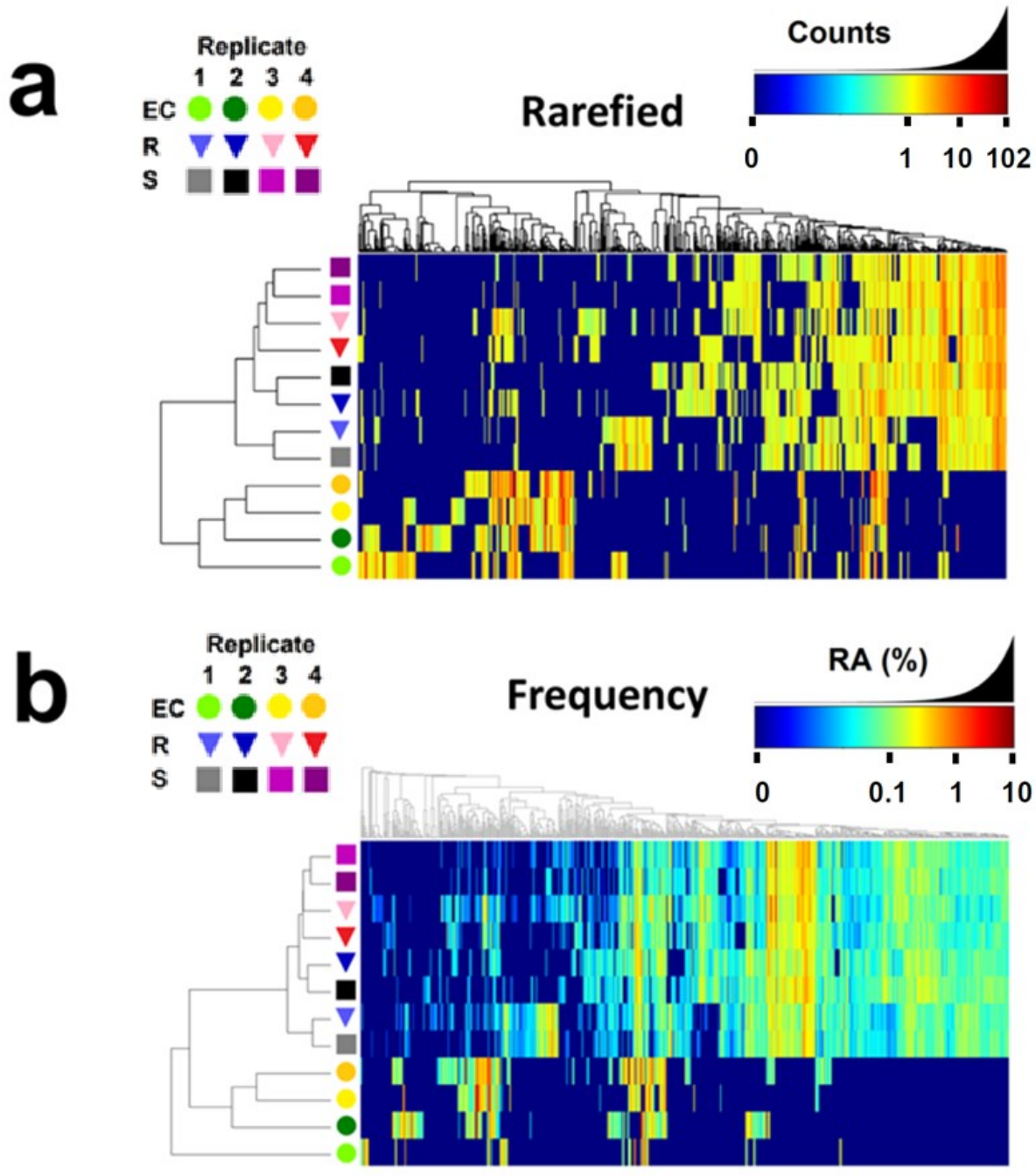
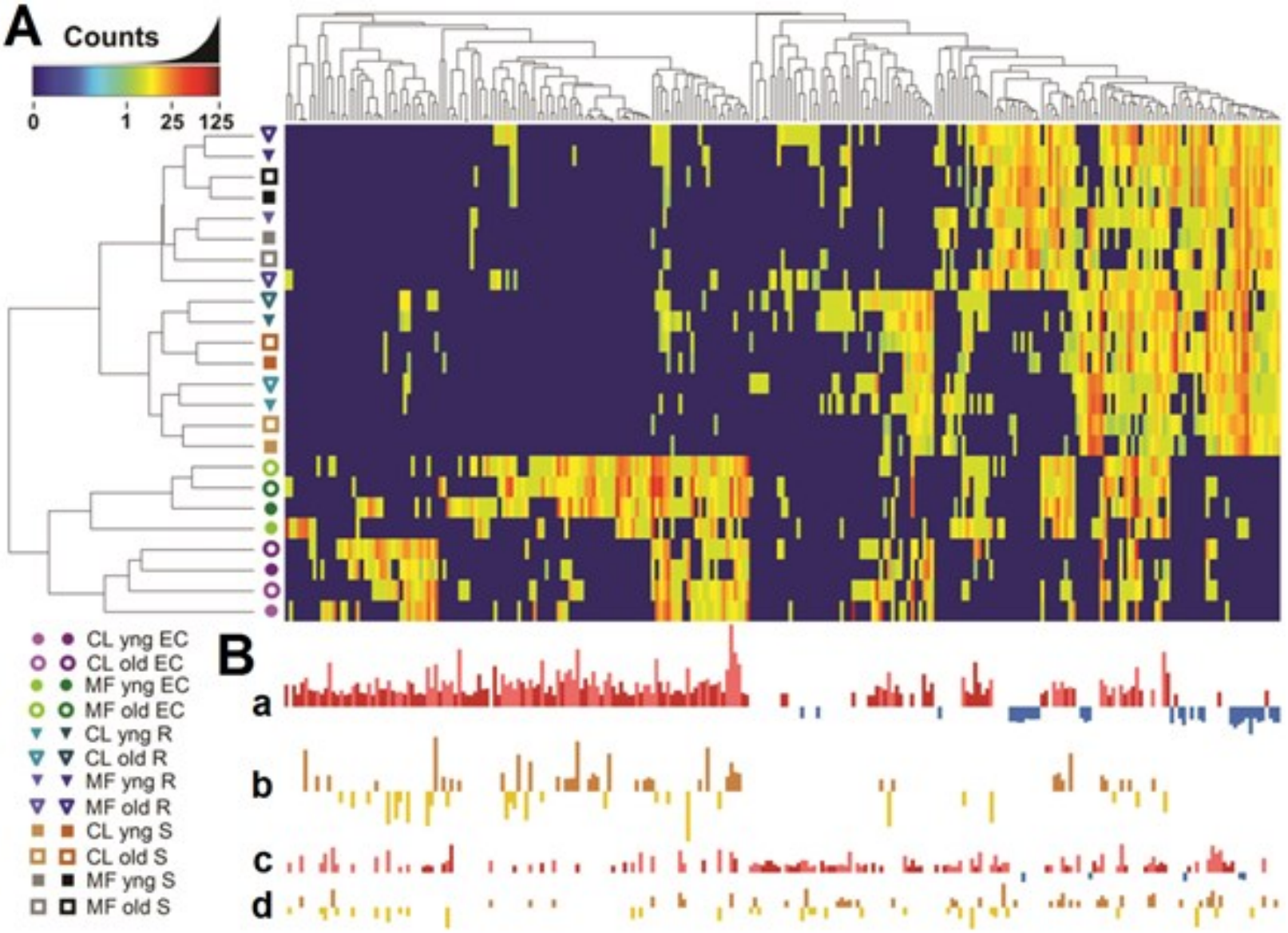

a
Replicate
1 2 3 4
EC
R
S
Rarefied
Counts
0
1
10
102
b
Replicate
1 2 3 4
EC
R
S
Frequency
RA (%)
0
0.1
1
10

Figure 2.7. OTUs identified from four independent biological replicates are reproducible.

Heat map displaying the reproducibility between four independent replicates at the yng developmental stage of bulk soil (squares), Col-0 R samples (triangles), and Col-0 EC samples (circles). Each symbol represents the median of six or more samples. All data were $\log_2$ transformed for visualization, but for ease of interpretation the quantities shown in the color key represent the original (untransformed) counts (in panel a) and frequencies (in panel b) for each color. Although all 778 measurable OTUs were included, some OTUs had a median of 0 in all Col-0 and soil groups shown and were removed from the display.


A
Counts
0
1
25
125
CL yng EC
CL old EC
MF yng EC
MF old EC
CL yng R
CL old R
MF yng R
MF old R
CL yng S
CL old S
MF yng S
MF old S
B
a
b
c
d

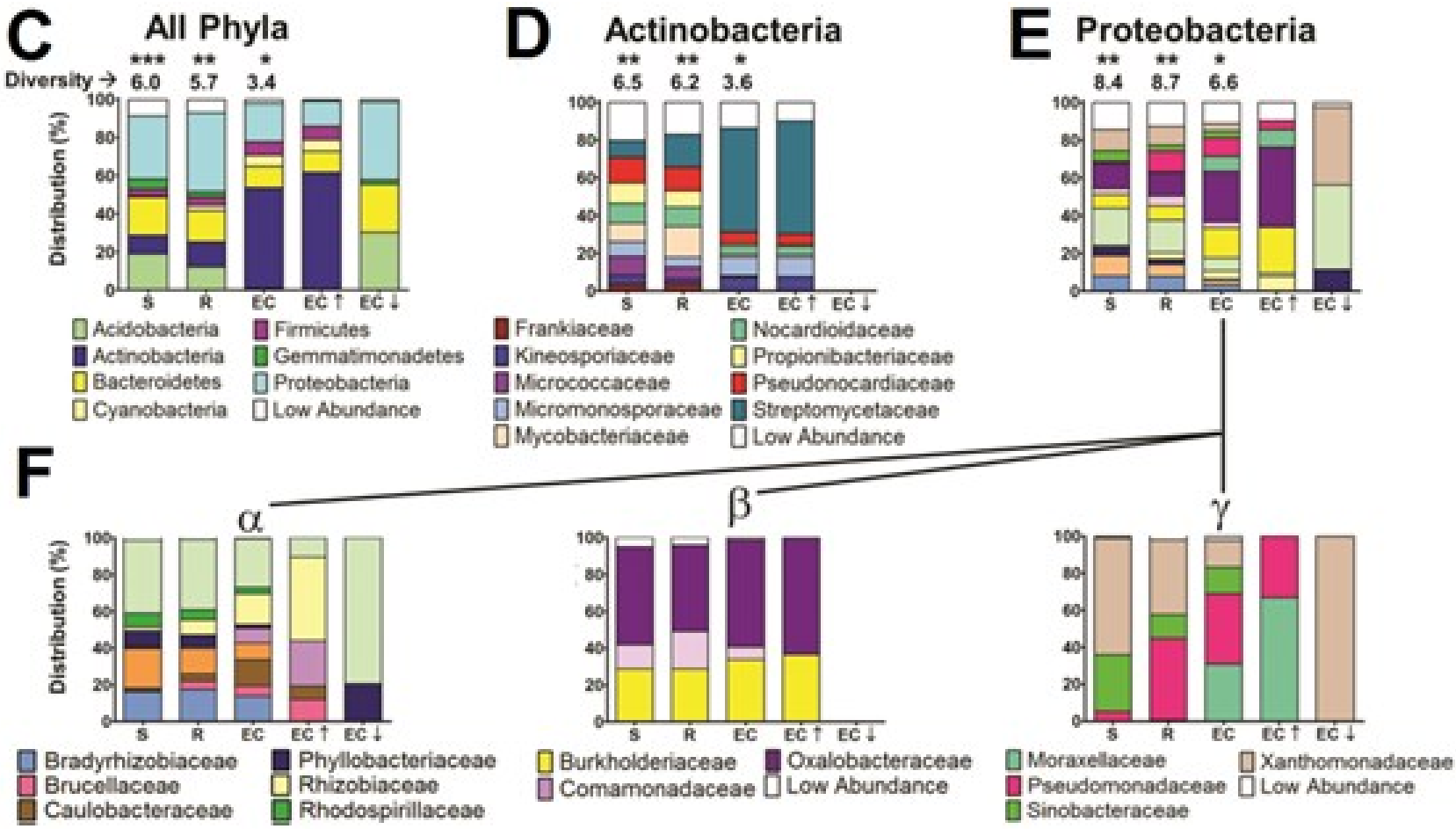

C
All Phyla
***
**
*
Diversity →
6.0
5.7
3.4
Distribution (%)
S
R
EC
EC ↑
EC ↓
Acidobacteria
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Gemmatimonadetes
Proteobacteria
Low Abundance
D
Actinobacteria
**
**
*
6.5
6.2
3.6
Frankiaceae
Kineosporiaceae
Micrococcaceae
Micromonosporaceae
Mycobacteriaceae
Nocardioidaceae
Propionibacteriaceae
Pseudonocardiaceae
Streptomycetaceae
Low Abundance
E
Proteobacteria
**
**
*
8.4
8.7
6.6
F
α
β
γ
Bradyrhizobiaceae
Brucellaceae
Caulobacteraceae
Phyllobacteriaceae
Rhizobiaceae
Rhodospirillaceae
Burkholderiaceae
Comamonadaceae
Oxalobacteraceae
Low Abundance
Moraxellaceae
Pseudomonadaceae
Sinobacteraceae
Xanthomonadaceae
Low Abundance

Figure 2.8. OTUs that differentiate the endophyte compartment and rhizosphere from soil (rarefied).

(A) Heat map showing OTU counts from the rarefied OTU table (log_2-transformed) from each of the 256 rhizosphere- and EC-differentiating OTUs present across replicates. Samples and OTUs are clustered on their Bray–Curtis similarities (group-average linkage). The key relates colours to the untransformed read counts. Different hues of the same colour correspond to different replicates as in Fig. 2.5.

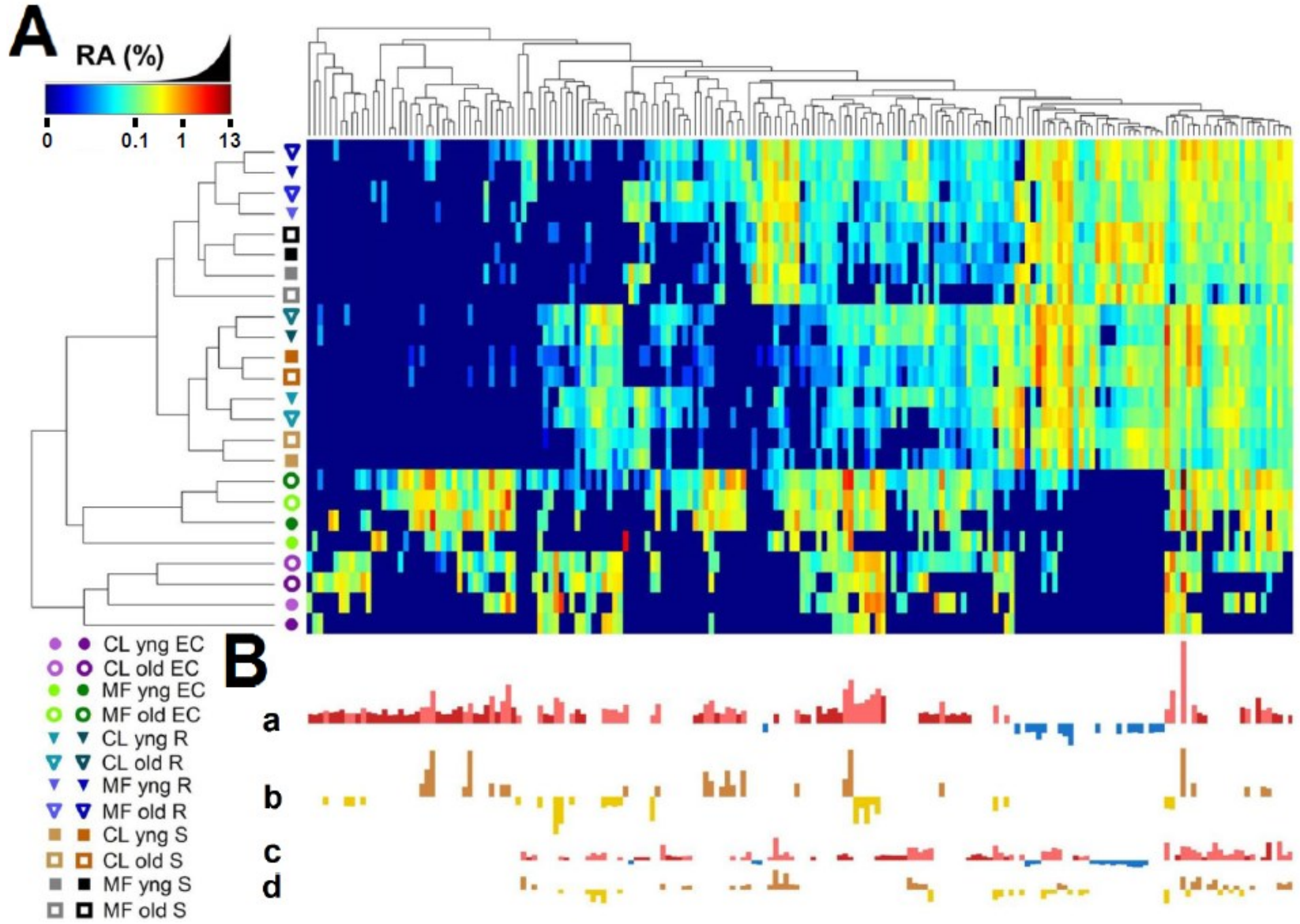
(B) The strength of GLMM predictions (best linear unbiased predictors) is represented by bar height. **a**, OTUs predicted as EC enriched (red, up) or EC depleted (blue, down). **b**, OTUs higher in the EC in Mason Farm soil than Clayton (brown, up) or higher in Clayton soil than Mason Farm (gold, down). OTUs in **a** that are not differentially affected by soil type are shown there in darker hues. **c**, OTUs predicted as rhizosphere enriched (as in **a**). **d**, OTUs higher in rhizosphere in one soil type (as in **b**).

(C) Histograms showing the distributions of phyla present in the 778 measurable OTUs in soil, rhizosphere and ECs compared with phyla present in the subset of EC OTUs enriched (EC ↑) or depleted (EC ↓) relative to soil. Shannon diversity (considering phyla as individuals) is given above each bar. A differential number of asterisks above the diversity values represents a significant difference ($P < 0.05$, weighted analysis of variance; Methods).

(D) Distribution of families present among the OTUs from the phylum Actinobacteria.

(E) Distribution of families present among the OTUs from the phylum Proteobacteria.

(F) Distribution of families present among the OTUs of three classes of the phylum Proteobacteria: Alphaproteobacteria (α), Betaproteobacteria (β) and Gammaproteobacteria (γ).


A
RA (%)
0
0.1
1
13
B
a
b
c
d
CL yng EC
CL old EC
MF yng EC
MF old EC
CL yng R
CL old R
MF yng R
MF old R
CL yng S
CL old S
MF yng S
MF old S

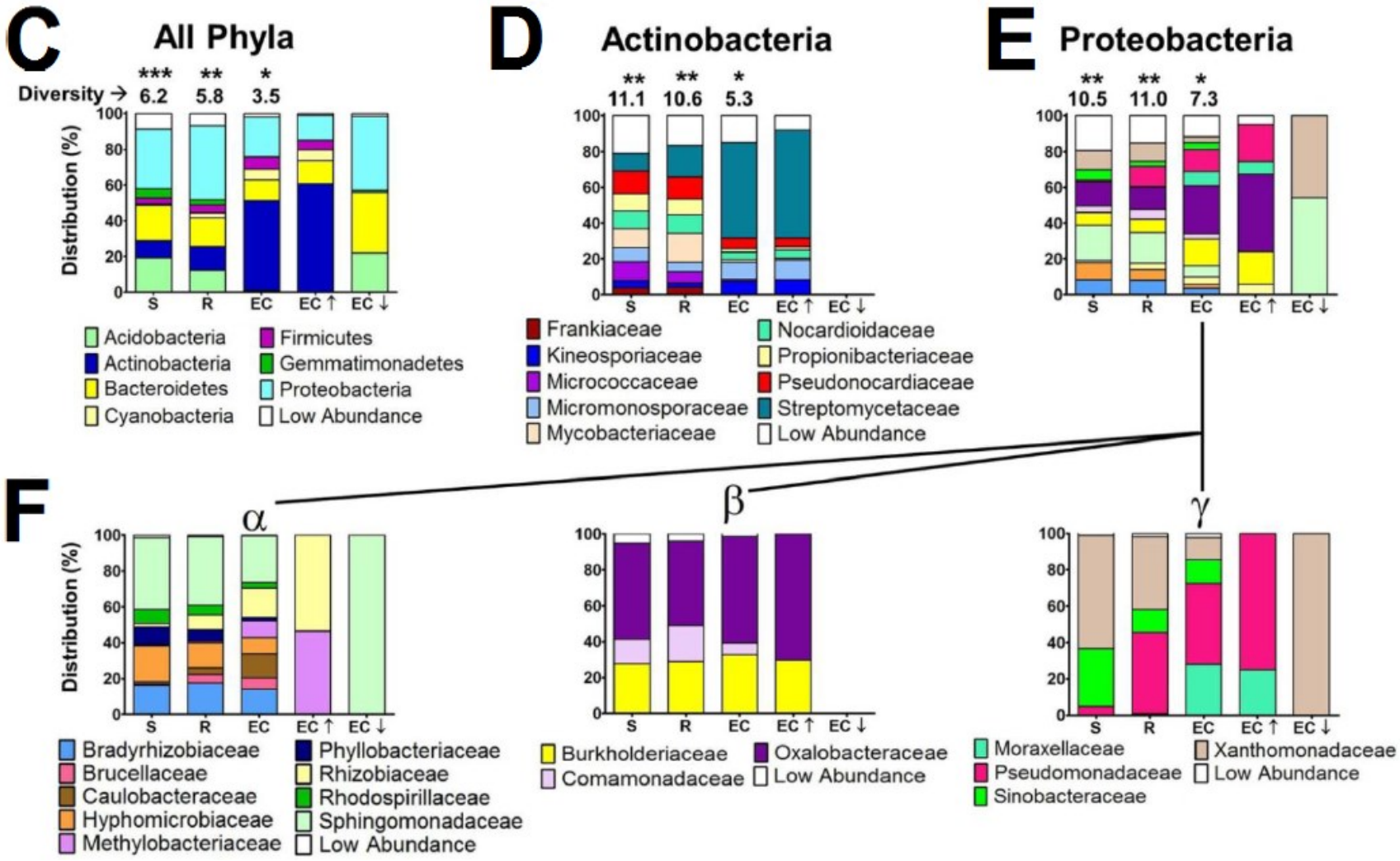

C
All Phyla
***
**
*
Diversity →
6.2
5.8
3.5
Distribution (%)
100
80
60
40
20
0
S
R
EC
EC ↑
EC ↓
Acidobacteria
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Gemmatimonadetes
Proteobacteria
Low Abundance
D
Actinobacteria
**
**
*
11.1
10.6
5.3
Frankiaceae
Kineosporiaceae
Micrococcaceae
Micromonosporaceae
Mycobacteriaceae
Nocardioidaceae
Propionibacteriaceae
Pseudonocardiaceae
Streptomycetaceae
Low Abundance
E
Proteobacteria
**
**
*
10.5
11.0
7.3
F
α
β
γ
Distribution (%)
Bradyrhizobiaceae
Brucellaceae
Caulobacteraceae
Hyphomicrobiaceae
Methylobacteriaceae
Phyllobacteriaceae
Rhizobiaceae
Rhodospirillaceae
Sphingomonadaceae
Low Abundance
Burkholderiaceae
Comamonadaceae
Oxalobacteraceae
Low Abundance
Moraxellaceae
Pseudomonadaceae
Sinobacteraceae
Xanthomonadaceae
Low Abundance

Figure 2.9. OTUs that differentiate the endophyte compartment and rhizosphere from soil (frequency).

(A) Heat map displaying the median RA (log_2 transformed) of each of 108 'R and EC-differentiating OTUs' present across experimental replicates, where samples and OTUs are clustered on their Bray Curtis Similarity (group average linkage). The color key relates the colors to the untransformed RAs.

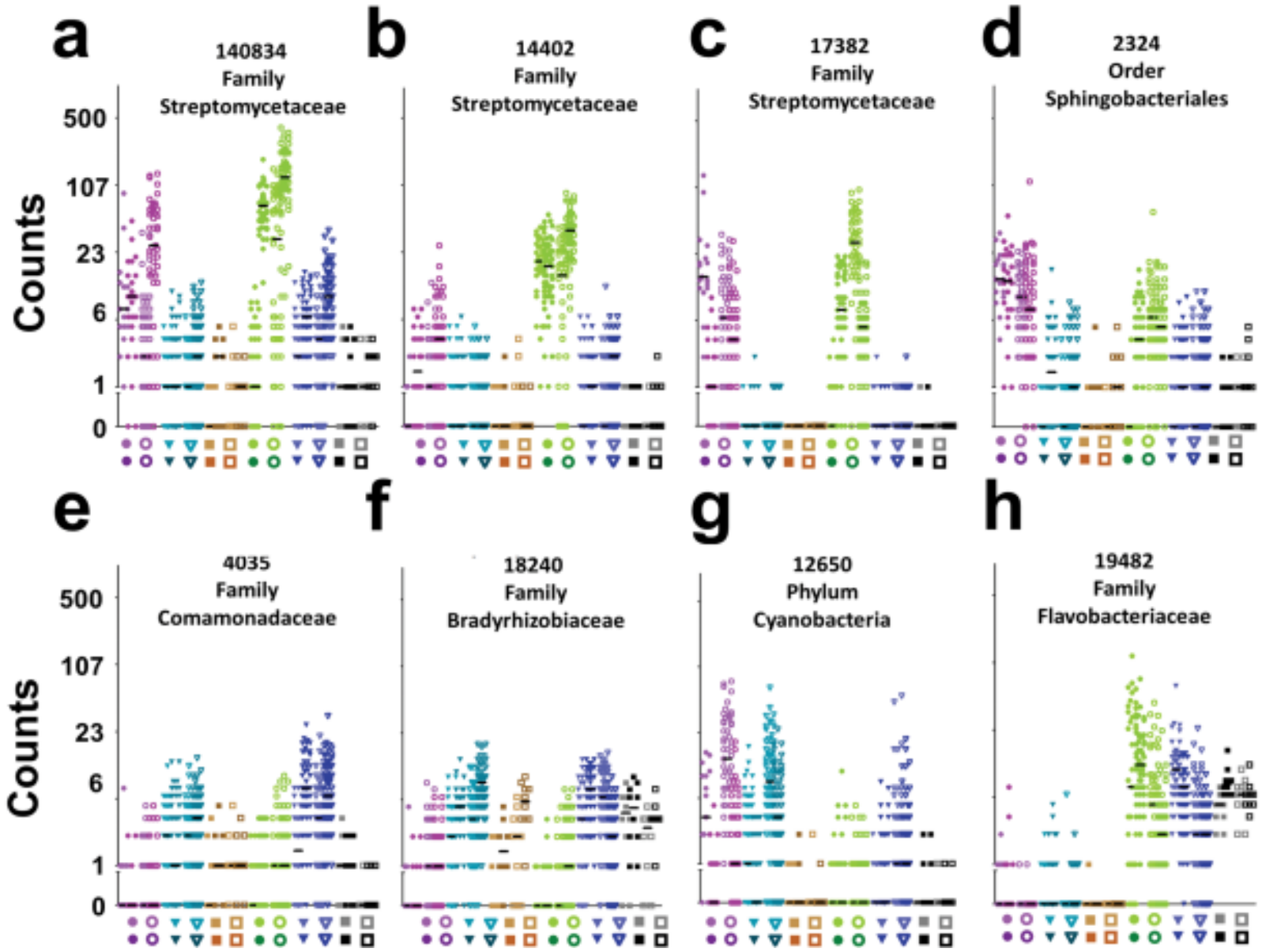
(B) The strength of the GLMM predictions (Best Linear Unbiased Predictors or BLUPs) is represented by the height of the bars. **a,** shows OTUs predicted as EC–enriched (red, up) or EC depleted (blue, down). **b,** shows OTUs found higher in the EC in MF soil than CL (brown, up) or higher in CL than MF (gold, down). OTUs in **i** that are not differentially affected by soil type as are shown in darker hues in **a**. **c,** OTUs predicted as R-enriched (as in **a** above). **d** OTUs higher in R in one soil type (as in **b**).

(C) Histogram displaying the distribution of the phyla present in the 778 measurable OTUs in soil (S), rhizosphere (R) and endophytic compartments (EC) compared to phyla present in the subset of EC OTUs enriched (EC-Up), or depleted (EC-Down) compared to soil. Shannon Diversity (considering phyla as individuals) is shown above. A differential number of asterisks above the Shannon Diversity values represents a significant difference ($p<0.05$, weighted ANOVA, Methods)

(D) Distribution of families present among the OTUs of the phylum Actinobacteria.

(E) Distribution of families present among the OTUs of the phylum Proteobacteria.

(F) Distribution of families present among the OTUs of three classes of the phylum Proteobacteria – Alpha (left), Beta (center), Gamma (right). Data in **(d-f)** are from both soil types, pooled (see Fig. 2.19 for each soil separately).


a
140834
Family
Streptomycetaceae
b
14402
Family
Streptomycetaceae
c
17382
Family
Streptomycetaceae
d
2324
Order
Sphingobacteriales
e
4035
Family
Comamonadaceae
f
18240
Family
Bradyrhizobiaceae
g
12650
Phylum
Cyanobacteria
h
19482
Family
Flavobacteriaceae
Counts
500
107
23
6
1
0

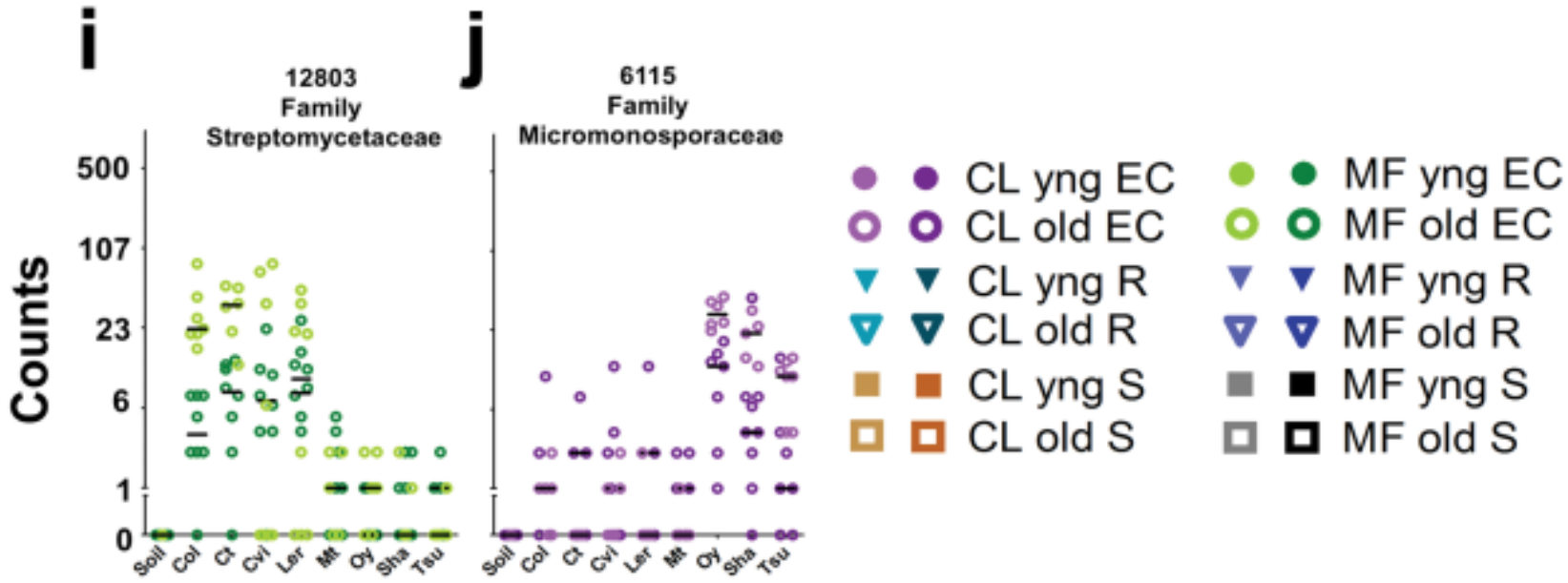
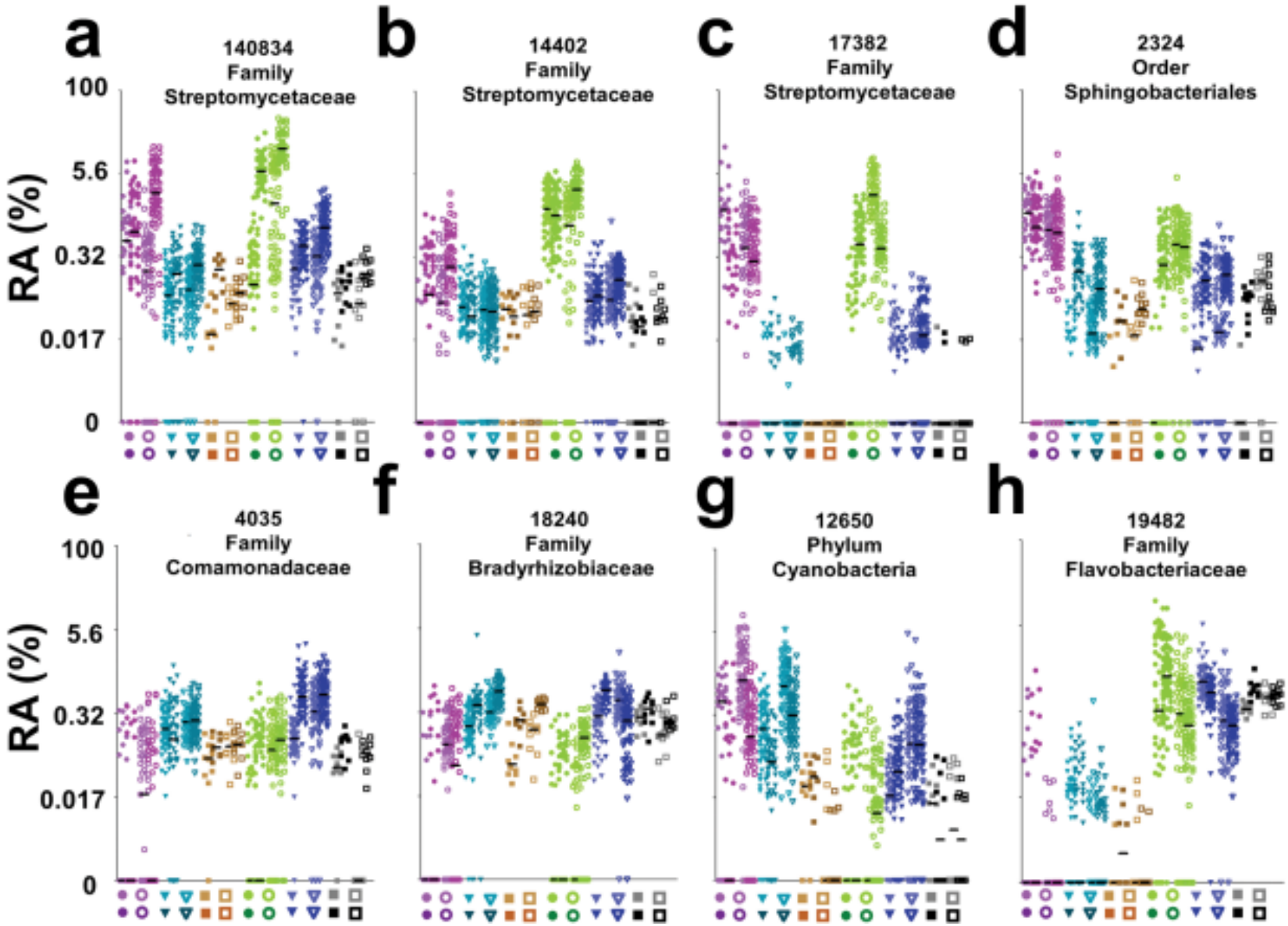


Figure 2.10. Dot Plots of Notable OTUs. (rarefied)

Counts for each OTU from the rarefied table were $\log_2$-transformed and the counts for each sample plotted as an individual symbol. The y axis is labelled with the actual (untransformed) counts. a–h, Each position on the x axis is labelled with a symbol to represent the sample group, and samples from that group are plotted in the column directly above. Biological replicates in the same column have different hues. The median of each replicate is shown with a horizontal black bar; some are invisible because they are at 0. i, j, Each x-axis position is labelled by Arabidopsis accession; samples from that accession are plotted above each label. Each OTU in the figure has model predictions in several categories.


a
140834
Family
Streptomycetaceae
b
14402
Family
Streptomycetaceae
c
17382
Family
Streptomycetaceae
d
2324
Order
Sphingobacteriales
e
4035
Family
Comamonadaceae
f
18240
Family
Bradyrhizobiaceae
g
12650
Phylum
Cyanobacteria
h
19482
Family
Flavobacteriaceae
RA (%)
100
5.6
0.32
0.017
0

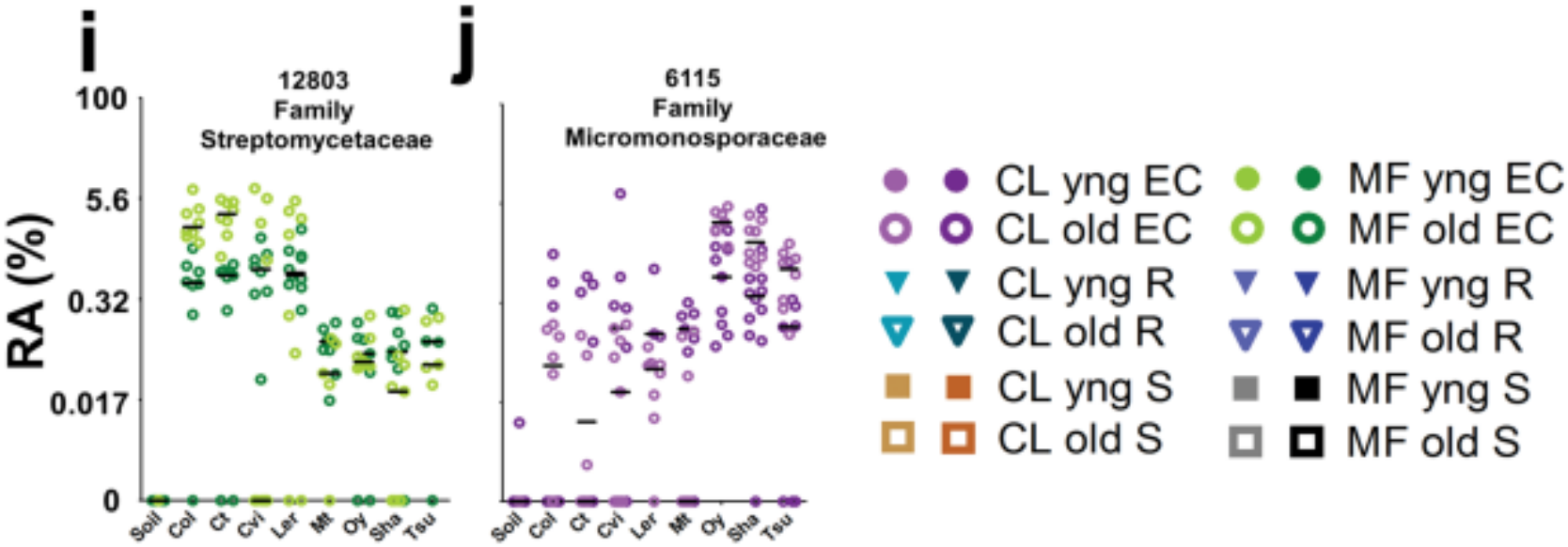


Figure 2.11. Dot Plots of Notable OTUs (frequency)

Relative abundance for each OTU (number at top of each panel) from the frequency-normalized table was $\log_2$ transformed and the abundance for each sample (y-axis) plotted as an individual symbol. The y-axis is labeled with the actual (untransformed) relative abundance values. In **a-h**, each position on the x-axis is labeled with a symbol to represent the sample group (legend, lower right), and samples from that group are plotted column-wise directly above. Biological replicates are shown in the same column with different hues. The median of each biological replicate is shown with a horizontal black bar; some may not be visible because they are at 0. In **i** and **j**, sample color is according to the legend, and each position on the x-axis is labeled by Arabidopsis accession, with samples from that accession plotted above each label. Each OTU in the figure has model predictions in several categories.

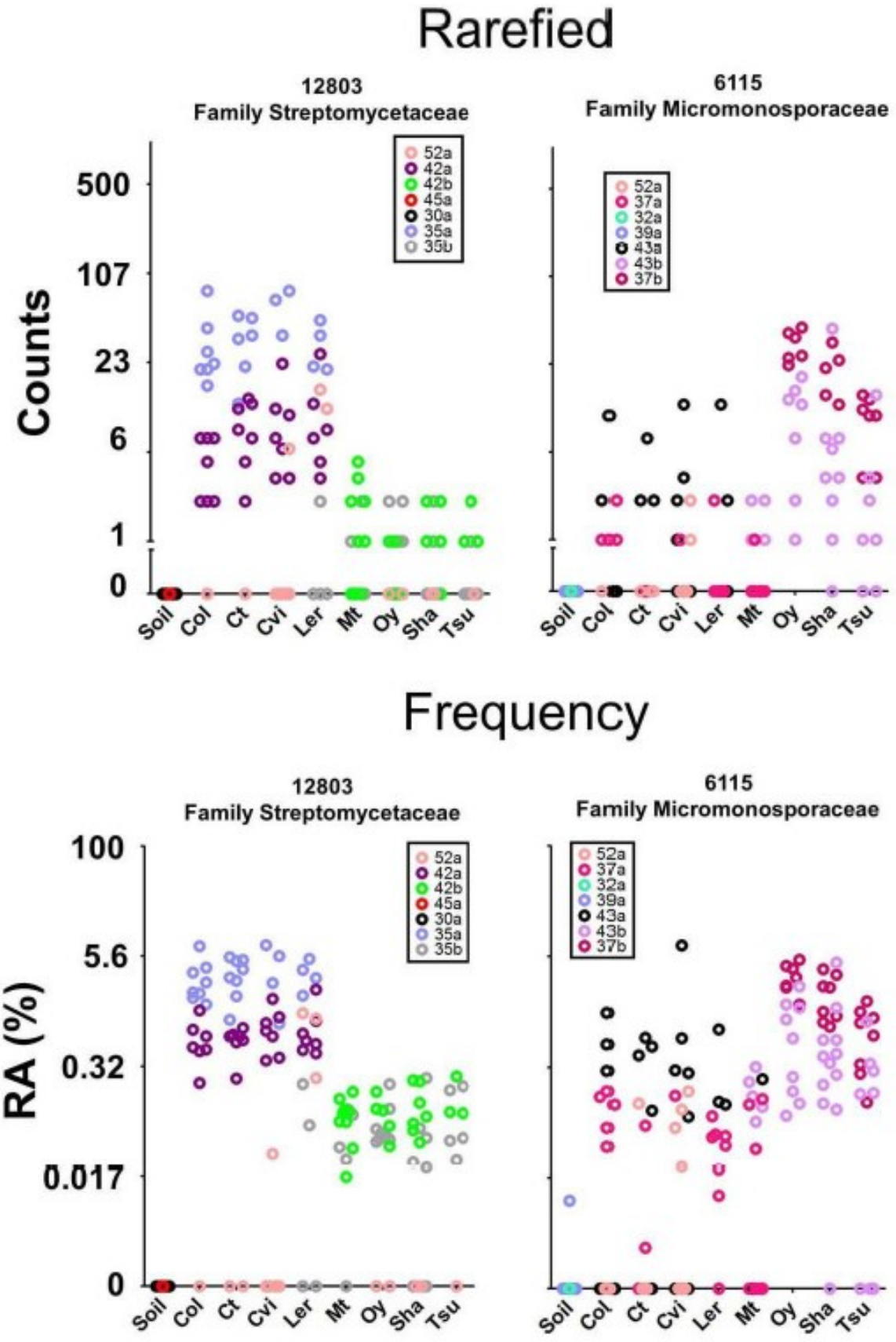

Rarefied
12803
Family Streptomycetaceae
6115
Family Micromonosporaceae
Counts
500
107
23
6
1
0
52a
42a
42b
45a
30a
35a
35b
52a
37a
32a
39a
43a
43b
37b
Soil
Col
Ct
Cvi
Ler
Mt
Oy
Sha
Tsu
Frequency
12803
Family Streptomycetaceae
6115
Family Micromonosporaceae
RA (%)
100
5.6
0.32
0.017
0
52a
42a
42b
45a
30a
35a
35b
52a
37a
32a
39a
43a
43b
37b
Soil
Col
Ct
Cvi
Ler
Mt
Oy
Sha
Tsu

Figure 2.12. Genotype-variable OTUs colored by sequence plate.

Displays the data from **Fig. 2.10i** (MF old EC, left) and **Fig. 2.10j** (CL old EC right), colored by sequence plate (instead of biological replicate as in Fig. 2.10) according to the legend within each plot. The top panel is based on rarefied data, as in Figure 2.10, and the bottom panel is based on the relative abundance, as in Fig. 2.11. (Note: 'a' and 'b' in our plate naming scheme do not represent different regions of the same plate. All 454 regions were modeled independently in the Full GLMM).

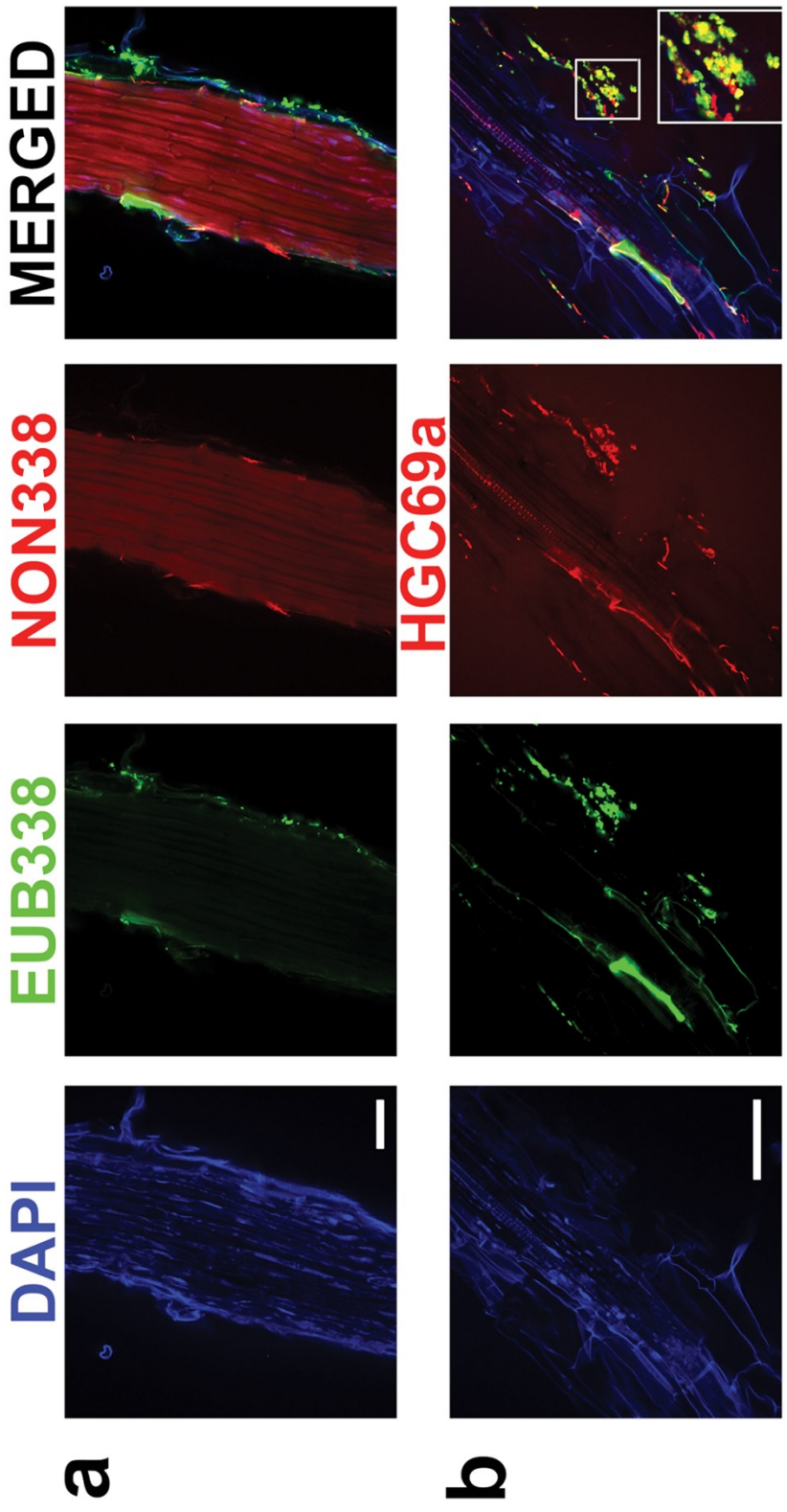

DAPI
EUB338
NON338
HGC69a
MERGED
a
b

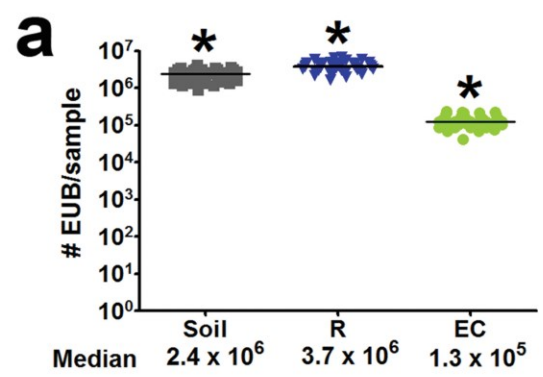
Figure 2.13. CARD–FISH confirmation of Actinobacteria on roots.

A single set of Mason Farm yng Col-0 roots were fixed and stained using CARD–FISH. DAPI, 49,6-diamidino-2-phenylindole. Double CARD–FISH was applied using the EUB338 eubacterial probe (green) and either

(a) theNON338 probe, which is the nonsense negative control of EUB338, or

(b) the HGC69a Actinobacteria probe.

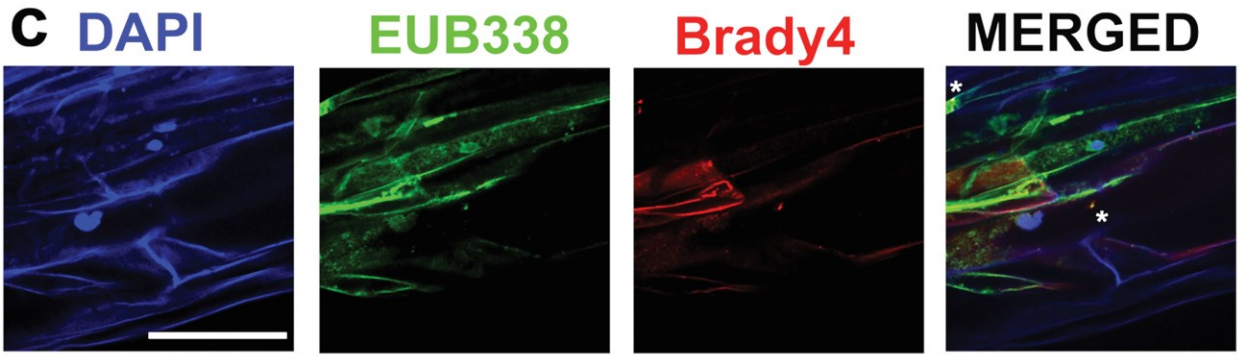
Inset, twofold enlargement of boxed region. Scale bars, 50 mm.



b

	% HGC69a	% Brady4
Soil	14.23 ± 2.46	3.05 ± 1.35
R	9.65 ± 2.12	2.65 ± 1.41
EC	20.38 ± 3.80	0.56 ± 0.56

Mean ± SEM



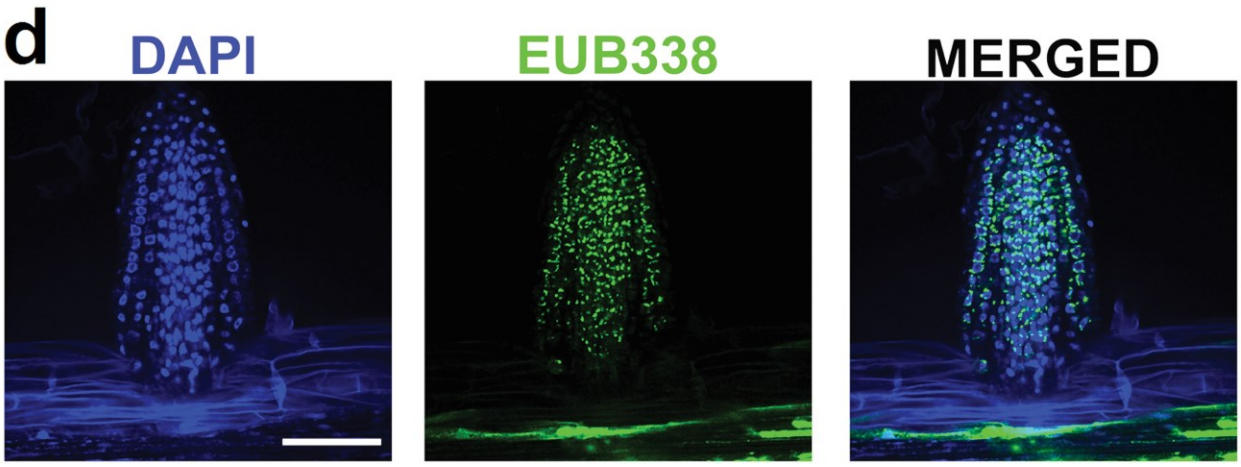


Figure 2.14. Quantification of microbes in the three sample fractions using CARD-FISH.

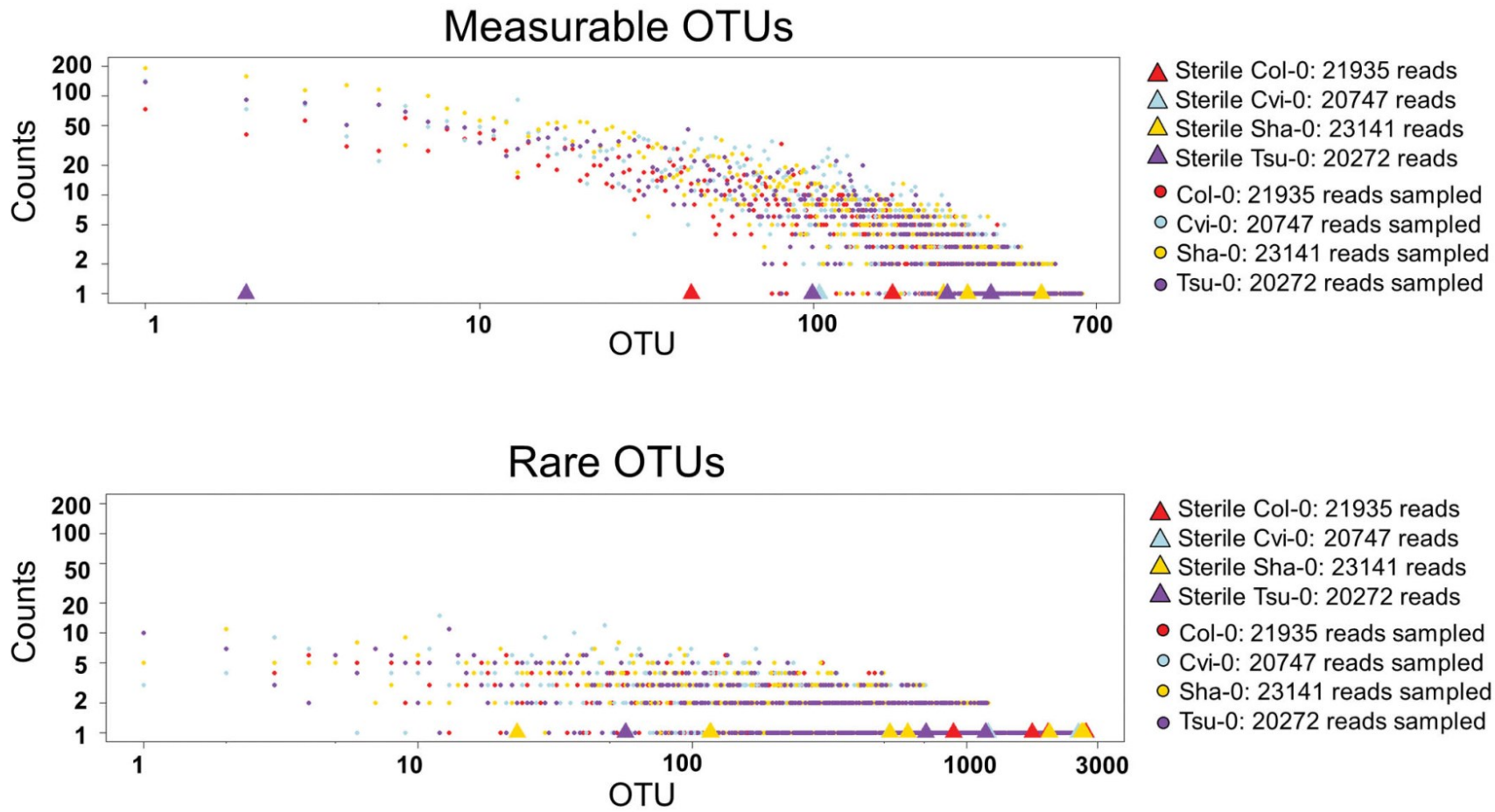
Four sets of Col-0 roots were pooled, processed, diluted, and put onto filters.

(a) CARD-FISH using the EUB338, eubacterial probe, was applied and counterstained with DAPI. The number of EUB positive signals co-localizing with a DAPI signal was counted and the number of EUB positive signals per sample was calculated. This is an estimate for the number of bacteria present in each of our samples that DNA was extracted from with bulk soil (n=40), rhizosphere (n=39), and endophytic compartment (n=40). * indicates statistical significance at p<1x1016 (ANOVA with post-hoc TukeyHSD) between each of the sample groups

(b) Using double CARD-FISH on filters made from equal concentration of the 3 sample fractions, we determined the % of DAPI positive eubacteria that are also co-localize with either the HGC69a (Actinobacteria) or Brady4 (Bradyrhizobiaceae) probes on filters made from bulk soil (n=10), rhizosphere (n=10), and endophytic compartment (n=10) samples. Actinobacteria was in higher abundance in EC samples and Bradyrhizobiaceae was in lower abundance in EC samples compared to soil and R samples as expected from our pyrotag sequencing data.

(c) Double CARD-FISH was applied using the EUB338, eubacterial probe (green) and the Brady4, Bradyrhizobiaceae probe (red), counterstained with DAPI (the asterisks indicate signals that are positive in all 3 channels).

(d) Newly forming lateral roots and root tips were found commonly to be heavily colonized. Scale bars represent 50 microns.



OTUs detected in >1 sterile sample

Rare OTU_9990: Unclassified (1 read each in sterile Cvi-0, Sha-0, and Tsu-0)

Rare OTU_17330: Phylum Cyanobacteria (1 read each in sterile Cvi-0, and Tsu-0)

Figure 2.15. Pyrosequencing of sterile seedlings as compared to non-sterile EC samples.

DNA was extracted from homogenates from gnotobiotic seedlings of the genotypes Col-0, Cvi-0, Sha-0, and Tsu-0 (from which no culturable microbes were found), using bacteriolytic DNA preps, and these were pyrosequenced and clustered into OTUs as part of our full dataset. 21935, 20747, 23141, and 20272 high quality reads were obtained from each gnotobiotic genotype, respectively (triangles). The same total number of total reads was sampled from using pooled EC data from the full dataset for these accessions (circles). Each position on the X axis represents an OTU in the full dataset (measurable OTUs on top, rare OTUs on bottom) and the position on the Y axis represents the number of sequence reads found in that OTU. Both axes are shown in log scale. Of the 86095 HQ reads obtained from both sterile plants and non-sterile plants, the majority were from chloroplast OTUs (not shown). Far more non-plant reads were obtained from the non-sterile plants (19093 of 86095, or 22%) vs. sterile plants (34 of 86095, or 0.04%), a difference approaching three orders of magnitude. The 34 reads from non-sterile plants were members of 31 OTUs (triangles – some overlap on the log-scale axis). No OTU in a sterile plant sample was represented by more than one read, and only two OTUs were shared by more than one of the accessions - both of these shared OTUs were not in the measurable set, and had poor taxonomic classification. 11 of these 31 OTUs were not represented in the non-sterile samples. Furthermore, by including extra unused barcodes in our mapping files, or by sequencing sterile water in excess, we have been able to occasionally 'detect' single representatives of OTUs in our dataset, demonstrating that technical noise can cause singletons (data not shown). While we cannot rule out that unculturable microbes survive surface sterilization and exist at extremely low abundance, we have no evidence that such microbes exist in *A. thaliana* roots.

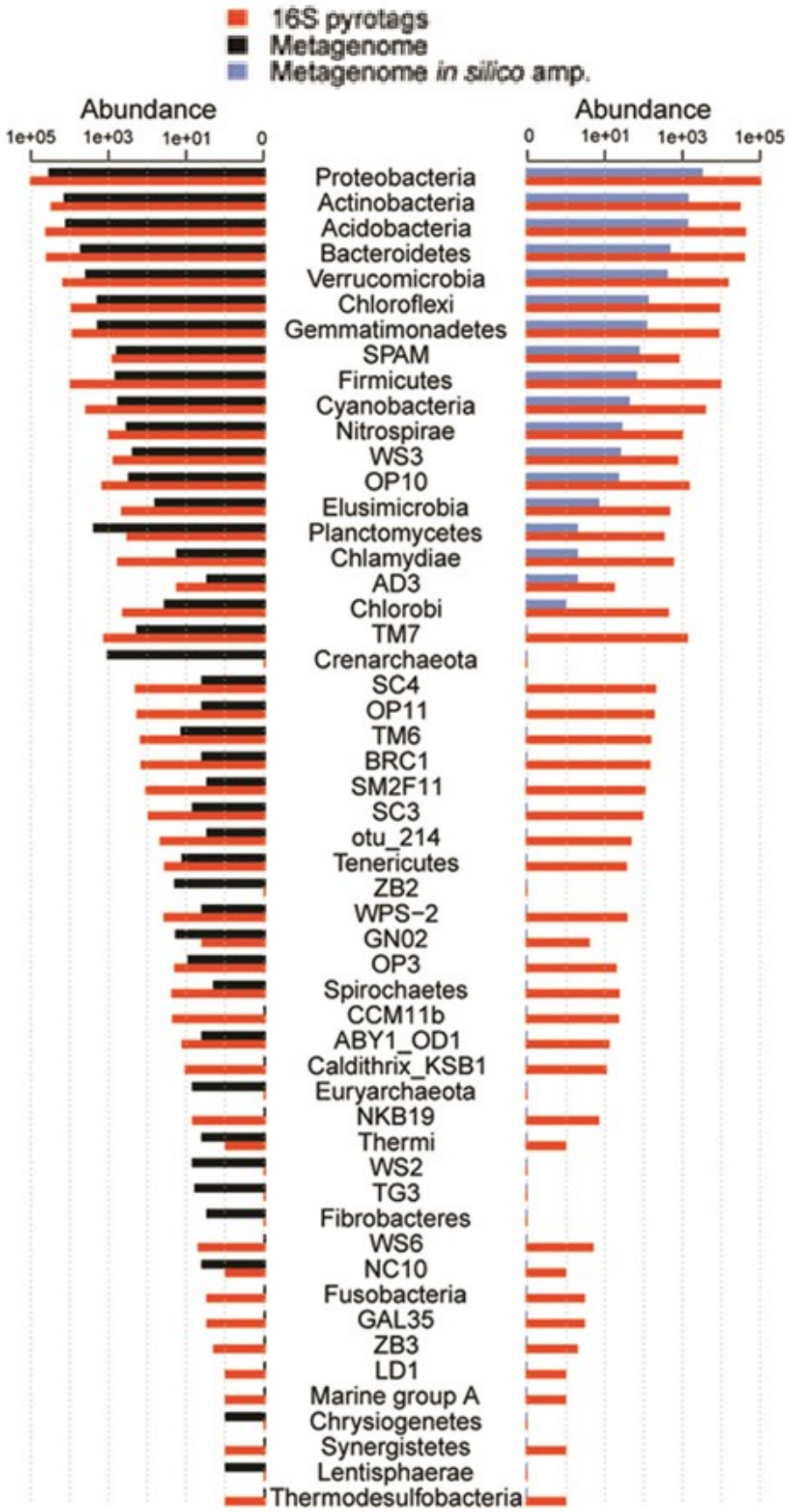

16S pyrotags
Metagenome
Metagenome *in silico* amp.
Abundance
1e+05
1e+03
1e+01
0
Abundance
0
1e+01
1e+03
1e+05
Proteobacteria
Actinobacteria
Acidobacteria
Bacteroidetes
Verrucomicrobia
Chloroflexi
Gemmatimonadetes
SPAM
Firmicutes
Cyanobacteria
Nitrospirae
WS3
OP10
Elusimicrobia
Planctomycetes
Chlamydiae
AD3
Chlorobi
TM7
Crenarchaeota
SC4
OP11
TM6
BRC1
SM2F11
SC3
otu_214
Tenericutes
ZB2
WPS−2
GN02
OP3
Spirochaetes
CCM11b
ABY1_OD1
Caldithrix_KSB1
Euryarchaeota
NKB19
Thermi
WS2
TG3
Fibrobacteres
WS6
NC10
Fusobacteria
GAL35
ZB3
LD1
Marine group A
Chrysiogenetes
Synergistetes
Lentisphaerae
Thermodesulfobacteria

Figure 2.16. Test for PCR bias in pyrotagging.

Relative abundance of 16S metagenomics and pyrotag reads

(a) To assess possible bias introduced by amplification for pyrotagging, we compared the taxonomic distribution of a metagenome library created without amplification with a corresponding pyrotag dataset. Both datasets are from Col-0 Mason Farm young samples. 16S rDNA reads from this metagenome library (One HiSeq lane; more than 400 million 150 bp paired-end reads) were extracted by alignment against the 16S Silva database (release 106). Aligned reads were then assigned a taxonomy using an RDP training set built with the Greengenes reference database (version: May 9th 2011). This allowed classification of 57,663 16S reads from the metagenome sample using a bootstrap threshold >=0.50. There is an excellent overall correlation between the relative abundance of pyrotags and metagenome 16S rDNA reads across the major phyla represented in the datasets. Only two major classes, Thaumarchaeota and Planctomycea, were not amplified by the 1114F-1392R primers. Slightly higher abundance of Actinobacteria and Betaproteobacteria was observed in pyrotag data than in metagenome 16S reads. This was investigated further.

(b) For those classes in which underrepresentation in the pyrotag data are observed (red class names in **a**, we used *in silico* PCR analyses using the Greengenes database as template and our pyrotags primer pair, allowing a maximum of 2 mismatches, to investigate at which taxonomic level the under-representation would be discerned . We show that Thaumarchaeota (class) and Planctomycea (class) may be misrepresented in our pyrotag data. Since the Greengenes database contains many sequences amplified with the 1392R primer and therefore lacks this primer's sequence, we removed all sequences shorter than 6449 (in absolute position) in our reference database to minimize false negative rate (*i.e.* sequences not amplifying because they are not long enough to match the 1392R primer sequence).

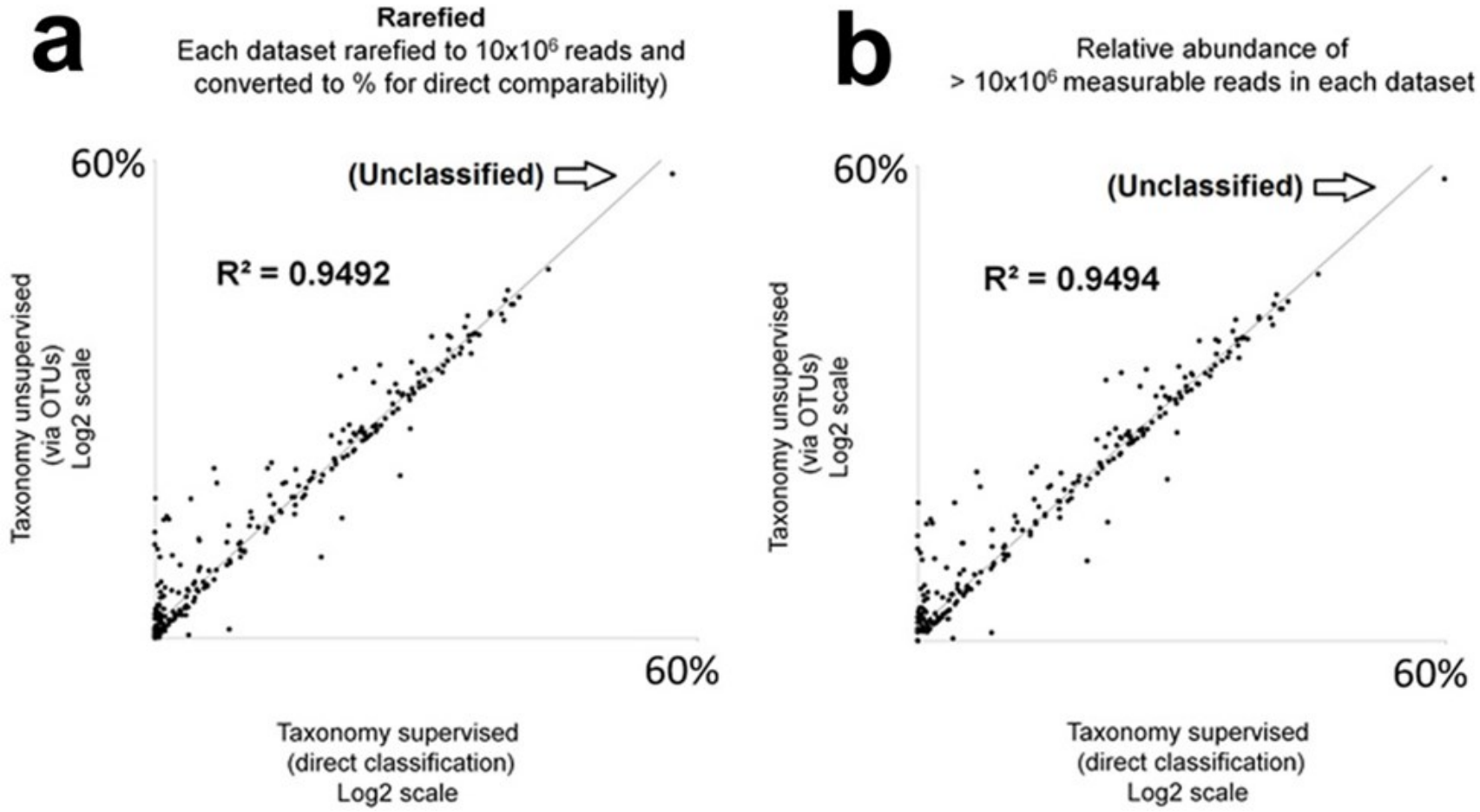

a
Rarefied
Each dataset rarefied to 10x10^6 reads and converted to % for direct comparability)
60%
(Unclassified)
R² = 0.9492
Taxonomy unsupervised (via OTUs) Log2 scale
60%
Taxonomy supervised (direct classification) Log2 scale
b
Relative abundance of > 10x10^6 measurable reads in each dataset
60%
(Unclassified)
R² = 0.9494
Taxonomy unsupervised (via OTUs) Log2 scale
60%
Taxonomy supervised (direct classification) Log2 scale

Figure 2.17. 16S taxonomy classification at the family level is robust to method.

For taxonomy-supervised classification, reads that passed default QIIME quality thresholds (but that were not clustered into OTUs) were trimmed to 220bp and were classified via RDP against Greengenes (Feb. 4 2011 version) training set to get family-level taxonomy. The abundance of each family was compared to the abundance of that family when the family assignments were assigned *after* the taxonomy-unsupervised grouping of reads into OTUs.

(a) The total reads from non-chloroplast families from both taxonomy-supervised and taxonomy-unsupervised methods were rarefied to 10,000,000 reads, and the reads per family are shown as the $\log_2$ transformed relative abundance of the total reads, whereas

(b) The relative abundance of each family using all non-chloroplast reads, omitting the rarefaction step. The scatterplots thus show the high correlation at the family level for supervised and unsupervised taxonomy assignment. The dataset used for this figure included extra samples not described here, and was clustered as a single .fasta using the default QIIME implementation of Uclust (Caporaso et al. 2010).

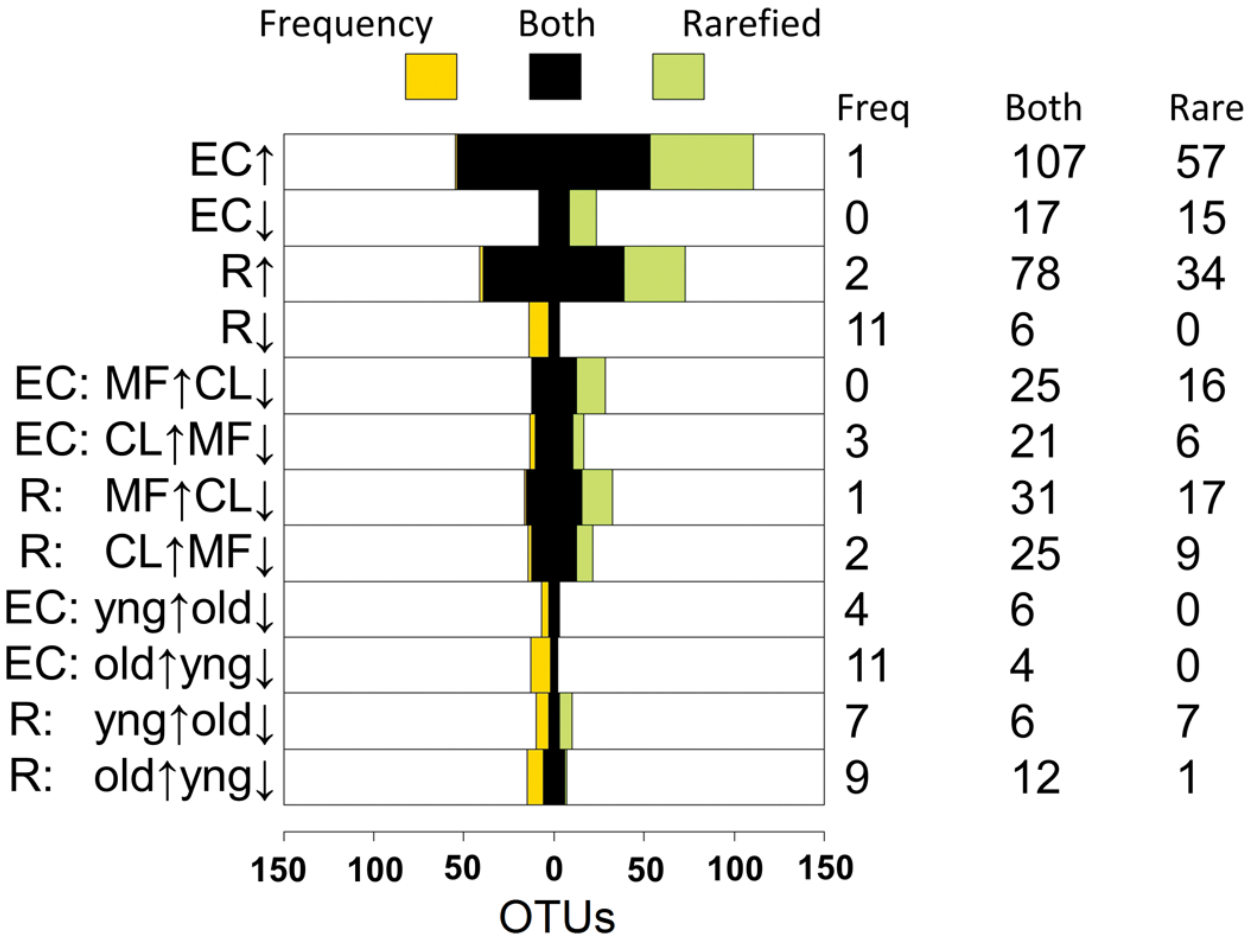


Figure 2.18. Overlap of GLMM predictions between rarefaction-normalized and frequency-normalized OTU tables.

The number of OTUs predicted by the full GLMM in each category that are unique to the frequency table is shown in orange. The number of OTUs predicted by the full GLMM in each category that are unique to the rarefied table are shown in green. The number of OTUs that were shared predictions in the two tables is shown in black.

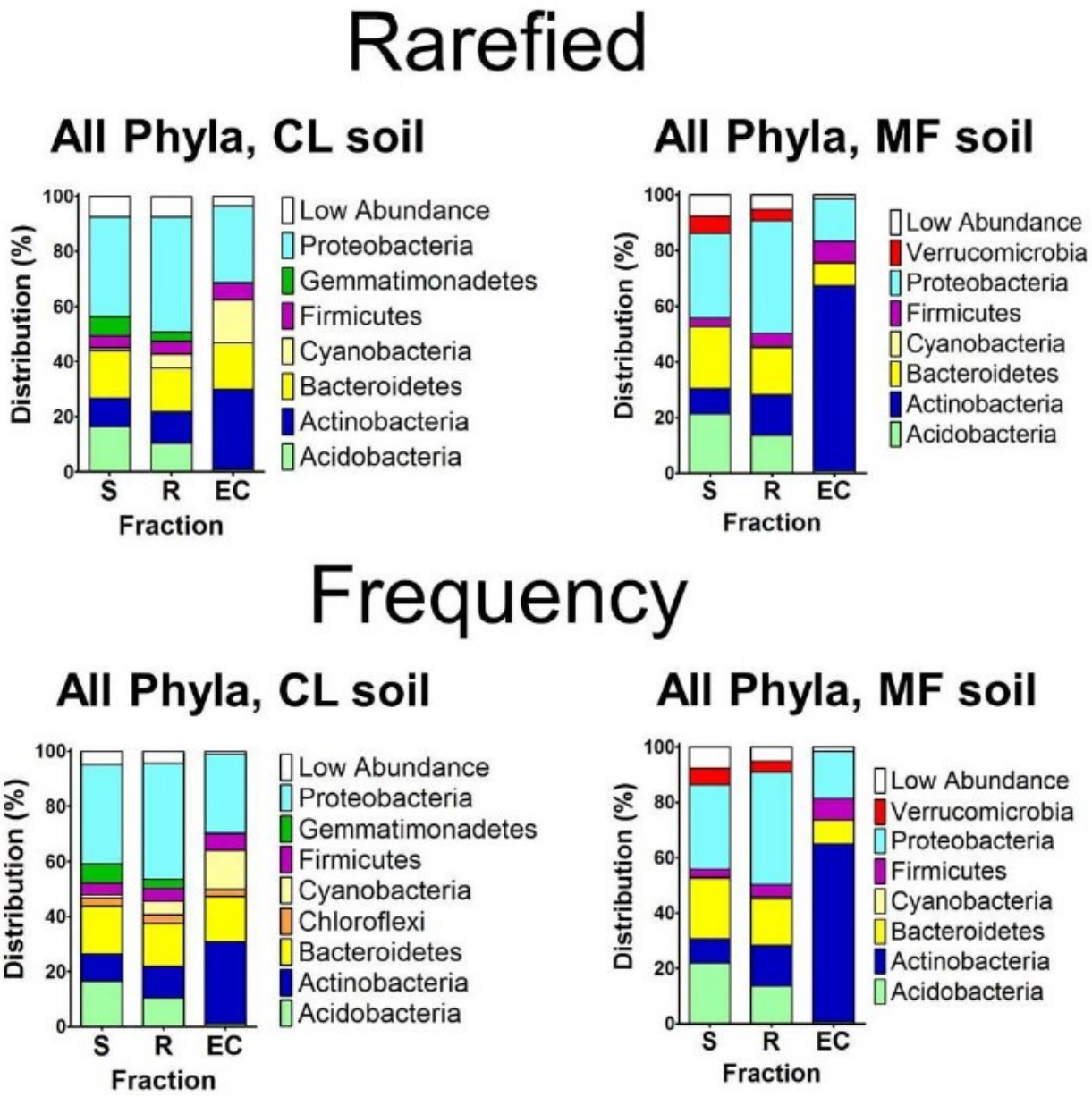


Figure 2.19. Phyla in each sample fraction by soil type.

Histogram displaying the distribution of the phyla present in the 778 measurable OTUs in soil (S), rhizosphere (R) and endophytic compartments (EC) with each soil type, MF and CL, considered independently. Rarefaction-normalized on top; frequency-normalized on bottom.

REFERENCES

Arumugam M, Raes J, Pelletier E, Le Paslier D, Yamada T et al. (2011) Enterotypes of the human gut microbiome. Nature 473(7346): 174-180.

Bates D, Maechler M, Bolker B (2011) lme4: Linear mixed-effects models using s4 classes (R package version 0.999375-42). Available: http://CRAN.R-project.org/package=lme4.

Benson AK, Kelly SA, Legge R, Ma F, Low SJ et al. (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of Sciences 107(44): 18933-18938.

Bulgarelli D, Rott M, Schlaeppi K, Ver Loren van Themaat E, Ahmadinejad N et al. (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature 488(7409): 91-95.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Meth 7(5): 335-336.

Chi F, Shen SH, Cheng HP, Jing YX, Yanni YG et al. (2005) Ascending migration of endophytic rhizobia, from roots to leaves, inside rice plants and assessment of benefits to rice growth physiology. Appl Environ Microbiol 71(11): 7271-7278.

DeDeyn GB, Cornelissen JHC, Bardgett RD (2008) Plant functional traits and soil carbon sequestration in contrasting biomes. Ecology Letters 11(5): 516-531.

Dennis PG, Miller AJ, Hirsch PR (2010) Are root exudates more important than other sources of rhizodeposits in structuring rhizosphere bacterial communities? FEMS Microbiology Ecology 72(3): 313-327.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL et al. (2006) Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol 72(7): 5069-5072.

Dodds PN, Rathjen JP (2011) Plant immunity: towards an integrated view of plant-pathogen interactions. Nat Rev Genet 11(8): 539-548.

Eickhorst T, Tippkötter R (2008) Improved detection of soil microorganisms using fluorescence in situ hybridization (FISH) and catalyzed reporter deposition (CARD-FISH). Soil Biology and Biochemistry 40(7): 1883-1891.

Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F et al. (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J 4(5): 642-647.

Firáková S, Šturdíková M, Múčková M (2007) Bioactive secondary metabolites produced by microorganisms associated with plants. Biologia 62(3): 251-257.

Gottel NR, Castro HF, Kerley M, Yang Z, Pelletier DA et al. (2011) Distinct microbial communities within the endosphere and rhizosphere of Populus deltoides roots across contrasting soil types. Appl Environ Microbiol 77(17): 5934-5944.

Hallmann J, Quadt-Hallmann A, Mahaffee WF, Kloepper JW (1997) Bacterial endophytes in agricultural crops. Canadian journal of microbiology 43(10): 895-914.

Hardoim PR, van Overbeek LS, Elsas JD (2008) Properties of bacterial endophytes and their proposed role in plant growth. Trends Microbiol 16(10): 463-471.

Inceoglu O, Al-Soud WA, Salles JF, Semenov AV, van Elsas JD (2011) Comparative analysis of bacterial communities in a potato field as determined by pyrosequencing. PLoS One 6(8): e23321.

Inceoglu O, Salles JF, van Overbeek L, van Elsas JD (2010) Effects of plant genotype and growth stage on the betaproteobacterial communities associated with different potato cultivars in two fields. Appl Environ Microbiol 76(11): 3675-3684.

Jones JDG, Dangl JL (2006) The plant immune system. Nature 444(7117): 323-329.

Kunin V, Hugenholtz P (2010) PyroTagger: A fast, accurate pipeline for analysis of rRNA amplicon pyrosequence data. The Open Journal.

Lane DJ (1991) 16S/23S rRNA sequencing. In: E. Stackebrandt MG, editor. Nucleic acid techniques in bacterial systematics: John Wiley & Sons, Chichester, United Kingdom

Loy A, Maixner F, Wagner M, Horn M (2007) probeBase--an online resource for rRNA-targeted oligonucleotide probes: new features 2007. Nucleic Acids Res 35(Database issue): D800-804.

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

Marschner H, Römheld V, Horst WJ, Martin P (1986) Root-induced changes in the rhizosphere: Importance for the mineral nutrition of plants. Zeitschrift für Pflanzenernährung und Bodenkunde 149(4): 441-456.

Masclaux C, Valadier M-H, Brugière N, Morot-Gaudry J-F, Hirel B (2000) Characterization of the sink/source transition in tobacco (Nicotiana tabacum

L.) shoots in relation to nitrogen management and leaf senescence. Planta 211(4): 510-518.

Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M et al. (2011) Deciphering the Rhizosphere Microbiome for Disease-Suppressive Bacteria. Science 332(6033): 1097-1100.

Motulsky HJ (2003) Prism 4 Statistics Guide −Statistical analyses for laboratory and clinical researchers. . San Diego, CA.: GraphPad Software, Inc.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR et al. (2011) vegan: Community Ecology Package.

Redford AJ, Bowers RM, Knight R, Linhart Y, Fierer N (2010) The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. Environmental Microbiology 12(11): 2885-2893.

Rodriguez R, Redman R (2008) More than 400 million years of evolution and some plants still can't make it on their own: plant stress tolerance via fungal symbiosis. J Exp Bot: 1109-1114.

S. Levey AW (2005) Natural variation in the regulation of leaf senescence and relation to other traits in Arabidopsis. Plant, Cell & Environment 28(2): 223-231.

Schulz BJE, Boyle CJC, Sieber TN, Schulz B, Boyle C (2006) What are Endophytes? Microbial Root Endophytes: Springer Berlin Heidelberg. pp. 1-13.

Spor A, Koren O, Ley R (2011) Unravelling the effects of the environment and host genotype on the gut microbiome. Nature reviews Microbiology 9(4): 279-290.

Sul WJ, Cole JR, Jesus Eda C, Wang Q, Farris RJ et al. (2011) Bacterial community comparisons by taxonomy-supervised analysis independent of sequence alignment and clustering. Proc Natl Acad Sci U S A 108(35): 14637-14642.

Van der Lelie D, Taghavia S, Monchya S, Schwendera J, Millerb L et al. (2009) Poplar and its Bacterial Endophytes: Coexistence and Harmony. Critical Reviews in Plant Sciences 28: 346-358.

van Elsas JD, Trevors JT, Starodub ME (1988) Bacterial conjugation between pseudomonads in the rhizosphere of wheat. FEMS microbiology letters 53(5): 299-306.

Warnes GR (2011) gplots: Various R programming tools for plotting data.

CHAPTER 3

Practical innovations for high-throughput amplicon sequencing[1]

INTRODUCTION

We describe improvements for sequencing 16S ribosomal RNA (rRNA) amplicons, a cornerstone technique in metagenomics. Through unique tagging of template molecules before PCR, amplicon sequences can be mapped to their original templates to correct amplification bias and sequencing error with software we provide. PCR clamps block amplification of contaminating sequences from a eukaryotic host, thereby substantially enriching microbial sequences without introducing bias.

MAIN

Microbes profoundly affect biological processes across Earth's ecological niches and are frequently identified through culture-independent methods using DNA purified directly from environmental samples (Lozupone and Knight 2007). Common PCR-based approaches target highly conserved rRNA genes, such as those encoding the 16S/18S and 28S subunits or the internal transcribed spacer (ITS) between them. These ubiquitous

---

[1]Lundberg DS*, Yourstone S*, Mieczkowski P, Jones CD, Dangl JL (2013) Practical innovations for high-throughput amplicon sequencing. Nat Meth 10(10): 999-1002.
* = contributed equally

genes have diverged enough that polymorphisms across their 'hypervariable regions' (Figure 3.1) allow taxonomic classification. Amplicon sequencing is an important and widely used tool for inferring the presence of taxonomic groups in microbial communities, but poor estimates result from sequencing errors and biases introduced during amplification. Inefficiencies also result from the amplification of nontarget DNA. Here we describe methods that make rRNA amplicon sequencing more accurate and cost-effective.

Accurate base-calling on Illumina platforms requires sequence diversity at each nucleotide position (Krueger et al. 2011). Because amplicon libraries often lack diversity at specific positions owing to sequence conservation, it is common to spike sequencing runs with sheared genomic DNA from the virus phiX174. We created sequence diversity in 16S amplicons using a mix of primers that have frameshifting nucleotides (Figures 3.2 and 3.3). Despite recent upgrades to Illumina's base-calling procedure, this strategy remains useful for maximizing data yield as it devotes the entire sequencing effort to the amplicon of interest (Figures 3.4 and 3.5).

PCR and sequencing introduce sequence errors and sampling bias (Patin et al. 2013). We adapted and validated a modified protocol that uniquely tags each template molecule with random nucleotides before PCR (Faith et al. 2013; Jabara et al. 2011; Kinde et al. 2011; Kivioja et al. 2011) (Figures 3.2 and 3.6a,b). Provided that there are enough random nucleotides, amplicons sharing the same tag are overwhelmingly likely to have originated from the same template molecule (the 'birthday paradox' (Sheward et al. 2012); Figure 3.7). Thus, by generating consensus sequences from each group of sequences sharing a molecule tag (MT), we can correct errors and infer the amplicon's probable template sequence (Figure 3.6f–h).

We verified that consensus sequences (ConSeqs) correct errors by amplifying a

clonal plasmid-borne 16S template (Figure 3.8). A dilution series ensured a variety of coverage depths for each MT (Figure 3.8a). We found that a sample of 15,000 ConSeqs had fivefold lower mean error than a sample of 15,000 untreated (nonconsensus) 16S sequences (Figure 3.8b and Materials and Methods).

We observed unexpectedly high numbers of singletons in the MT depth distributions for samples prepared from diluted templates, suggesting that some singletons arise from MT mutations in lower-quality reads. Consistent with this, the error rate among 15,000 singletons was more than twice that for untreated sequences. We also observed a lower error rate among the 4,777 available 'perfect ConSeqs' constructed from three or more reads with identical sequence sharing an MT, compared with all ConSeqs. Interestingly, this rate was not 0 because either all sequences in these perfect ConSeqs carried the same error, the template plasmid had some level of polymorphism that was accurately captured or a combination of these.

Operational taxonomic unit (OTU) clustering is a common approach both to corral noisy 16S sequence data into groups approximating microbial species and to reduce computational complexity (Patin et al. 2013). Using data from the clonal 16S template, we clustered either 30,000 untreated sequences or 30,000 ConSeqs into OTUs using both 97% and 99% identity thresholds. ConSeqs clustered at 97% formed two OTUs, with the second OTU containing only six sequences (Figure 3.8c). Untreated sequences at 97%, on the other hand, produced 66 OTUs, two of which were sufficient to capture 95% of the data. ConSeqs clustered at 99% formed 42 OTUs, and the first two OTUs contained 95% of the data, whereas untreated sequences produced 683 OTUs and required 66 OTUs to capture 95% of the data. Thus, ConSeqs were more homogenous than untreated sequences and tolerated stricter OTU definitions, a result suggesting that ConSeqs can be used to provide a more accurate picture of true microbial alpha diversity(Patin et al. 2013).

We applied our approach to samples amplified from pooled bulk wild Mason Farm soil DNA ('soil') and pooled root endophyte compartment DNA grown in that soil (Lundberg et al. 2012) ('root EC'; Materials and Methods). All 16S reads were processed into untreated sequences, ConSeqs and singletons as above, as well as a mix of 'ConSeqs plus adjusted singletons' (CASs), in which the singletons were downsampled in proportion to the ConSeqs collapse ratio (the ratio of the number of all ConSeqs to the number of all constituent sequences used to compute them). CASs thus retain the majority of singletons from template-overloaded samples, in which singletons contain the majority of high-quality reads; but they retain fewer singletons from dilute samples, in which the singletons are enriched for lower-quality outcasts. We generated OTUs at 97% and 99% identity thresholds and used rarefaction curves to observe the microbial richness (Figure 3.9a). Within both root EC and the more complex soil communities, ConSeqs and CASs performed similarly and gave estimates of microbial richness lower than those of untreated sequences. This effect was particularly apparent at 99% clustering, but it was also evident at 97%, again demonstrating that ConSeqs correct overestimates of microbial alpha diversity (Patin et al. 2013).

MT treatments enhanced the technical reproducibility of independently amplified samples. Our data set comprised 12 pairs of root EC replicates and 12 pairs of soil replicates (Materials and Methods). The OTU abundances of all samples were regressed against those of their replicates, and the coefficient of determination R^2 was graphed (Figure 3.9b). Low-abundance OTUs were the least correlated (Benson et al. 2010; Bulgarelli et al. 2012; Lundberg et al. 2012); as these were removed, R^2 increased quickly. Even before low-abundance OTUs were dropped, ConSeqs and CASs were more reproducible than untreated sequences and singletons, and their R^2 plateaued more quickly. Singletons formed many more small OTUs than did other classes, especially at 99% clustering. Thus, relatively more of the irreproducible singleton data were discarded at lower OTU abundance

thresholds than for other MT classes, which explains the more rapid increase in technical reproducibility for singletons than for untreated sequences.

We compared our method directly to that of the Earth Microbiome Project (EMP), which uses primers without MTs (Caporaso et al. 2012). Using both methods, we prepared libraries of the same sample composition, including independent soil samples from two sites, root EC samples from individual plants grown in one of the soils and the clonal 16S template used above (Materials and Methods). Major beta diversity conclusions from both methods were the same; the sample types grouped similarly after we performed principal-coordinates analysis based on weighted UniFrac distances (Figure 3.10). Also, the same clades formed on the basis of hierarchical clustering by Bray-Curtis dissimilarity. However, there were fewer OTUs using our method, which is consistent with our initial data (Figures 3.8a and 3.9a). Evidence that the extra OTUs are noise comes from the clonal 16S template, which formed one OTU with our method, as opposed to several with the EMP method.

Next we tackled a problem encountered when investigating microbial communities associated with a eukaryotic host, wherein 16S sequences originating from the host's genome, plastid or mitochondria can account for >80% of the sequences obtained (Bulgarelli et al. 2012; Lundberg et al. 2012; Sakai and Ikenaga 2013). Although modification of the bases in the 'universal' amplicon primers can mitigate amplification of the contamination, this can also lead to bias (Sim et al. 2012). We instead developed peptide nucleic acid (PNA) PCR clamps (von Wintzingerode et al. 2000): synthetic oligomers that bind tightly and specifically to a unique signature in the contaminant sequence and physically block its amplification (Ray and Nordén 2000; Sakai and Ikenaga 2013; Tanaka et al. 2010; Troedsson et al. 2008) (Figure 3.11 and Materials and Methods). We designed PNAs to suppress plant host plastid and mitochondrial 16S contamination (Figure 3.12) and tested them using 24 samples amplified from pooled root EC DNA samples, in which ~85%

of 16S sequences post-PCR were either plastid or mitochondria (Figure 3.13a). Combining both PNAs in the same reaction blocked both types of contaminant and yielded approximately eightfold more bacterial 16S rRNA sequence as a fraction of total sequences.

Owing to an effective PNA-dependent template reduction, the mean number of sequences sharing an MT that were aligned and used to calculate each ConSeq was ~2.5-fold larger in the 12 samples containing anti-plastid PNA (pPNA; $P = 0.026$, permutation test of the means) (Figure 3.13a). Neither the presence of pPNA or anti-mitochondrial PNA (mPNA) nor the related increase in the number of sequences per alignment affected clustering of root EC samples by bacterial families or OTUs (Figure 3.13b and Figure 3.14a). There was also not a significant effect on the relative abundance of individual bacterial families or OTUs when the 12 samples amplified with each PNA were compared to the 12 samples amplified without it ($Q > 0.05$ for all permutation tests on the means with false discovery rate (FDR) correction; Figures 3.14a and 3.15a and Materials and Methods). Using the same PNA concentrations for PCR of extremely diverse bulk soil (Lundberg et al. 2012), we observed that PNAs had no effect on clustering of samples by bacterial families or OTUs (Figure 3.13c and Figure 3.14b) or the abundances of families or OTUs ($Q > 0.05$ for all permutation tests on the means with FDR correction; Figures 3.14b and 3.15b), with one exception that was likely a false positive.

Both the pPNA and mPNA sequences are conserved among higher plants and should function well for most plant microbiome projects (Figure 3.16). Many studies have demonstrated the potential of PNAs for a variety of research questions using low-resolution molecular methods (Chow et al. 2011; Ray and Nordén 2000; Sakai and Ikenaga 2013; Terahara et al. 2011; Troedsson et al. 2008; von Wintzingerode et al. 2000), but a proof-of-concept study using deep sequencing has been lacking. A recent study showed the effectiveness of PNAs designed to block plastid and mitochondrial sequences for plant

microbiome analysis using T-RFLP (Sakai and Ikenaga 2013). However, the authors considered only primer annealing-blocking regions that overlapped with conserved 16S primers, which limited the number of candidate PNAs and likely their target specificity.

Sequence features in the molecule tagging–frameshifting (MT-FS) primers can be used as additional barcodes. For example, nonintersecting sets of frameshifting primers on two samples sharing the same PCR barcode—or better, conventional barcoding bases in the MT-FS primers—allowed samples to be distinguished with >99.9% accuracy. Each MT-FS barcode, or even unrelated template-tagging primers such as ITS region primers, can be used with the universal PCR barcodes, thereby enhancing the cost-effectiveness of our approach (Figures 3.17 and 3.18).

We also provide our validated MTToolbox: user-friendly software to merge overlapping paired-end reads, recognize and trim primer sequences, and process molecular tags into ConSeqs. MTToolbox is compatible with data produced by the related Safe-SeqS (Kinde et al. 2011) and LEA-Seq (Faith et al. 2013) techniques. Downloads and source code can be accessed through SourceForge (http://sourceforge.net/projects/mttoolbox/), and user manuals and documentation can be found at https://sites.google.com/site/moleculetagtoolbox/).

In summary, our methods provided higher sequencing accuracy and technical reproducibility while increasing flexibility and savings. In the case of a MiSeq run of 96 root EC samples in which the PNAs were applicable, the combination of frameshifts, combinatorial barcoding and PNA yielded substantial cost reductions and provided greater flexibility to investigate new amplicons. These techniques can be adopted à la carte for a particular amplicon project and sequencing platform. The benefits of frameshifting and template tagging were independently described in a metagenomics context during the revision of this work (Faith et al. 2013), attesting to the need for improved amplicon sequencing methods.

MATERIALS AND METHODS

Cloned 16S template

We amplified a 16S rRNA gene from a Mycobacterium sp. using primers 27F and 1492R and 25 PCR cycles, cloned the PCR product into pENTR/D-TOPO (Invitrogen) and selected a single transformed Escherichia coli colony. Plasmid DNA was prepped from a 3 mL culture using standard alkaline lysis, purified by silica column, quantified using a NanoDrop 1000 (Thermo Scientific) and sequenced using an ABI3130 genetic analyzer using 515F and 806R variable region 4 (V4) primers. The forward and reverse reads were overlapped and merged using Sequencher (http://genecodes.com/). Primer sequences were recognized and removed, thereby generating a high-quality sequence (Figure Specific Details).

Root EC, soil and leaf DNA extraction and quantification

Mason Farm root endophyte compartment DNA (root EC), Mason Farm bulk soil DNA (soil), and Clayton bulk soil DNA (Clayton soil) were collected and extracted as previously described in Lundberg et al. 2012. All Arabidopsis DNA was made from the *Arabidopsis thaliana* Col-0 reference accession. *A. thaliana* and *Oryza sativa* leaf DNA were prepared in the same manner as root EC DNA, except that a similar quantity of whole leaves was prepped fresh, without sonication, bleaching or any other treatment to remove epiphytes. DNA templates were quantified using PicoGreen fluorescent dye (Invitrogen) and a fluorescence plate reader exciting at 475 nm and reading at 530 nm. Leaf DNA could not be reliably quantified, as it showed fluorescence at the limits of detection, and was therefore added without dilution in the template-tagging reactions (described below). For the individual samples used in the comparison of our method (Run C) to the Earth Microbiome Project (EMP) method (Run D), approximately 50 ng/μL was used for each sample.

Peptide nucleic acid (PNA) design

To identify candidate PNA oligo sequences, we fragmented in silico the full length *A. thaliana* plastid and mitochondrial 16S sequences into short k-mers for k of length 9, 10, 11,

12 and 13, and we queried for exact matches against the 4 February 2011 version of the Greengenes 16S training set comprising 35,430 unique, high-quality full-length bacterial sequences (Figure 3.12). *A. thaliana*–specific k-mers falling between the 515F and 806R 16S rRNA primers (V4 region) were considered candidates and were lengthened as necessary to increase the predicted melting temperatures and were screened for design characteristics (Terahara et al. 2011; von Wintzingerode et al. 2000).

A successful elongation arrest PNA clamp is generally between 13 bp and 17 bp and has an annealing temperature above that of the PCR primer whose extension it blocks and a melting temperature above that used for the extension cycle (Terahara et al. 2011). We designed 17-mer sequences to block the plastid and mitochondria, each with a predicted melting temperature around 80 °C (Table 3.1f). Melting temperature, problematic hairpins, GC content and other design considerations were calculated using the Life Technologies PNA designer (http://www6.appliedbiosystems.com/support/pnadesigner.cfm).

The anti-mitochondrial PNA (mPNA) 5′-GGCAAGTGTTCTTCGGA-3′ and the anti-plastid PNA (pPNA) 5′-GGCTCAACCCTGGACAG-3′ (Table 3.1f) were ordered from PNA Bio. Lyophilized PNA was resuspended in sterile water to a stock concentration of 100 μM. For PNA concentrations that were repeatedly tested, working stocks of 5 μM, 15 μM, 25 μM and 40 μM were prepared in water. All stocks were stored at −20 °C and heated to 65 °C before use to resolubilize any precipitate.

Primer design

All primers longer than 45 bases were Ultramers from Integrated DNA Technologies, purified by standard desalting. Shorter primers, such as the sequencing primers, were ordered from Eurofins MWG Operon and purified by the QuickLC method. Forward and reverse molecule tagging–frameshifting (MT-FS or Bc-MT-FS for nonbarcoded and barcoded, respectively) V4 16S primers and universal barcoding PCR primers are diagrammed and listed in Figure 3.2.

MT-FS primers and their barcoded versions, Bc-MT-FS primers, were designed with the frameshift and barcoding bases occurring within the molecular tag regions to break up the stretch of random bases and minimize unpredictable features related to annealing and secondary structure. We used 2-bp linkers to buffer the template-annealing 515F and 806R

portions of the MT-FS primers from the rest of the primer. Ideal linkers have low homology to known microbial sequences, creating a short stretch of mispairing. Our linker sequences for the V4 16S region differ from those used in the EMP method (Caporaso et al. 2012) but are equally valid choices on the basis of the lack of matches to the Greengenes database (Figure 3.18).

The molecule-tagging ITS2 primers are similar but are of an earlier design that uses nine random bases for the forward primer and four random bases for the reverse primer. No frameshifting variants of the ITS2 primers were used. Forward and reverse molecule-tagging ITS2 primers are listed below:

>ITS9F
GCCTCCCTCGCGCCATCAGAGATGTGTATAAGAGACAGNNNNNNNNNTTGAACGCAGC
RAAIIGYGA

>ITS4R
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTNNNNGATCCTCCGCTTATTGATATG
C

The 9-bp barcodes we used for the universal barcoding PCR primers were adapted from the 12-bp Golay barcodes used by Caporaso and colleagues (Caporaso et al. 2012). Of the 2,168 published Golay barcodes, we chose a subset of 96 that had a balanced mix of all bases at each position. We then extracted just the first 9 bases of these 12-bp barcodes; in our set of 96 barcodes of 9 bp, three or more SNPs would be needed to transform any one barcode into another. We chose to trim the Golay barcodes from 12 to 9 in order to shorten the primers; deeper barcoding can be accomplished by adding mini-barcodes in the MT-FS primers, such as the 3-bp barcodes we chose (Figure 3.2b), and combining each mini-barcode used during template tagging with the full suite of 96 universal barcodes in PCR.

Template tagging with molecular tagging–frameshifting primers

Template DNA was tagged with the MT-FS primers in two reactions: one for the reverse MT-FS primers and a subsequent reaction for the forward MT-FS or Bc-MT-FS primers, as described below. The purpose of using the tagging primers in two separate reactions, one for each primer, was to reduce the possibility of formation of difficult-to-

remove heterodimers between the long MT-FS primers. The shorter reverse MT-FS primers were used to tag the template first because removal of shorter primers during PCR cleanup is more efficient. Although the use of separate tagging reactions discourages heterodimers, it is not strictly necessary; and in practice both forward- and reverse-tagging primers can be used in a single two-cycle template-tagging reaction with good results (not shown).

For reverse V4 16S tagging in Run B, the primary MiSeq run we analyzed, we prepared two working stocks of reverse MT-FS V4 16S primer in water, where each working stock contained an equimolar mix of three of our six primers such that the concentration of the mixed stock was 0.5 μM. These working stocks we designate “V4R_2-4-6” (806R_f2, 806R_f4, and 806R_f6) and “V4R_1-3-5” (806R_f1, 806R_f3, and 806R_f5). For Run C, which we used to compare our method directly to the EMP method, we used a mix of all six reverse MT-FS primers (“V4R_mix1-6”) such that the concentration of the mixed stock was again 0.5 μM.

We used the KAPA 2G Robust HS PCR Kit with dNTPs (KK5518, Kapa Biosystems) in a 25 μL including 5 μL Kapa Enhancer, 5 μL Kapa Buffer A, 2 μL of 0.5 μM reverse-tagging primer mix (“V4R_1-3-5” or “V4R_2-4-6” for Run B or “V4R_mix1-6” for Run C), 0.5 μL Kapa dNTPs, 0.25 μL Kapa Robust Taq and 12.5 μL DNA template with water.

To minimize pipetting variation of small volumes, we used master mixes to prepare reagents whenever possible. Samples were incubated in a thermocycler using a program of denaturing at 95 °C for 1 min, reverse–MT-FS primer annealing at 50 °C for 2 min, and extension at 72 °C for 2 min, followed by a cooldown to 4 °C. The newly synthesized reverse-tagged strands, as well as the original DNA template molecules to which they were annealed, were cleaned to remove primers and PCR reagents with Agencourt AMPure XP beads (Beckman Coulter) using the manufacturer's protocol with the exception of an altered bead-to-DNA ratio: we used 15 μL of beads to clean the 25 μL of tagged template because this ratio (0.6:1) allowed size selection that more effectively eliminated the long tagging primers (data not shown). The DNA was eluted in 11 μL of water.

The cleaned, reverse-tagged DNA was next tagged with forward primers. For Run B, we made two forward MT-FS working stocks of three frameshift variants each (Figure 3.2b), which we designate “V4F_2-4-6” (515F_f2, 515F_f4, and 515F_f6) and “V4F_1-3-5” (515F_f1, 515F_f3, and 515F_f5). For Run C, we made two forward Bc-MT-FS working

stocks of six frameshift variants each, where each Bc-MT-FS mix differed by its 3-bp barcode (Figure 3.2b). We designate these “V4F_TGA_mix1-6” and “V4F_ACT_mix1-6.”

For samples to which PNA was applied, PNA was included in reactions in only the forward-tagging step, as the PNA blocks the extension of the forward-tagging primers. The 25-µL forward-tagging reaction included 5 µL Kapa Enhancer, 5 µL Kapa Buffer A, 2 µL of 0.5 µM forward-tagging primer mix (“V4F_1-3-5” or “V4F_2-4-6” for Run B or “V4F_TGA_mix1-6” or “V4F_ACT_mix1-6” for Run C), 0.5 µL Kapa dNTPs, 0.25 µL Kapa Robust Taq, 2.5 µL PNA working stock (containing pPNA, mPNA, both mPNA and pPNA, or water) and 10 µL reverse-tagged DNA from above.

Samples were incubated in a thermocycler using a program of denaturing at 95 °C for 1 min, PNA annealing at 78 °C for 10 s, forward tagging–primer annealing at 50 °C for 2 min and extension at 72 °C for 2 min, followed by a cooldown to 4 °C. The DNA, now tagged with both forward- and reverse-tagging primers, was cleaned with Agencourt beads using 17.5 µL of beads to clean the 25 µL of tagged template. A marginally more conservative bead-to-DNA ratio of 0.7:1 was used to clean the dual-tagged template as compared to single-tagged template because the overall length of dual-tagged template (<500 bp) is shorter than that of single-tagged template (>1 kbp). The dual-tagged DNA was eluted in 16 µL of water.

ITS tagging was similar to that for V4 16S, except that there was only one reverse primer in the 0.5 µM reverse working stock and only one forward primer in the 0.5 µM forward working stock.

PCR using tagged templates (our method)

We performed PCR in a 50-µL reaction mix, in which the reverse primer differed for each individually barcoded sample. The mix included 25 µL Kapa HiFi HotStart ReadyMix (KK2602, Kapa Biosystems), 2.5 µL PCR_F forward primer (from 5 µM working stock), 2.5 µL PCR_R_bc reverse primer (from 5 µM working stock), 5 µL mixed PNA working stock or water, and 15 µL DNA from the forward template–tagging step.

The PCR program was denaturation at 95 °C for 45 s followed by 34 cycles of denaturation at 95 °C for 15 s, PNA annealing at 78 °C for 10 s, primer annealing at 60 °C

for 30 s, extension at 72 °C for 30 s and then a cooldown to 4 °C. All samples were cleaned with Agencourt beads using 35 μL of beads to clean the 50-μL PCR (0.7:1). DNA was eluted in 50 μL water.

PCR using untagged templates (EMP method)

We used the primers and protocol available at http://www.earthmicrobiome.org/, with some exceptions to improve direct comparability with our method. The first exception to the published protocol is that we used 2× Kapa HiFi Ready Mix for the PCR, which is the same polymerase we used for the PCR in our method. The second exception is that we altered the thermocycling conditions to be more similar to ours (with the exception of the primer annealing temperature) and to include a PNA annealing step. The altered EMP thermocycling conditions were denaturing at 95 °C for 45 s followed by 35 cycles of denaturation at 95 °C for 15 s, PNA annealing at 78 °C for 10s, primer annealing at 50 °C for 30 s, extension at 72 °C for 30 s and ending with a cooldown to 4 °C. All samples were cleaned with Agencourt beads using 35 μL of beads to clean the 50-μL PCR (0.7:1). DNA was eluted in 50 μL of water.

Quantification of PCR products and library mixing.

From all cleaned PCR reactions, 1 μL was quantified in 96-well plate format using PicoGreen fluorescent dye (Invitrogen) and a fluorescence plate reader exciting at 475 nm and reading at 530 nm. The PCR reactions were mixed at equimolar ratios to make a pooled library for each run. For analysis purposes, in a setup run (Run A) and our primary run (Run B), we included from each run all potentially sequenceable material from low-yield and negative-control samples: low-quality material enriched for primer dimers and other abnormal amplicons that decrease the overall quality of the run. In Run C and Run D, samples with DNA below the detection limit were not used.

The mixed libraries were purified once more using Agencourt beads at a 0.7:1 bead-to-library ratio and were eluted in half the original volume to concentrate the final libraries. Each final library was quantified in triplicate using PicoGreen, and the values were averaged to reach a library quantification.

Library denaturation, dilution and sequencing

The final library was diluted to 4 nM, assuming an average amplicon length, including adaptors, of 448 bp. To denature the DNA, we mixed 5 µL of the 4 nM library with 5 µL of 0.2 N fresh NaOH and incubated 5 min at room temperature. 990 µL of chilled Illumina HT1 buffer was added to the denatured DNA and mixed to make a 20 pM library. Finally, 275 µL of the 20 pM library was mixed with 725 µL of chilled HT1 buffer to make a 5.5 pM sequenceable library, which was kept on ice until use. We noticed that 5.5 pM gave us a cluster density of between 700 K/mm^2 and 900 K/mm^2, which gave the best balance of quantity (which improves with higher cluster density) and quality (which improves with lower cluster density). The Illumina recommended range is 500 K/mm^2–1,200 K/mm^2. A 500-cycle v2 MiSeq reagent cartridge was thawed for 1 h in a water bath, inverted ten times to mix the thawed reagents, and stored at 4 °C a short time until use.

For sequencing in Run A, Run B and Run C using our method, the custom Illumina Nextera P1 primer ("Read1_seq"), was used as the forward sequencing primer for read 1 and was prepared by mixing 3 µL of 100 µM stock into 597 µL HT1 buffer to make a 0.5 µM solution. A MiSeq v2 flow cell was rinsed with water and ethanol and polished dry with lens paper. The 5.5 pM library was loaded into the "Load Sample" well, and the custom Nextera primer solution was loaded into port 18 of the reagent cartridge. The "Settings" section of the sample sheet was modified to include "C1" as the "CustomRead1PrimerMix" and "5′-AGATCGGAAGAGCACACGTC-3‴ as the adaptor. Read 2 was sequenced with the TruSeq read 2 sequencing primer already present in the reagent cartridge ("Read2_seq"), and the barcode read was sequenced with the TruSeq Index Read Sequencing Primer ("Barcode_seq"). The sample sheet along with sample names and the corresponding reverse complement of each nine-nucleotide barcode sequence was uploaded onto the MiSeq instrument before each run. The machine does not use the final base of the barcode read for annotation, and so each sample was associated with an 8-bp read sequence. Sequences for Read1_seq, Read2_seq, and Barcode_seq are shown below.

>Read1_seq
GCCTCCCTCGCGCCATCAGAGATGTGTATAAGAGACAG

>Read2_seq
GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

>Barcode_seq
GATCGGAAGAGCACACGTCTGAACTCCAGTCAC

For Run A and Run B, we applied a feature in Real-Time Analysis (RTA v1.17.22) that allowed the machine to use a hardcoded matrix and phasing calculations. This modification improved the performance of low diversity libraries. In order to do this we altered the MiSeqConfiguration.xml file (this modification required assistance from an Illumina field application specialist). For Run C, we upgraded our machine to the new version of Real-Time Analysis (RTA v1.17.28) and used the default feature of the upgrade without additional hardcoded matrix or phasing modifications.

For sequencing in Run D (EMP method), all custom sequencing primers were prepared by mixing 3 µL of 100 µM primer stock into 597 µL HT1 buffer to make a 0.5 µM solution. The custom primer “EMP_Read1_seq” was used as the forward sequencing primer for read 1 and was loaded into port 18 of the reagent cartridge. “EMP_Read2_seq” was used as the forward sequencing primer for read 2 and was loaded into port 20. “EMP_barcode_seq” was used to sequence the sample barcode and was loaded into port 19. The Settings section of the sample sheet was modified to include “C1” as the “CustomRead1PrimerMix,” “C2” as the “CustomIndexPrimerMix,” and “C3” as the “CustomRead2PrimerMix.” The sample sheet along with sample names and the corresponding reverse complement of each 12-nucleotide barcode sequence was uploaded onto the MiSeq instrument before the run. The machine does not use the final base of the barcode read for annotation, and so each sample was associated with an 11-bp read sequence. As with Run C, Run D was completed using Real-Time Analysis (RTA v1.17.28) without additional software modifications. Sequences for EMP_Read1_seq, EMP_Read2_seq, and EMP_Barcode_seq are shown below.

>EMP_Read1_seq
TATGGTAATTGTGTGCCAGCMGCCGCGGTAA

>EMP_Read2_seq
AGTCAGTCAGCCGGACTACHVGGGTWTCTAAT

>EMP_barcode_seq
ATTAGAWACCCBDGTAGTCCGGCTGACTGACT

Demultiplexing

Standard preprocessing and demultiplexing of PCR barcodes were performed with Consensus Assessment of Sequence and Variation (CASAVA) software (Illumina, v.1.8.2), allowing for 0 mismatches to the sample barcodes.

Raw sequence processing (our method)

Paired-end overlapping and merging, as well as recognition of pattern-matching sequences and MT processing, were performed using MTToolbox, a freely available software package hosted by SourceForge (https://sourceforge.net/projects/mttoolbox/). Documentation and user manuals can be accessed via the MTToolbox web page (https://sites.google.com/site/moleculetagtoolbox/).

Paired ends were overlapped with FLASH (Magoč and Salzberg 2011) using parameters "-m 30 -M 250 -x 0.25 -p 33 -r 250 -f 310 -s 20" for all V4 16S samples and "-m 20 -M 250 -x 0.25 -p 33 -r 250 -f 400 -s 20" for all ITS samples. In the overlapping region, the bases with the highest quality score were chosen for the merged reads, with bases from Read1 preferred in the case of ties (Figure 3.6e).

In Run B, merged sequences in each sample were then matched to expected patterns for either V4 16S amplicons or ITS amplicons using regular expressions. Because the merged V4 amplicons in Run C contained barcodes on the template-tagging Bc-MT-FS primers (Figure 3.2b), a slightly modified regular expression was used. These expressions select sequences without ambiguous bases or errors in priming sequences.

From the pattern-matching sequences, the sequence fragment 5′ to the forward linker and the fragment 3′ to the reverse linker were extracted and concatenated to form that sequence's molecular tag (MT), and the sequence occurring between the forward and reverse template–specific primers was extracted for analysis (Figure 3.6f). We did not analyze sequences corresponding to the primers because we observed high sequence variability at the wobble bases, even when amplifying a clonal template, which indicated that the wobble base observed in the sequence is a poor indicator of the primed sequence (data not shown).

Each unique MT observed in a sample was considered a unique MT category (Figure 3.6g). Sequences sharing the same MT were classified as belonging to the same category, and for each category containing two or more sequences, a multiple sequence alignment was built using command line ClustalW (Larkin MA 2007) with parameters "-output=gde -outorder=input -case=upper -query -quicktree" (Figure 3.6h). A consensus sequence was calculated from the multiple sequence alignment by choosing the most common base at each position. For MT categories containing only two sequences (and for all other ties), the base with the highest average quality score was chosen; and if a tie could still not be resolved, an IUPAC base was used to indicate the tie in the consensus sequence. For each sample, a FASTA file of consensus sequences was built, with each consensus sequence given a composite name including the sample of origin, or "P_number_ID" followed by the MT of that consensus. For example: >P0_GGCTGACTTTAC-GGCAGTCAAT [Sequence].

MT categories in each sample that contained only one sequence (category depth = 1) could not be represented by a consensus, and the sequences in these categories, or 'singletons', were kept in a separate FASTA file with each sequence given a composite name including the sample the sequences came from, the sequence number within the corresponding sample, the MT sequence and the original read ID. For example:

>P0_20176 GAGTAGGAATA-TCTAT UNC20:76:000000000-A315U:1:1101:14750:16671:N:0:GGCGCTTA

[Sequence]

Raw sequence processing (EMP method)

Paired ends were overlapped with FLASH21 using parameters "-m 30 -M 250 -x 0.25 -p 33 -r 250 -f 310 -s 20."

EMP sequencing primers provide data between the highly conserved areas bound by the 515F and 806R primers; thus, regular expressions for these primers cannot be used to identify pattern-matching sequences. Therefore, we define high-quality sequences in the context of EMP data as sequence that successfully overlaps and merges and does not have ambiguous bases.

Operational taxonomic unit (OTU) formation

OTUs were built using OTUpipe, a collection of USearch (http://www.drive5.com/) commands encapsulated in a bash script that clusters sequences on the basis of their nucleotide identity and that removes chimeras that can form during PCR. First, FASTA files from samples to be clustered were concatenated into one file. OTUpipe was then run with nondefault parameters ABSKEW = 3 and MINSIZE = 1. For 99% OTU clustering, the following nondefault parameters were used: PCTID_ERR = 99, PCTID_OTU = 99, PCTID_BIN = 99. We did not make OTUs at higher than 99% because a single bacterial genome can harbor several copies of the 16S rRNA gene that differ on average by 0.55% (Pei et al. 2010), meaning that at identity thresholds higher than 99%, a single bacterium would form several OTUs even if error was eliminated.

OTU table construction

OTUs were built into OTU tables, and their taxonomy was assigned using functions in QIIME 1.5.0 (Caporaso et al. 2010). The OTUpipe output file “readmap.uc” was transformed into a QIIME cluster file by running the QIIME script “readmap2qiime.py,” generating the text file “qiime_otu_clusters.txt.” This file was passed to the QIIME script “make_otu_table.py” to make a Biological Observation Matrix (BIOM) OTU table. Finally, the BIOM table was converted to a classic format OTU table using the QIIME script “convert_biom.py.”

Assigning taxonomy to OTUs

Taxonomy was assigned to bacterial OTUs using the RDP classifier trained on the most recent (4 February 2011) Greengenes 97% identity taxonomy representatives and was accomplished by running the QIIME 1.5.0 script “assign_taxonomy.py” on OTU representative sequences using “greengenes_tax_rdp_train.txt” as the ID to taxonomy mapping file, “gg_97_otus_4feb2011.fasta” as the reference sequences and the parameter “-c 0.5.” Helpful instructions for running the QIIME scripts can be found by searching for the script name on the QIIME website (http://www.qiime.org/).

Owing to a focus on bacterial taxa, RDP trained on Greengenes did a poor job of recognizing plastid and mitochondrial sequences in our data. Rather than editing the training set, we further recognized plant contaminant OTUs by using BLAST to compare the representative sequences to a custom database containing the Arabidopsis 18S rRNA sequence as well as plastid and mitochondria 16S rRNA sequences from Arabidopsis and other plants. We used BLAST with an E value of 0.00001 and a percent identity of 94.

Predicting pPNA and mPNA utility across diverse plant families

The pPNA and mPNA sequences were tested for exact matches to representative chloroplast and mitochondrial 16S sequences from diverse plant species found in NCBI GenBank (Figure 3.16).

Subsampling

Normalization of FASTA files and all other subsampling was performed using the sample() function in the “base” library of R (http://www.r-project.org/). Rarefaction of OTU tables was performed using the function rrarefy() the “vegan” library of R, which also makes use of the sample() function.

Permutation tests

All permutation tests involved 24 samples and asked whether the mean value of 12 samples in “condition low” was lower than the mean value of 12 samples in “condition high.” For each permutation test, the values from the 24 samples were randomly assigned into two groups of 12 using the sample() function in the base library of R, and the difference in the means of these groups was taken. This was repeated 10,000 times per test to form the probability distribution for each test. The P value was the fraction of 10,000 permutations in which the observed difference in the means would be as large due to chance.

A nonparametric test on the means was chosen in preference to a parametric t-test because of relatively low group size of 12 samples, which prevents accurate estimation of

the underlying probability distributions and is not sufficiently large to make the assumption of normality under the Central Limit Theorem.

Correction for multiple testing

Permutation tests were used to test whether the relative abundances of bacterial families and bacterial OTUs were lower in PNA samples than in control samples, for all families and OTUs above the threshold (see Figure Specific Details). The green and red histograms of uncorrected P values display the results of these permutation tests for pPNA and mPNA (Figures 3.14 and 3.15). The P values within each histogram were corrected for multiple testing using the Benjamini-Hochberg false discovery rate (FDR) method as implemented by the p.adjust() function in the "stats" library of R, and the number of tests that were included in each application of the FDR method is shown beneath each P value histogram.

Chi-squared tests

The green and red histograms of uncorrected P values display the results of permutation tests for pPNA and mPNA, respectively (Figures 3.14 and 3.15a,b). For root EC families and OTUs, and for soil families, there were ~100 or fewer tests, and ten bins were used for the P value histogram (Figures 3.14a and 3.15a,b). For soil OTUs, there were 1,010 tests for each PNA, and 20 bins were used for higher resolution of the distribution histogram (Figure 3.14b). Each histogram was compared to the null flat distribution (equal number of P values in each bin of the histogram) using a Chi-squared test with 9 degrees of freedom for histograms with 10 bins or 19 degrees of freedom for histograms with 20 bins. Chi-squared tests were performed using the function chisq.test() in the stats library of R.

Accession codes

Sequence Read Archive: ERP003492

FIGURE-SPECIFIC DETAILS

Figure 3.1: Variable regions in the 16S rRNA gene

All sequences without unambiguous bases in the Greengenes training set of full length 16S sequences (29,846 sequences) were aligned to the pre-aligned Greengenes core set using PyNAST with default parameters as implemented in the QIIME script "align_seqs.py". The *E. coli* 16S sequence (PMID: CP002967.1) was also aligned. For each base position (non-gap position) in the *E. coli* alignment, the number of A, C, T, G, and gap characters in the corresponding position of all sequences in the Greengenes alignment was counted and the Shannon diversity for this position was calculated and graphed (light blue vertical needles). The moving average of the Shannon diversity (thick waving black line in was calculated by taking the mean Shannon diversity at each *E. coli* base position considering a 50 bp sliding window stretching 25 bases 5' and 25 bases 3' of the base position considered – for this reason the black line is not graphed for the first 25 and the last 25 bases of the alignment. We note that this interpretation does not show the Shannon diversity at positions in the alignment for which the *E. coli* sequence shows a gap. Location of the hypervariable regions was based on mapping information in Chakravorty et al., 2007(Chakravorty et al. 2007). Degenerate regular expressions of common primers were used to map primer locations to the *E.coli* reference on the x-axis. The charts were produced with the plot() and points() functions in the "base" library of R(Team 2012).

Figure 3.3b: Library diversity simulation

We simulated in *silico* a PCR template composed of 1,000 identical copies of a single V4 16S sequence, as well as a more realistic template composed of 1,000 real bacterial V4 16S sequences from a root EC sample. To mimic the effect of using frameshifting primers to PCR each template, subsets of the 1,000 sequences were randomly assigned to equally-sized groups to which six frameshifting treatments of 0-5 additional 5' bases were applied. To visualize the effect of mixing in phiX174 genomic DNA post-PCR, the phiX174 genome [NCBI GenBank ID: NC_001422] was randomly fragmented and the fragments were used to replace specific fractions of the 1,000 V4 16S sequences in the simulated PCRs. For each treatment of frameshifts and / or phiX174, the first 250 bp of each sequence was considered. Shannon diversity at each base position was calculated from the number of A, C, T, and G bases present at that position. The charts were produced with the boxplot() and points() functions in the "base" library of R(Team 2012).

Figure 3.4a-b, 3.5a-d: Q Score histograms, plots, and heatmaps, and base diversity per cycle

Illumina Q scores are equivalent to 10 times the log10 of the reciprocal of the error rate. Q score histograms, plots, and heatmaps, and the graph of % base at each cycle, were generated from raw data on the MiSeq machine using Sequence Analysis Viewer version 1.8.11

Figure 3.4c: Error rate for clonal 16S samples

Identical to Figure 3.8b, except that only reads from Run A and Run B processed by method 1, (NT), were used.

Figure 3.7: Monte Carlo simulation of MT uniqueness

A custom R script was written to generate 100,000 oligonucleotide (A, C, T, or G) *N*-mers each for N's of 10, 11, 12, 13, and 14. For each *N-mer* length, the number of non-unique oligos in the set was divided by 100,000 to give the fraction of non-unique oligos, and then multiplied by 100 to give the percentage that is graphed. This process was also repeated for depths of 75,000, 50,000, and 25,000 *N*-mers. The entire simulation was then repeated 4 additional times, and all 5 replicates for each *N*-mer length were graphed. The chart was produced with the geom_line() function in the "ggplot2" library of R(Wickham 2009).

Figure 3.8a: Copy number per MT for clonal 16S samples

Pattern-matching sequences from Run B of all the clonal 16S template samples, including both replicates of the no dilution, 50× dilution, and 100× dilution samples were rarefied to 40,000 sequences per sample. The sequences were then categorized by their MT (as in Figure 3.6g), but were not made into ConSeqs. A histogram was plotted of the number of sequences falling at each discrete MT category depth. The chart was produced with the geom_density() and geom_line() functions in the "ggplot2" library of R(Wickham 2009).

Figure 3.8b: Error rate for clonal 16S samples

Pattern-matching sequences of the clonal 16S template samples from only the 50× dilution and 100× dilution samples, as well as their replicates for a total of four samples were

gathered from Run B. The MT and amplified template sequences were extracted (as in Figure 3.6f). The sequences were processed four ways to form four comparison groups:

1) *"NT", or no tag*, contained a mix of all pattern-matching sequences from all four samples, regardless of the MT.

2) *"ConSeqs", or ConSeqs of two or more sequences*, in which pattern-matching sequences in each sample were categorized by their MT and ConSeqs were constructed from the multiple sequence alignments. All ConSeqs made from MT categories containing 2 or more sequences were pulled from each sample and pooled.

3) *"S", or singletons*, in which pattern-matching sequences in each sample were categorized by their MT, and all the sequences with a unique single-copy molecular tag were pulled from each sample and pooled.

4) *"PConSeqs", or perfect ConSeqs made from three or more sequences*, in which pattern-matching sequences in each sample were categorized by their MT and ConSeqs were constructed from the multiple sequence alignments as described above. Only alignments of three or more sequences in which all constituent sequences were 100% identical were considered, and the ConSeqs (in this case PConSeqs) of these perfect alignments were pulled from each sample and pooled.

The four comparison groups were then each rarefied to 15,000 sequences, with the exception of the PConSeqs, which were a rarer class and used in full because only 4,777 were available in the run. The sequences were all aligned to a common set of pre-aligned templates using PyNAST, with default parameters as implemented in the QIIME script "align_seqs.py". The Sanger sequence of the clonal 16S template (sequence below), trimmed to the region between the 515F and 806R primers, was also aligned using PyNAST.

>Mycobacterium_16S_clone
TACGTAGGGTCCGAGCGTTGTCCGGAATTACTGGGCGTAAAGAGCTCGTAGGTGGTTTGTCGCGTTGT
TCGTGAAAACTCACAGCTTAACTGTGGGCGTGCGGGCGATACGGGCAGACTTGAGTACTGCAGGGGAG
ACTGGAATTCCTGGTGTAGCGGTGGAATGCGCAGATATCAGGAGGAACACCGGTGGCGAAGGCGGGTC
TCTGGGCAGTAACTGACGCTGAGGAGCGAAAGCGTGGGGAGCGAACAGG

For each aligned comparison group, positions that were gaps in all aligned sequences *and* the Sanger reference sequence were not considered. In every case, gaps in the PyNAST comparison group alignments matched gaps in the Sanger alignment, indicating that the frequency of insertion errors in this sequence was extremely low. Next, for each base in the aligned Sanger sequence, the SNPs and gaps for all other sequences in the 15,000 sequences (or 4,777 for PConSeqs) of the comparison group were counted. This value was divided by 15 (or 4.777 for PConSeqs) to generate the errors per thousand (ept) at each base. The mean error rates per thousand were calculated by taking the mean of the per-base errors per thousand across each of 253 bases of the sequence. The chart was produced with the geom_line() function in the "ggplot2" library of R(Wickham 2009).

Figure 3.8c: OTU analysis of clonal 16S samples

Pattern-matching sequences from Run B of the clonal 16S template samples from only the 50× dilution and 100× dilution samples were unprocessed (NT) or processed into ConSeqs (ConSeqs). Each comparison group was rarefied to 30,000 sequences per sample and these sequences were clustered into OTUs at 97% or 99% identity. The OTUs were ordered by their relative abundance, and the number of sequences in each ranked OTUs is graphed for each comparison group. To determine the number of OTUs necessary to represent 95% of the data, the sequences in the OTUs were summed, starting with the most abundant, until 28,500 sequences (95% of 30,000) were accounted for. The chart was produced with the geom_line() and geom_point() functions in the "ggplot2" library of R(Wickham 2009).

Figure 3.9a: Rarefaction curves of different MT treatments

Pattern-matching sequences were gathered from Run B and the MT and amplified template sequences were extracted (as in Figure 3.6f). The sequences were processed four ways to form four comparison groups:

Comparison groups:

1) *"NT", or no tag*, as described for Figure 3.8b and Figure 3.4c.

2) *"ConSeqs", or ConSeqs of two or more sequences*, as described for Figure 3.8b.

3) *"S", or singletons*, as described for Figure 3.8b.

4) "CAS", *or ConSeqs with adjusted singletons*. For each sample, sequences were categorized by their MT, and ConSeqs were constructed from the multiple sequence alignments. The *ConSeqs collapse ratio*, or the number of ConSeqs divided by the number of constituent sequences in the multiple sequence alignments, was calculated. Next, the singletons were quantified. The number of singletons was multiplied by the *ConSeqs collapse ratio*, rounded to the nearest integer, and then the singles were down-sampled to this integer. This adjustment thus keeps the ratio of singles to all other sequences constant, even as all other sequences are collapsed into their ConSeqs.

OTUs tables were formed from each comparison group, using 97% and 99% identity thresholds for clustering. Because the number of sequences in each comparison group varied substantially, with the NT group having many more sequences than the other groups, the FASTA files containing sequences from each comparison group were each normalized to 500,000 sequences. Each comparison group was then clustered independently at 97% or 99% sequence identity to produce 4 OTU tables. Plastid and mitochondrial OTUs were removed computationally, and bacterial reads for root EC and soil samples across all tables were pooled, producing a soil pool and a root EC pool per OTU table. These pools were rarefied at intervals of 1,000 sequences and the number of OTUs observed at each depth was plotted. The chart was produced with the geom_line() and geom_point() functions in the "ggplot2" library of R(Wickham 2009).

Figure 3.9b: Progressive drop-out analysis of technical reproducibility

The same four OTU tables representing the four comparison groups were used as in Figure 2a, with four exceptions. First, plastid and mitochondrial OTUs were *not* removed computationally. Second, in each OTU table, we considered the technical reproducibility of 12 pairs of root EC samples and 12 pairs of soil samples, for a total of 24 pairs, where each member of a pair was independently template-tagged, treated with water or PNA, and amplified. These 24 pairs were chosen because these samples had good sequencing depth and reasonably diverse microbial composition. Third, each sample was rarefied to a common inter-table depth. Fourth, within each table, the more deeply-sequenced pair member for each of the 24 technical replicate pairs was rarefied to the number of sequences

of the less-sequenced sample in the pair, such that the sequencing depth of the pair members was equal.

For each comparison group, the relative abundance of each OTU in one technical replicate pair member was log10-transformed to correct for heteroscedasticity and plotted against the log10-transformed relative abundance of that same OTUs in the other technical replicate pair member. This was repeated for all 24 pairs on the same set of axis, generating a densely-populated linearly-correlated scatterplot for each comparison group, similar to that previously published(Benson et al. 2010; Bulgarelli et al. 2012; Lundberg et al. 2012). The R^2 coefficient of determination was then calculated for the scatterplot and graphed.

The low-abundance OTUs in an OTU table either represent rare but real sequences, or sequence errors, and are less-reproducible than larger OTUs. We dropped OTUs from the scatterplot that did not meet the threshold abundance (x-axis) in at least one pair member in at least one of the 24 pairs, and recalculated R^2 at each threshold, generating the upward-sloping curves. The chart was produced with the geom_line() and geom_point() functions in the "ggplot2" library of R(Wickham 2009).

Figure 3.10: Principal Coordinate Analysis of Weighted Unifrac distances

ConSeqs from our method in Run C, or high quality sequences from the EMP method in Run D, were clustered into OTUs with OTUpipe as described above using a 97% identity threshold, forming a separate OTU table for each run. Each sample in the OTU table from our method was rarefied to 1,200 ConSeqs, while each sample in the OTU table from the EMP method was rarefied to 1,200 high quality sequences. For each run, the OTU representative sequences were aligned to a common set of pre-aligned templates using PyNAST, with default parameters as implemented in the QIIME script "align_seqs.py". The full alignments were then filtered and clustered into phylogenetic trees using the QIIME script "filter_alignment.py" followed by "make_phylogeny.py". The phylogenetic trees and the OTU tables were used in the QIIME script "beta_diversity.py" to return, for each OTU table, a pairwise matrix of weighted Unifrac distances between all samples. Principal Coordinate Analysis ordination was performed using the pcoa() function in the "ape" library of R(Paradis et al. 2004), and the first two principal coordinates were plotted using the geom_point() function in the "ggplot2" library of R(Wickham 2009).

Figure 3.13a, left: Relative abundance of contaminant sequences

Pattern-matching sequences in Run B were processed into ConSeqs which were clustered at 97% identity to form an OTU table. The twelve root EC samples with the PNA titrations and their technical replicates were extracted from this OTU table and rarefied to the smallest sample of the 24 (6,880 sequences). The relative abundance of bacterial sequences, plastid sequences, mitochondrial sequences, and other sequences were expressed as a percentage. The stacked bar chart was produced with the geom_bar() in the "ggplot2" library of R(Wickham 2009).

Figure 3.13a, right: Mean number of sequences per multiple sequence alignment

Pattern-matching sequences from root EC and soil samples, the same used in Figure 3.13a, left, were gathered from Run B and the MT and amplified template sequences were extracted (as in Figure 3.6f). The sequences were processed into ConSeqs, but just prior to formation of the ConSeqs from the multiple sequence alignments, the number of sequences in all multiple sequence alignments was counted. The average number of sequences per multiple sequence alignment per sample is graphed. The bar chart was produced with the geom_bar() in the "ggplot2" library of R(Wickham 2009).

Figure 3.13b and 3.13c, and Figures 3.10 and 3.14: Heatmaps

Pattern-matching sequences from Run B were processed into ConSeqs, which were clustered at 97% identity to form an OTU table. All contaminant OTUs were removed, leaving only bacterial OTUs. For root EC heatmaps, 12 root EC samples and their technical replicates were extracted from this OTU table and rarefied to the smallest sample of the 24 (1,092 bacterial ConSeqs). For soil heatmaps, 12 soil samples and their technical replicates were extracted from this OTU table and rarefied to the smallest sample of the 24 (11,593 bacterial ConSeqs). For Figure 3.13b and Figure 3.13c, the bacterial OTUs in each table were then reclassified at the family level, and OTUs from the same bacterial family were combined to convert the OTU table into a family-level table. Bacterial families that did not have an abundance of 5 ConSeqs in at least one of the 24 samples were removed to avoid visualizing rare families prone to sampling artifacts. For Figure 3.14, the bacterial OTUs were *not* reclassified at the family level, and OTUs that did not have an abundance of 5 ConSeqs in at least one of the 24 samples were removed. For better visualization in all heatmaps, abundances were transformed to log2 *per mille* $\log_2(1000x+1)$ prior to color assignment – this transformation is reflected in the color key. *The log2 transformation was*

for visualization only and transformed data was not used for statistical tests. All heatmaps were made using the function heatmap.2() from the "gplots" library of R(Warnes 2011). Hierarchical clustering of rows and columns in the heatmaps is based on Bray-Curtis dissimilarity and uses group-average linkage.

Figure 3.12: Exhaustive search for PNA oligo candidates

Method described under "Peptide Nucleic Acid (PNA) design". The black histogram of *k*-mer matches to the database was made using the plot() function in the "base" library of R(Team 2012), with the abline() function used to add the vertical red lines. Degenerate regular expressions of common primers were used to map primer locations to the plastid or mitochondrial sequence along the x-axis.

Figures 3.14 and 3.15: Bacterial family and OTU relative abundance for different PNA treatments

In panel a (root EC) and b (soil) for both Supplementary Figures, the relative abundance of each bacterial family or OTU was compared between the 12 samples amplified using either pPNA (left) or mPNA (right) and the remaining 12 samples not containing pPNA or mPNA respectively. The test used was a permutation test on the means (Online Methods). Histograms of *P*-values were made with the hist() function in the "base" library of R(Team 2012). The distribution of *P*-values was compared to the null flat distribution using a Chi-squared test (Online Methods).

Figure 3.16: Use of PNA on *A. thaliana* and *O. sativa* leaf DNA

In **a**, the chloroplast and mitochondrial 16S sequences used to determine if pPNA and mPNA respectively were likely to function were taken from NCBI GenBank.

In **b**, pattern-matching sequences in Run B were processed into ConSeqs which were clustered at 97% identity to form an OTU table. Sixteen leaf samples from *A. thaliana* and *O. sativa* were extracted from this OTU table and rarefied to the smallest sample of the 16 (161 sequences). The relative abundance of bacterial sequences, plastid sequences, mitochondrial sequences, and other sequences were expressed as a percentage. The stacked bar chart was produced with the geom_bar() in the "ggplot2" library of R(Team 2012). Next, pattern-matching sequences in Run B were processed into NT-sequences which were clustered at 97% identity to form an OTU table, and the 16 leaf samples were extracted. The number of all NT sequences in each sample, without normalization, is

graphed in the dark blue bars. Contaminant OTUs were removed and the number of usable bacterial reads, without normalization, is graphed in brown bars. The dark blue and brown bar plots were produced with the function barplot() in the “graphics” library of R(Wickham 2009).

Supplementary Figure 16: Internal Transcribed Spacer (ITS) amplicons

Pattern-matching sequences from Run B were processed into ConSeqs, which were clustered at 97% identity to form an OTU table. Root EC samples amplified with V4 16S primers and root EC samples amplified with ITS2 primers were pooled and each pool was rarefied to a common value of 14,112 ConSeqs. The number of bases in each OTU was calculated (OTU length) for ITS and 16S OTUs, and then the number of OTUs at each OTU length was graphed in panel a using the plot() function in the “base” library of R(Team 2012), as was the rank-abundance curve in panel b.

Supplementary Figure 17: Primer linkers

Two bases 5-prime of the 515F primer and two bases 3’ of the 806R primer were extracted from all sequences without unambiguous bases in the Greengenes 97% training set of full length 16S sequences (29,846 sequences) that matched expected patterns for V4 amplicons, and the frequency of each base at all four positions was graphed. The figure represents the + strand, and so the reverse complement of the linkers in both our 806R primers and the Earth Microbiome Project(Caporaso et al. 2012) primers are displayed.

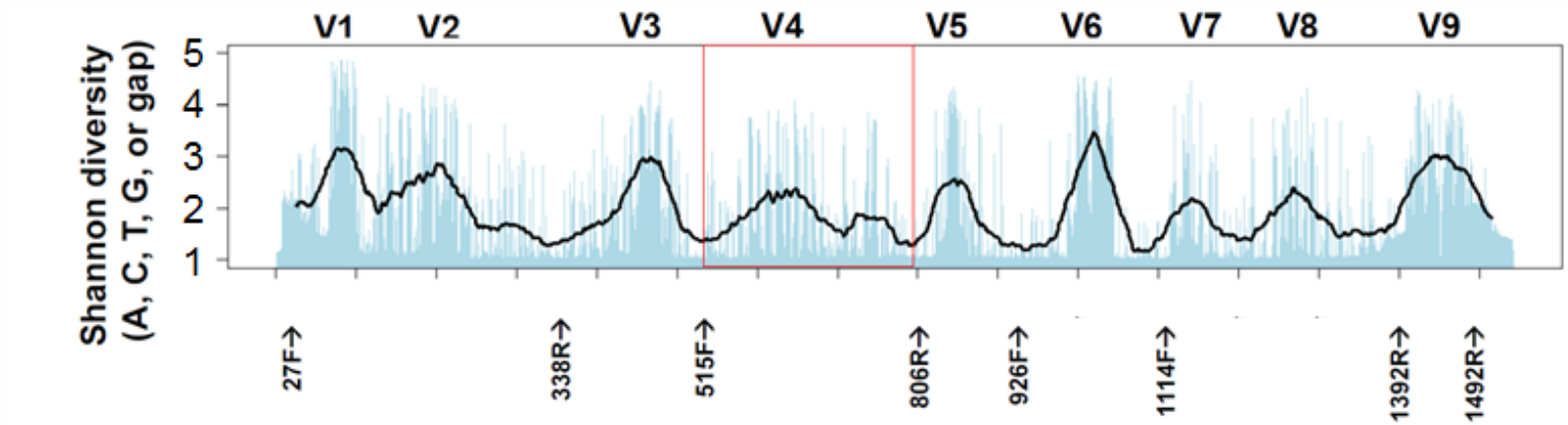


Figure 3.1. Reference Map of the 16S rRNA gene.

Map shows variable regions V1-V9 (above chart) and the locations of common primers (based on conventional *E. coli* numbering, below chart). For each base present in *E. coli*, the Shannon Diversity of bases or gaps for that position is graphed in light blue histograms. The average Shannon Diversity based on a 50 bp sliding window is charted as a black line, displaying the classic 16S variable regions. The variable region V4 used in this study is boxed in red. Diversity was calculated by comparison to the Greengenes 97% representatives (most recent Feb. 4 2011 version) database of full length 16S sequences (Methods).

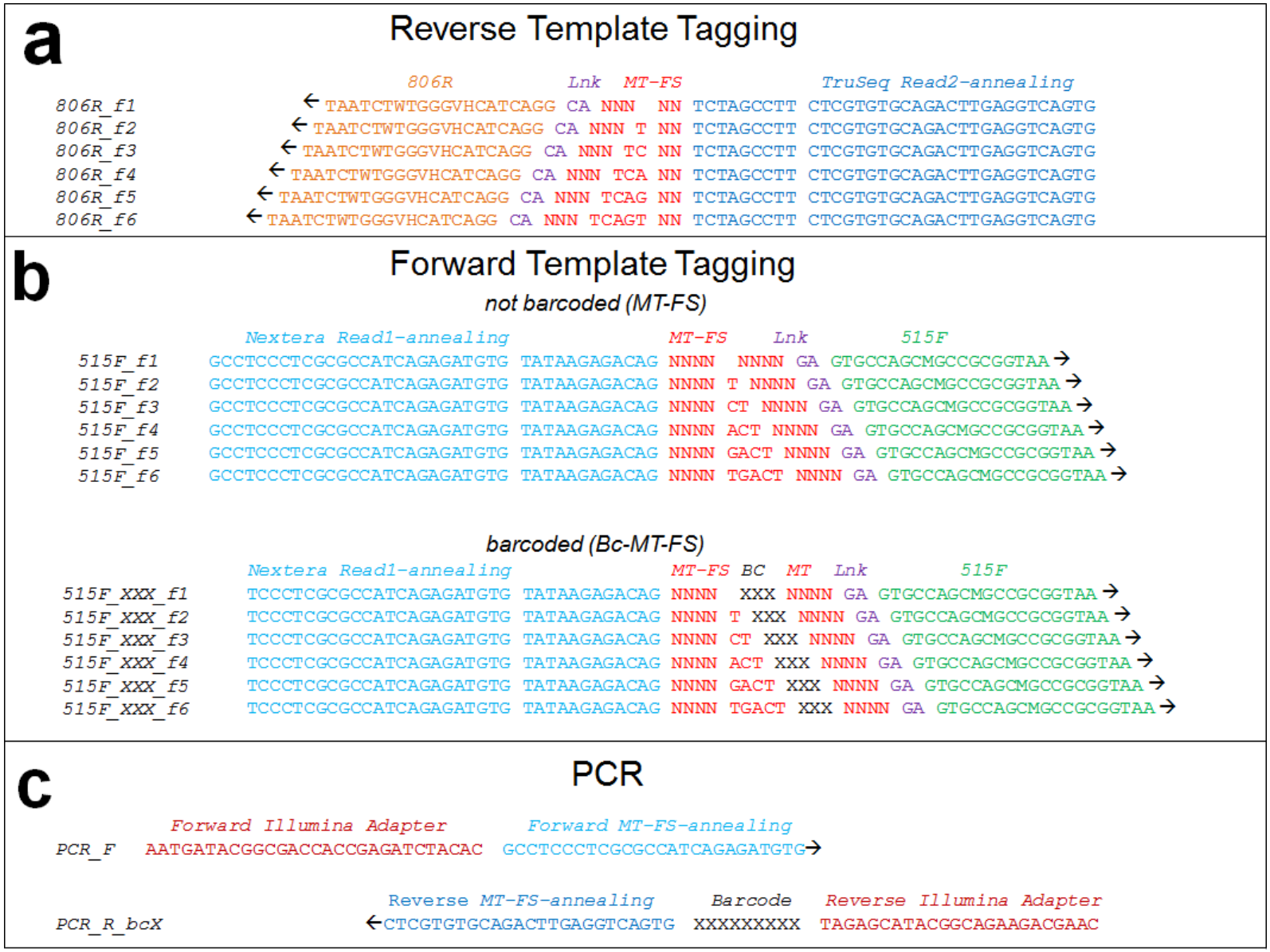

a
Reverse Template Tagging
806R
Lnk
MT-FS
TruSeq Read2-annealing
806R_f1 ← TAATCTWTGGGVHCATCAGG CA NNN NN TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG
806R_f2 ← TAATCTWTGGGVHCATCAGG CA NNN T NN TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG
806R_f3 ← TAATCTWTGGGVHCATCAGG CA NNN TC NN TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG
806R_f4 ← TAATCTWTGGGVHCATCAGG CA NNN TCA NN TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG
806R_f5 ← TAATCTWTGGGVHCATCAGG CA NNN TCAG NN TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG
806R_f6 ← TAATCTWTGGGVHCATCAGG CA NNN TCAGT NN TCTAGCCTT CTCGTGTGCAGACTTGAGGTCAGTG
b
Forward Template Tagging
not barcoded (MT-FS)
Nextera Read1-annealing
MT-FS
Lnk
515F
515F_f1 GCCTCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_f2 GCCTCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN T NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_f3 GCCTCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN CT NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_f4 GCCTCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN ACT NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_f5 GCCTCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN GACT NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_f6 GCCTCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN TGACT NNNN GA GTGCCAGCMGCCGCGGTAA →
barcoded (Bc-MT-FS)
Nextera Read1-annealing
MT-FS
BC
MT
Lnk
515F
515F_XXX_f1 TCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN XXX NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_XXX_f2 TCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN T XXX NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_XXX_f3 TCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN CT XXX NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_XXX_f4 TCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN ACT XXX NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_XXX_f5 TCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN GACT XXX NNNN GA GTGCCAGCMGCCGCGGTAA →
515F_XXX_f6 TCCCTCGCGCCATCAGAGATGTG TATAAGAGACAG NNNN TGACT XXX NNNN GA GTGCCAGCMGCCGCGGTAA →
c
PCR
Forward Illumina Adapter
Forward MT-FS-annealing
PCR_F AATGATACGGCGACCACCGAGATCTACAC GCCTCCCTCGCGCCATCAGAGATGTG →
Reverse MT-FS-annealing
Barcode
Reverse Illumina Adapter
PCR_R_bcX ← CTCGTGTGCAGACTTGAGGTCAGTG XXXXXXXXX TAGAGCATACGGCAGAAGACGAAC

Figure 3.2. Schematic of molecular tagging - frameshifting template tagging primers.

(a) MT-FS V4 16S reverse template-tagging primers.

(b) Forward “MT-FS” V4 16S template-tagging primers (top), and forward barcoded “Bc-MT-FS” V4 16S template-tagging primers (bottom), where “XXX” is a three base pair barcode. MT-FS = Molecular tag and frameshifting bases. Lnk = Linker. “N” = MT random sequence **(c)** PCR primers.

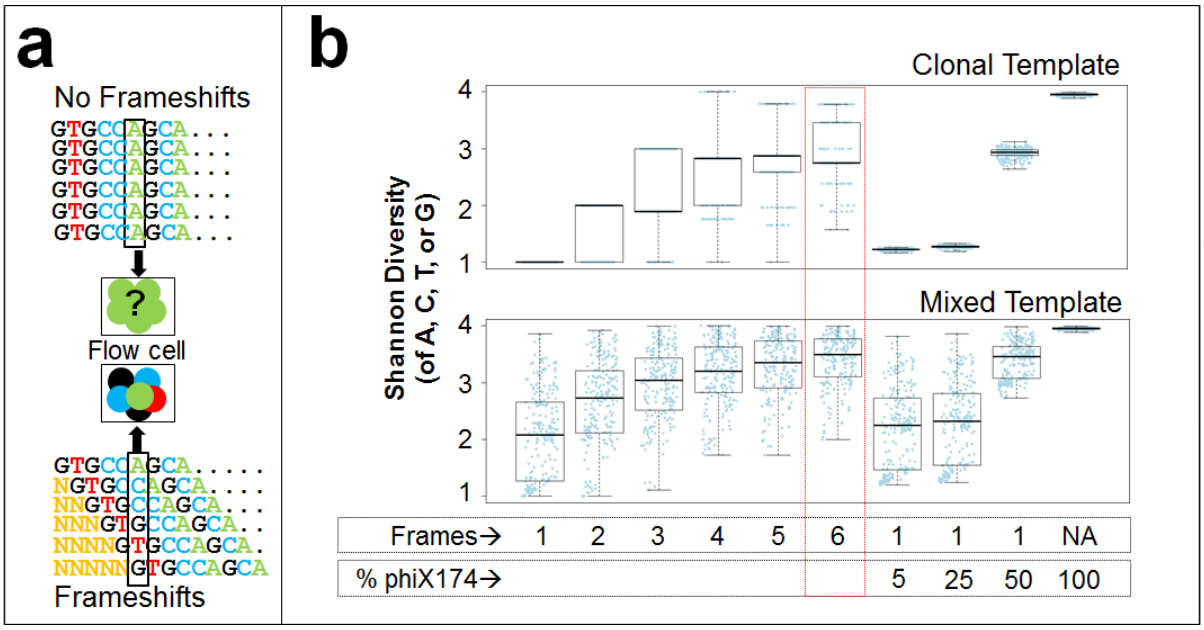


Figure 3.3. Frameshifting primers enhance library diversity.

(a) Schematic showing that frameshifts can impose diversity on a low-diversity library.

(b) Diversity per sequenced base for simulated libraries made from a perfect clonal template (top) or a low-complexity template of 1000 real V4 bacterial 16S rRNA sequences (bottom). For each simulated library, subsets of 1000 sequences were randomly assigned to equally-sized groups to which six frameshifting treatments of 0-5 additional 5' bases were applied, creating between 1 and 6 frames ("Frames", below xaxis). Some libraries received simulated fragments of phiX174 genomic DNA in place of a fraction of the 1000 V4 16S sequences ("%phiX174", below x-axis). For each library, the Shannon diversity for each of the first 250 sequences was graphed (light blue dots), and the distribution summarized with a box-and-whiskers plot showing the extremes, upper and lower quartiles, and the median. Six frameshifts and no phiX174 were used in the remainder of this study (red box).

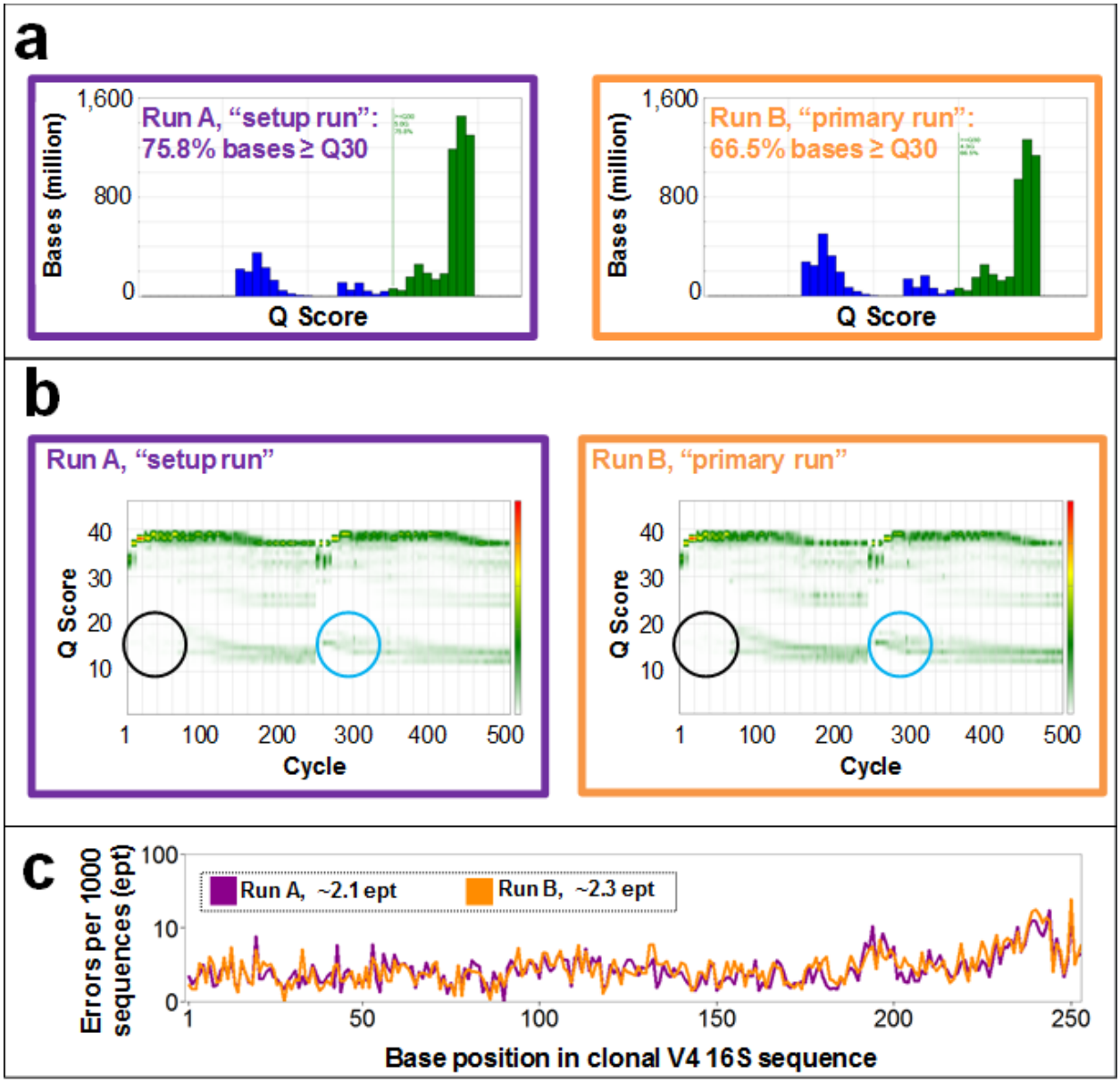

a
Run A, "setup run":
75.8% bases ≥ Q30
Bases (million)
1,600
800
0
Q Score
Run B, "primary run":
66.5% bases ≥ Q30
Bases (million)
1,600
800
0
Q Score
b
Run A, "setup run"
Q Score
40
30
20
10
1
100
200
300
400
500
Cycle
Run B, "primary run"
Q Score
40
30
20
10
1
100
200
300
400
500
Cycle
c
Errors per 1000 sequences (ept)
100
10
0
Run A, ~2.1 ept
Run B, ~2.3 ept
1
50
100
150
200
250
Base position in clonal V4 16S sequence

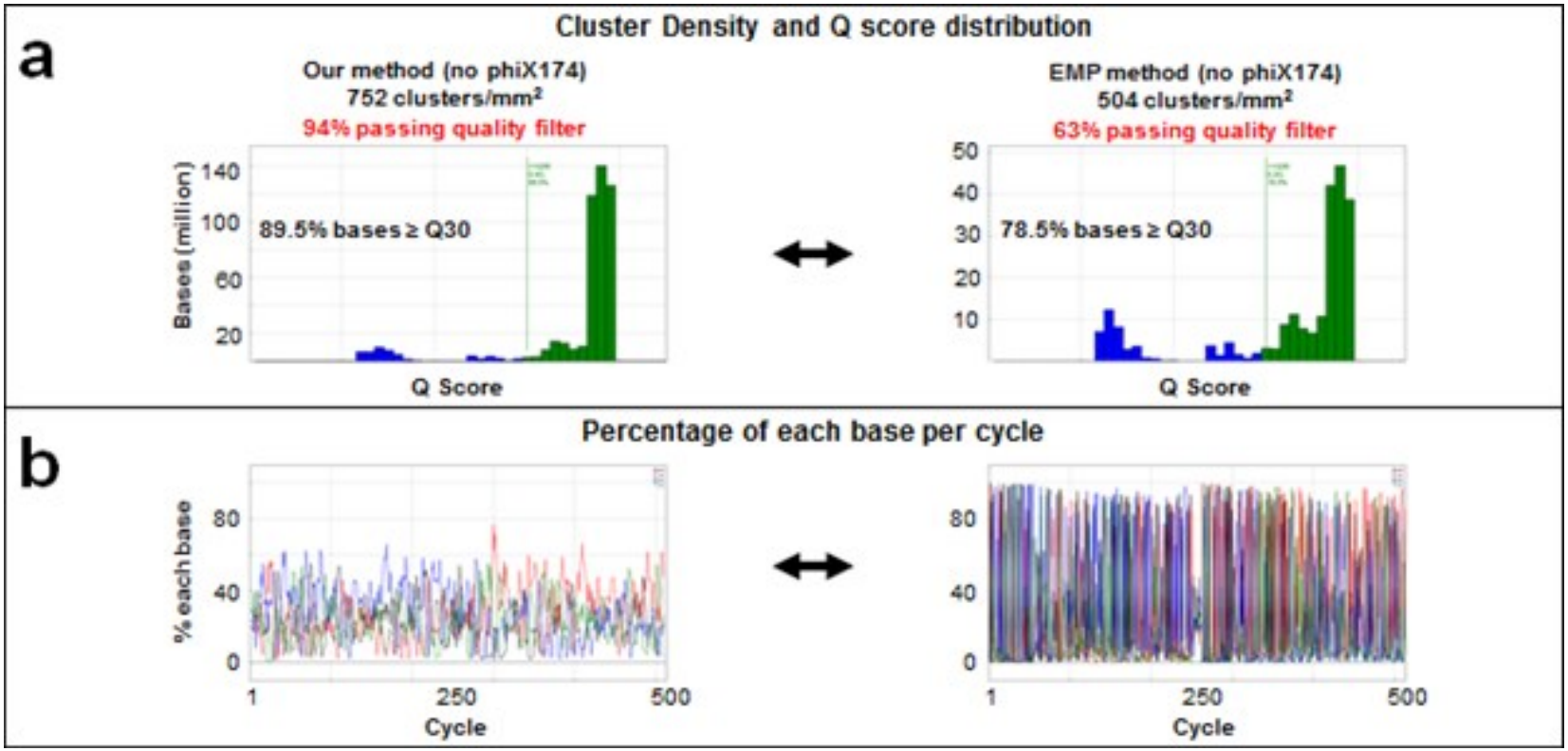
Figure 3.4. MiSeq run quality for Run A (setup run) and Run B (primary run).

Run A, a setup run, met Illumina quality specifications of sequencing pure phiX174 DNA and Run B, the primary run we analyze, came close.

(a) Illumina MiSeq performance specifications for a 2 × 250 run of phiX174 is >75% of total bases above Q30 (not per cycle). A setup run without any phiX174 DNA, but containing a sample composition differing only in the initial concentration of several templates and library mixing (Materials and Methods), met the advertised specifications based on the machine's statistics (top, purple box). The primary run we analyze (bottom; orange box), made up of a nearly identical composition of samples, was close. This was despite deliberate inclusion in these runs of all potentially-sequenceable material from low-yield and negative control samples.

(b) Q Score heatmaps for setup run A (left; purple) and primary run B (right; orange). Both runs show sustained high quality, with diminishing quality towards the end of each run, and lower quality at the beginning of Read2 than of Read1 (circles).

(c) Analysis of error rate across merged reads of a plasmid-borne clonal 16S rRNA template sample present in both runs reveals that the sequencing quality is similar in both runs. The mean error rate for pattern-matching (Materials and Methods) in each run is ~2.2 errors per thousand (ept) (color key), or Q27, with the error rate increasing towards the 3' end of the read representing the non-overlapping portion of read 2, as expected.


a
Cluster Density and Q score distribution
Our method (no phiX174)
752 clusters/mm²
94% passing quality filter
89.5% bases ≥ Q30
Bases (million)
140
100
60
20
Q Score
EMP method (no phiX174)
504 clusters/mm²
63% passing quality filter
78.5% bases ≥ Q30
50
40
30
20
10
Q Score
b
Percentage of each base per cycle
% each base
80
40
0
1
250
500
Cycle
80
40
0
1
250
500
Cycle

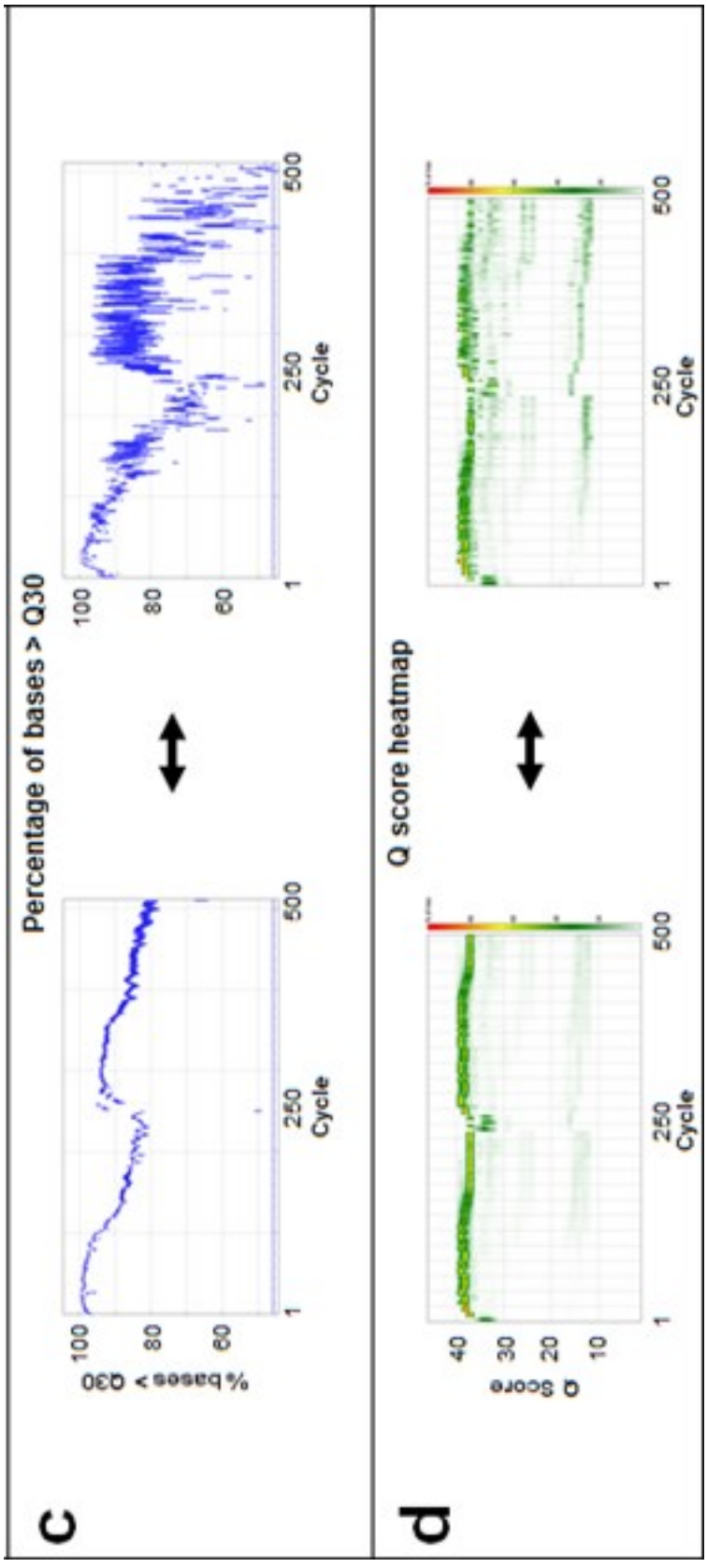

c
Percentage of bases > Q30
% bases > Q30
100
80
60
1
250
500
Cycle
d
Q score heatmap
Q Score
40
30
20
10

Figure 3.5. MiSeq run quality for Run C (our method) and Run D (Earth Microbiome Project method).

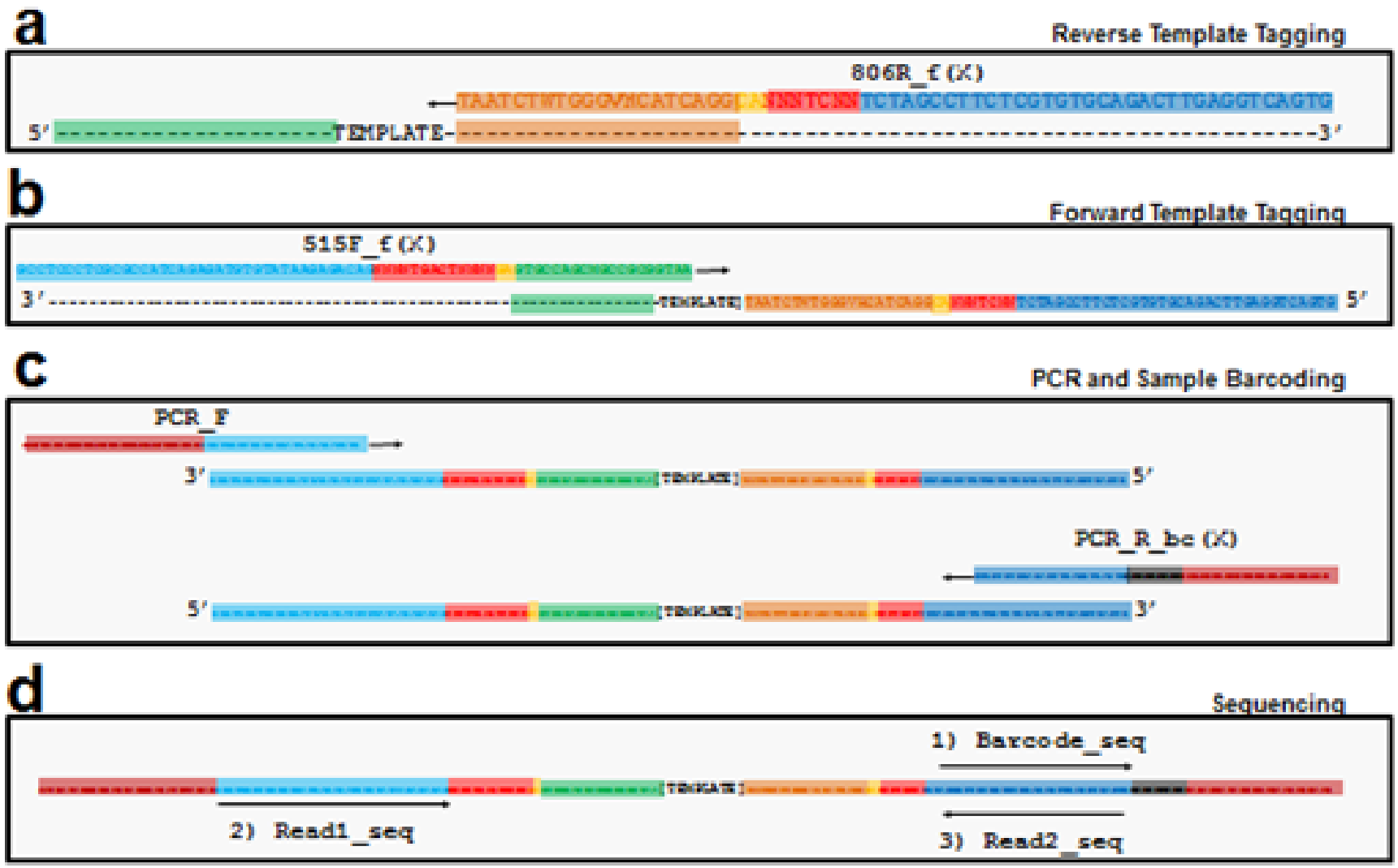
The runs were consecutive, on a machine that had the Illumina May 2013 software upgrade to Real-Time Analysis v1.17.28. The recommended 5% phiX174 spike was not used for either run. Our method (left) and the EMP method (right) were each used in parallel to amplify 16S rRNA from the same set of samples (Materials and Methods). Amplicons from each method were mixed to make two independent libraries.

(a) The EMP library was loaded at a lower cluster density than the library prepared by our method – although this is expected to reduce crowding and improve cluster recognition, significantly fewer clusters passed the machine's quality filter. Of the high quality clusters, the percent of bases above Q30 was higher for the library prepared by our method. Both "nano" runs had more bases above Q30 than Run B used for the majority of analysis, likely the combined consequence of faster cycling due to the "nano" reagent kit, the software upgrade, and the fact that low-quality samples such as blanks were not mixed into the libraries, though they were in Run B.

(b) As predicted from simulation (Figure 3.3b), observed base diversity is much higher for our method, resulting in no base approaching 100% representation in each cycle. In contrast, the EMP method results in much lower diversity.

(c) The percentage of bases above Q30 on a per-cycle basis demonstrates a faster drop in quality for the EMP method for both read 1 (cycle 1-250) and read 2 (cycle 251-500).

(d) Q score heatmaps demonstrating the full distribution of Q scores per cycle.


a
Reverse Template Tagging
806R_f(X)
TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG
5'
TEMPLATE
3'
b
Forward Template Tagging
515F_f(X)
3'
TEMPLATE
5'
c
PCR and Sample Barcoding
PCR_F
3'
5'
PCR_R_bc(X)
5'
3'
d
Sequencing
1) Barcode_seq
2) Read1_seq
3) Read2_seq

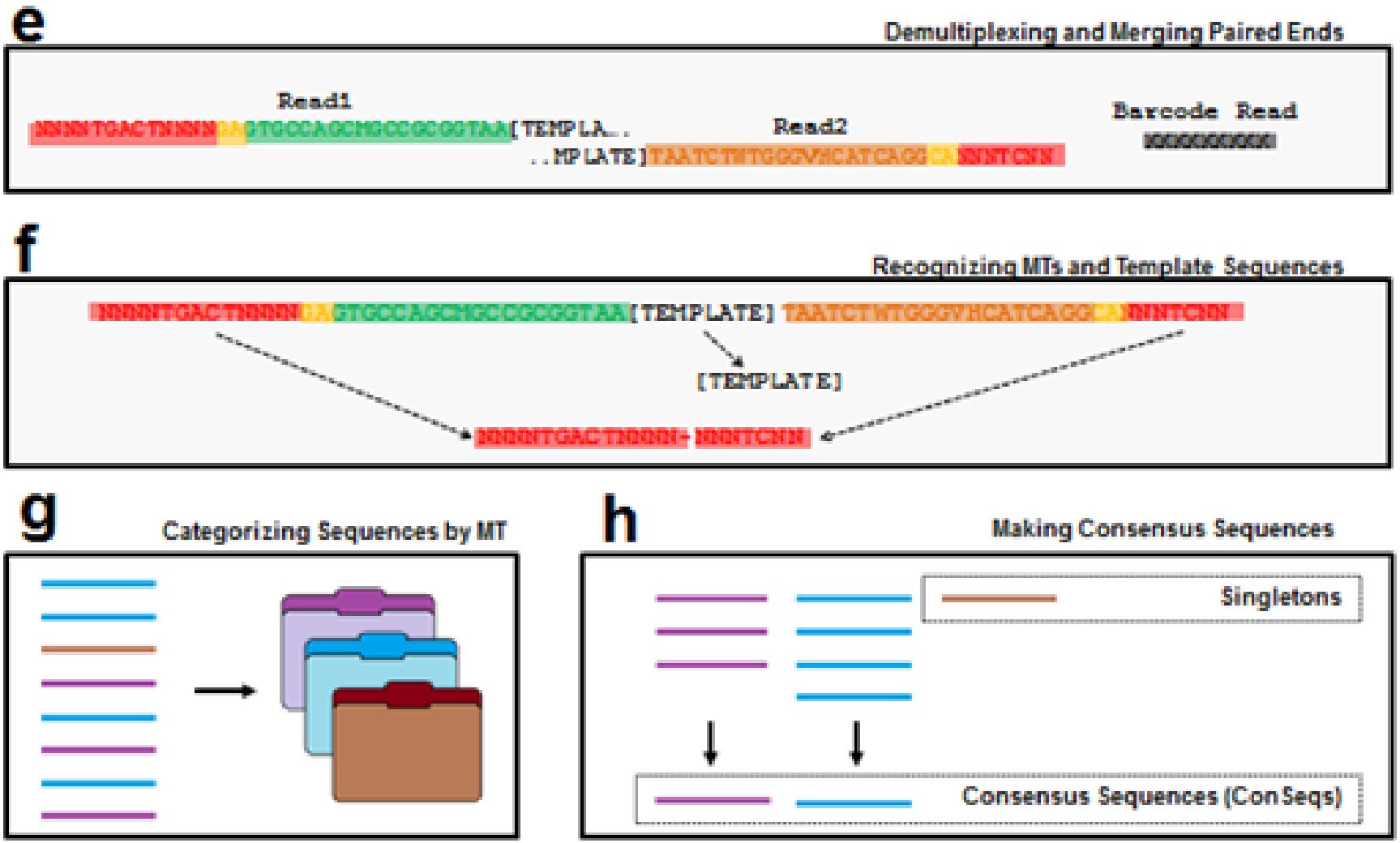

e
Demultiplexing and Merging Paired Ends
Read1
Read2
Barcode Read
f
Recognizing MTs and Template Sequences
[TEMPLATE]
g
Categorizing Sequences by MT
h
Making Consensus Sequences
Singletons
Consensus Sequences (ConSeqs)

Figure 3.6. Template Tagging, PCR, sequencing, and Molecular Tag (MT) processing workflow.

Primer components colored as in Figure 3.2.

(**a**) Template is tagged with reverse MT-FS primers using one extension cycle, and residual primer is removed.

(**b**) The reverse-tagged template is tagged with forward MT-FS primers using one extension cycle, and residual primer is removed.

(**c**) Dual-tagged template is amplified using universal primers that add sample barcodes. Residual primers are removed and samples are quantified and mixed to a final library.

(**d**) Amplicons are sequenced in three reads. First, the 9 bp sample barcodes are read following priming with “Barcode_seq”. The 250 bp forward read is sequenced following priming with “Read1_seq”, and the 250 bp reverse read is sequenced following priming by “Read2_seq”.

(**e**) All sequenced are de-multiplexed based on the “Barcode_seq” read which captures the sample barcode. For each sample, Read1 and Read2 are merged.

(**f**) Regular expressions find all sequences in the set of merged sequences that match the expected patterns, and then extract the MT and template sequence from these pattern-matching sequences.

(**g**) Sequences (colored lines) sharing the same molecular tag sequence (color) are grouped into the same MT category (each colored folder).

(**h**) Sequences in the same MT category are aligned and a consensus sequence is built to represent that MT category. Singleton MT categories are kept in a separate file from consensus sequences.

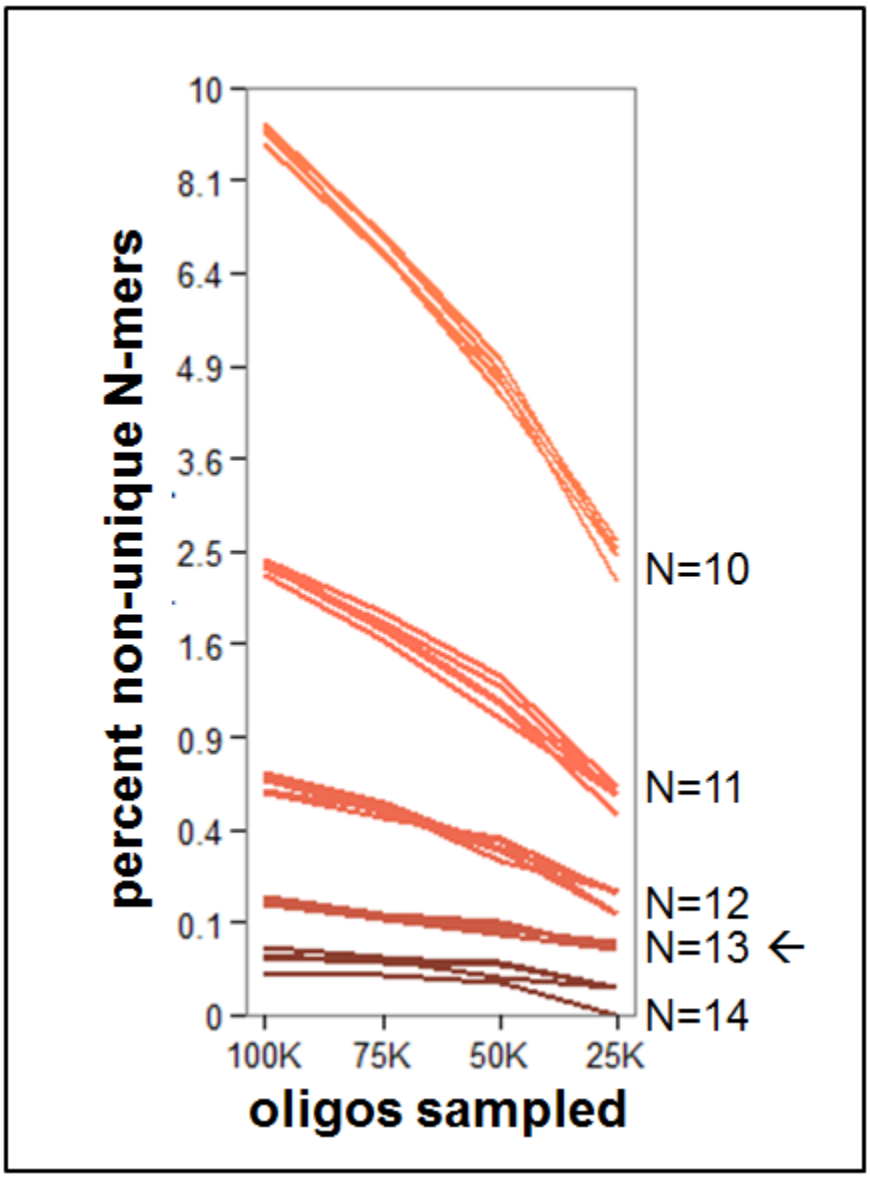

10
8.1
6.4
4.9
3.6
2.5
1.6
0.9
0.4
0.1
0
percent non-unique N-mers
N=10
N=11
N=12
N=13 ←
N=14
100K
75K
50K
25K
oligos sampled

Figure 3.7. A MT of 13 random bases is sufficiently unique.

Monte Carlo simulation at four sampling depths showing the percentage of non-unique oligonucleotide (A, C, T, or G) *N*-mers for *N*'s of 10, 11, 12, 13, and 14. The simulation was repeated 5 times (multiple lines within each hue). A randomer of $N = 13$ (second line from bottom) has about 140 non-unique oligos for every 100,000 sampled (~0.1%), which group into 70 duplicates. In the case of a template-overloaded sample sequenced to a depth of 100,000 reads or greater, these duplicate tags will lead to the unwanted classification of unrelated sequences as originating from the sample template. The consensus sequence made from the multiple sequence alignments will favor the overrepresented MT, often correcting the problem. Furthermore, each multiple sequence alignment can be assigned a quality score based on the average deviation of each sequence in the alignment from the consensus sequence for that alignment. Because multiple sequence alignments made from falsely-grouped independent templates will in general have worse alignment scores, these can be removed from the dataset by thresholding the worst alignments. Choice of random-mer length must be a balance between uniqueness on the one hand, versus costs in terms of sequence length and oligo chaos caused by longer lengths of *N*. It is more important to minimize non-unique *N*-mers than attempt to eliminate them; samples for which deep sequencing is needed can be multiplexed over several barcodes to increase depth, allowing unique molecular tagging without increasing random-mer length.

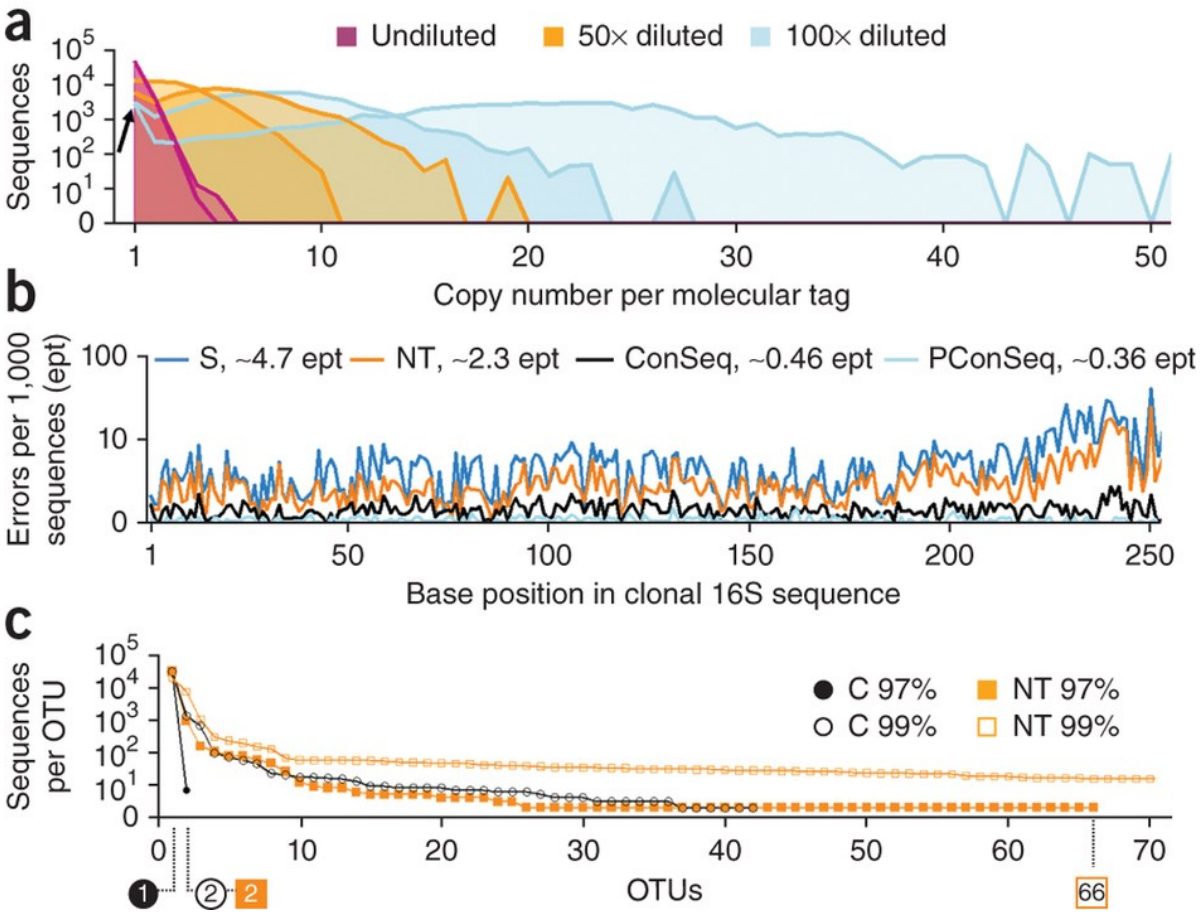

a
Undiluted
50× diluted
100× diluted
Sequences
Copy number per molecular tag
b
S, ~4.7 ept
NT, ~2.3 ept
ConSeq, ~0.46 ept
PConSeq, ~0.36 ept
Errors per 1,000 sequences (ept)
Base position in clonal 16S sequence
c
C 97%
NT 97%
C 99%
NT 99%
Sequences per OTU
OTUs
66

Figure 3.8. Molecular tagging reduces sequence error for a clonal template.

Molecular tagging reduces sequence error for a clonal template.

(a) Diluting template increases the coverage within each MT. Shown are two replicates each (overlaid in the same color) of undiluted, 50× diluted and 100× diluted clonal 16S template. All six samples were rarefied to 40,000 sequences, and the number of sequences collapsed into each MT was graphed as a density distribution for each sample. We noted more singleton MTs than expected by a unimodal Poisson distribution for the diluted samples (arrow).

(b) Per-base error rates per 1,000 sequences were measured in pooled data from the 50× and 100× diluted template samples. We compared no-MT sequences (NT); ConSeqs from two or more sequences with identical MTs (ConSeq); perfect ConSeqs, for which all sequences in the alignments of three or more sequences were identical (PConSeq); and singleton MTs (S). Mean error per thousand (ept) for each MT treatment is shown in the color key.

(c) 30,000 ConSeqs (C) or untreated sequences (NT) were clustered into OTUs at both 97% and 99% identity thresholds. Rank-abundance curves demonstrate the number of sequences per OTU. The position of the colored boxes and circles below the x axis, and the numbers in each, show the number of ranked OTUs necessary to represent 95% of the sequences for each condition.

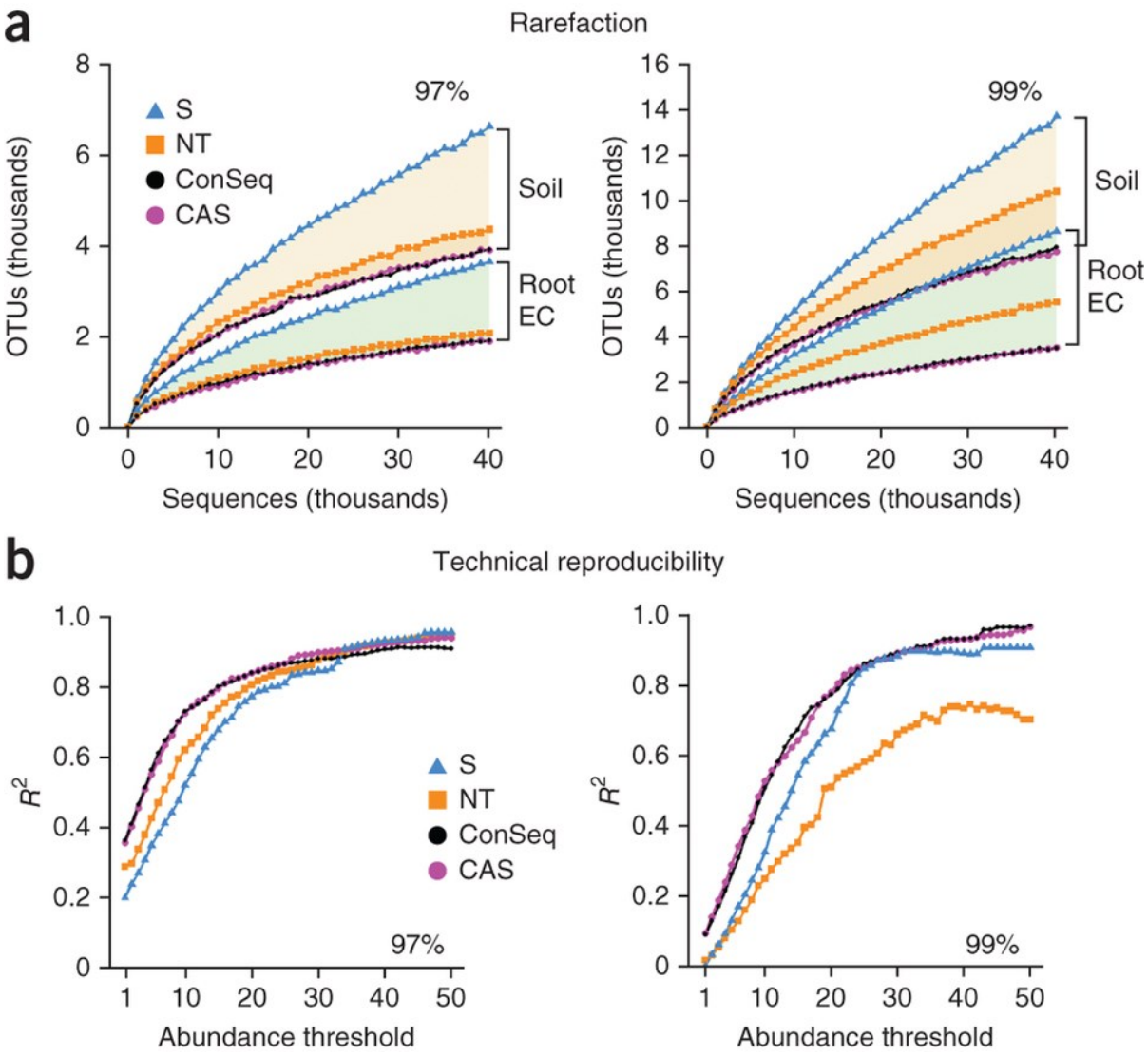

a
Rarefaction
97%
99%
S
NT
ConSeq
CAS
Soil
Root
EC
OTUs (thousands)
Sequences (thousands)
b
Technical reproducibility
R^2
Abundance threshold

Figure 3.9. Molecular tagging lowers estimates of alpha diversity and improves technical reproducibility.

(a) 16S sequences with no MTs (NT), ConSeqs from two or more sequences with identical MTs (ConSeq), singleton MTs (S) and a combination of ConSeqs and a downsampled fraction of the residual singletons (CAS) were rarefied before OTU formation and clustered independently into OTUs at 97% (left) and 99% (right) identity. Bacterial reads from root EC or soil samples were pooled, producing a soil pool and a root EC pool per MT treatment at each identity threshold. These pools were rarefied at intervals of 1,000 sequences, and the number of OTUs observed at each depth were plotted. Beige shading connects soil samples; green shading connects EC root samples.

(b) Progressive drop-out analysis displaying the coefficient of determination (R^2) of 24 intra-run technical replicates as OTUs with low read numbers are discarded. OTU tables are the same as in **a**, with the exception that plastid and mitochondrial OTUs were not removed.

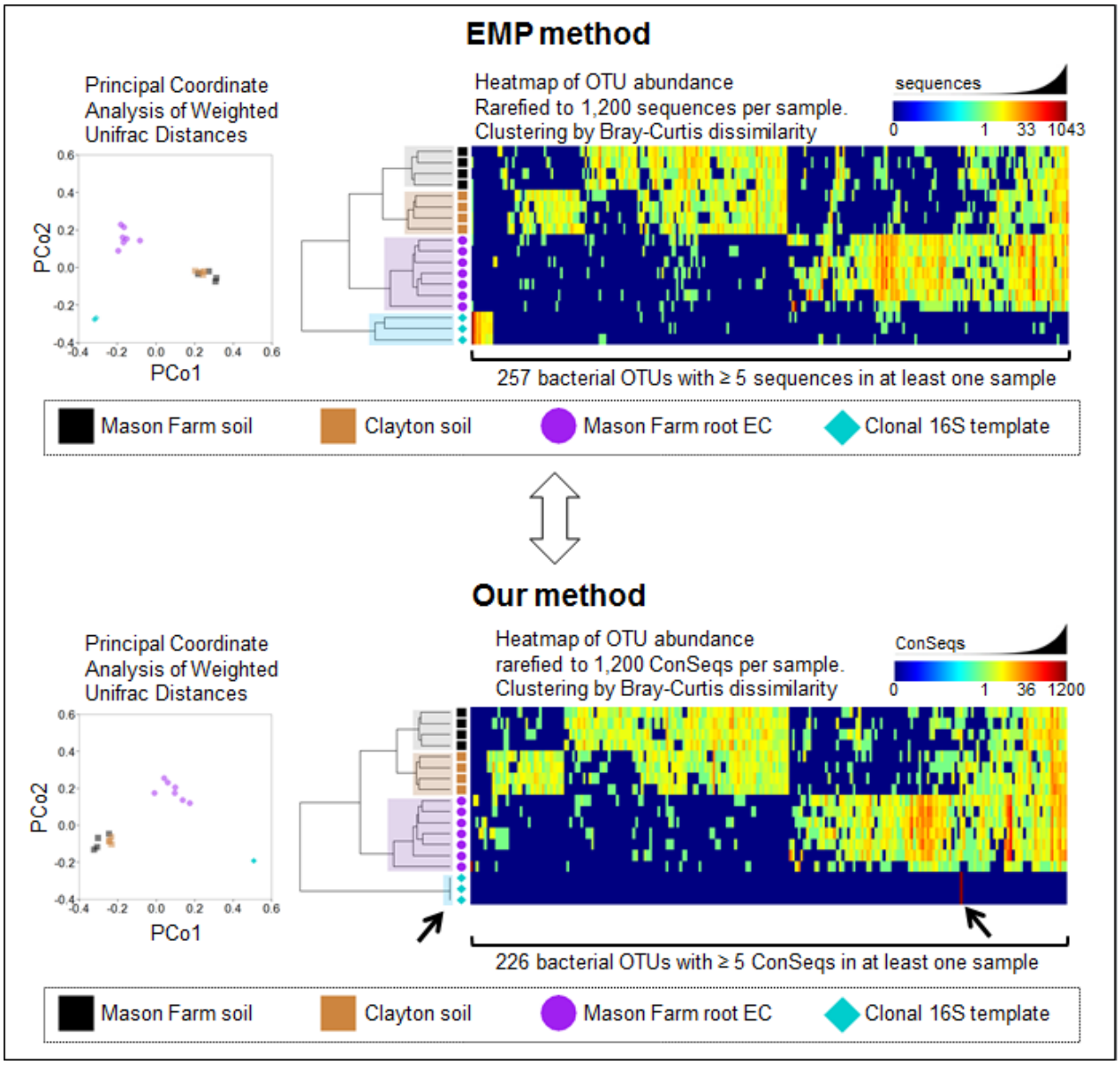

EMP method
Principal Coordinate Analysis of Weighted Unifrac Distances
PCo2
PCo1
Heatmap of OTU abundance
Rarefied to 1,200 sequences per sample.
Clustering by Bray-Curtis dissimilarity
sequences
0 1 33 1043
257 bacterial OTUs with ≥ 5 sequences in at least one sample
Mason Farm soil
Clayton soil
Mason Farm root EC
Clonal 16S template
Our method
Principal Coordinate Analysis of Weighted Unifrac Distances
PCo2
PCo1
Heatmap of OTU abundance
rarefied to 1,200 ConSeqs per sample.
Clustering by Bray-Curtis dissimilarity
ConSeqs
0 1 36 1200
226 bacterial OTUs with ≥ 5 ConSeqs in at least one sample
Mason Farm soil
Clayton soil
Mason Farm root EC
Clonal 16S template

Figure 3.10. Beta diversity conclusions from our method vs. the Earth Microbiome Project (EMP) method.

Four independent Mason Farm soil samples (back squares), four indepdent Clayton soil samples (brown squares), seven Mason Form root endopyte compartment samples from separate plants (puple circles) and 3 technical PCR replicates of a cloned 16S template were each phylotyped using the EMP method (top) or our method (bottom) (Materials and Methods). OTUs were formed at 97% identity and all samples were rarefed to 1,200 sequences or 1,200 ConSeqs. Principal coordinates analysis based on weighted unifrac distances (left) demonstrates that for both methods, the first two principal coordinates capture a similar separation of sample types. For heatmap visualization, the OTUs were thresholded such that only those OTUs contaning at least 5 sequences or ConSeqs in at least one sample are displayed. Heatmap rows and columns are ordered based on unsupervised clustering by Bray-Curtis dissimilarity. The heirarchical clustering results in the same separation of sample types as the Unifrac ordination for both methods, demonstrating that the same major beta-diversity conclusions can be reached with both methods. However, the ConSeqs from our method represent less noise, clearly evident from the single OTU formed for the clonal 16S template. In contrast, the EMP method produed produce several low-abundance OTUs from the clonal template, and 31 more OTUs overall using the same thresholding parameters (x-axis of heatmap, Materials and Methods).

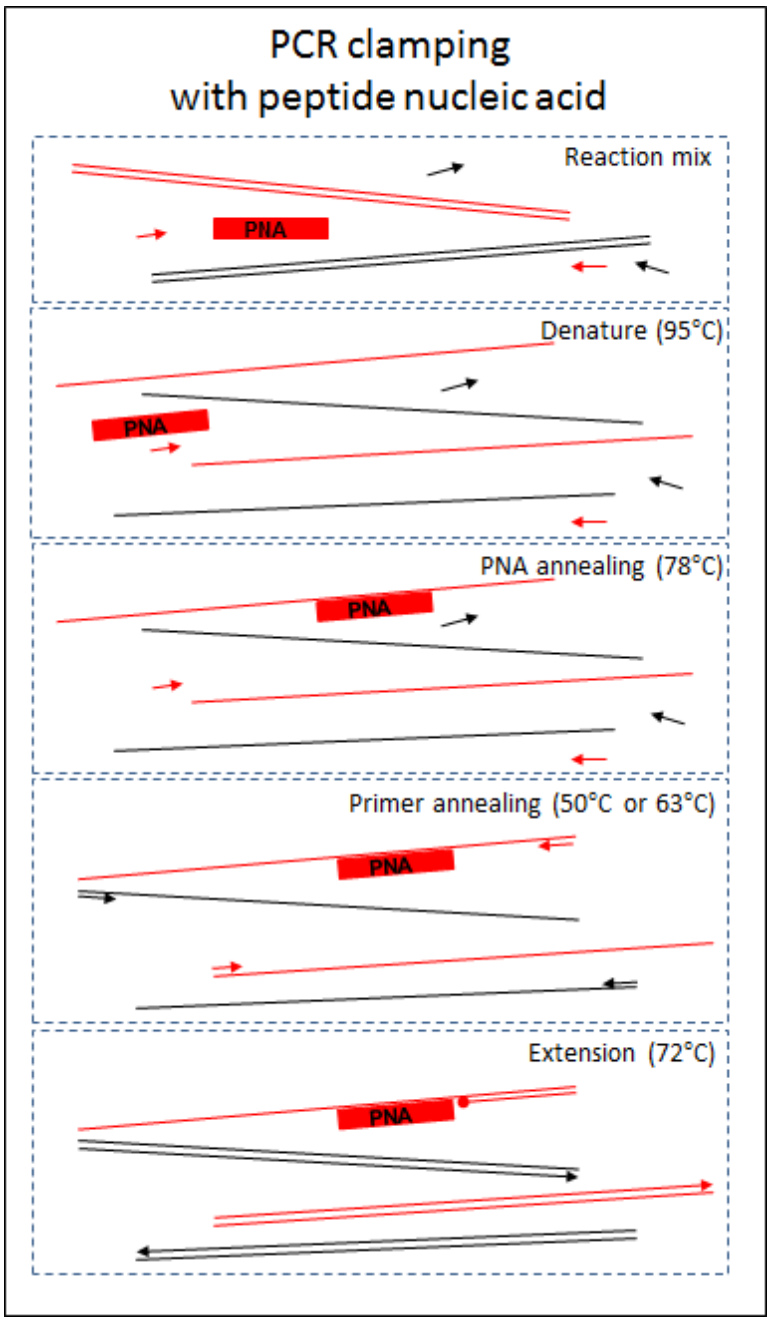


Figure 3.11. PNA schematic

PNA functions as an additive in the PCR reaction mix (top). After denaturation, PNA anneals specifically to templates via base pairing. As long as the PNA has a higher melting temperature than the primers, it anneals to template prior to the primers (middle). Depending on design, PNA either directly blocks primer annealing or blocks extension of the nascent strand.

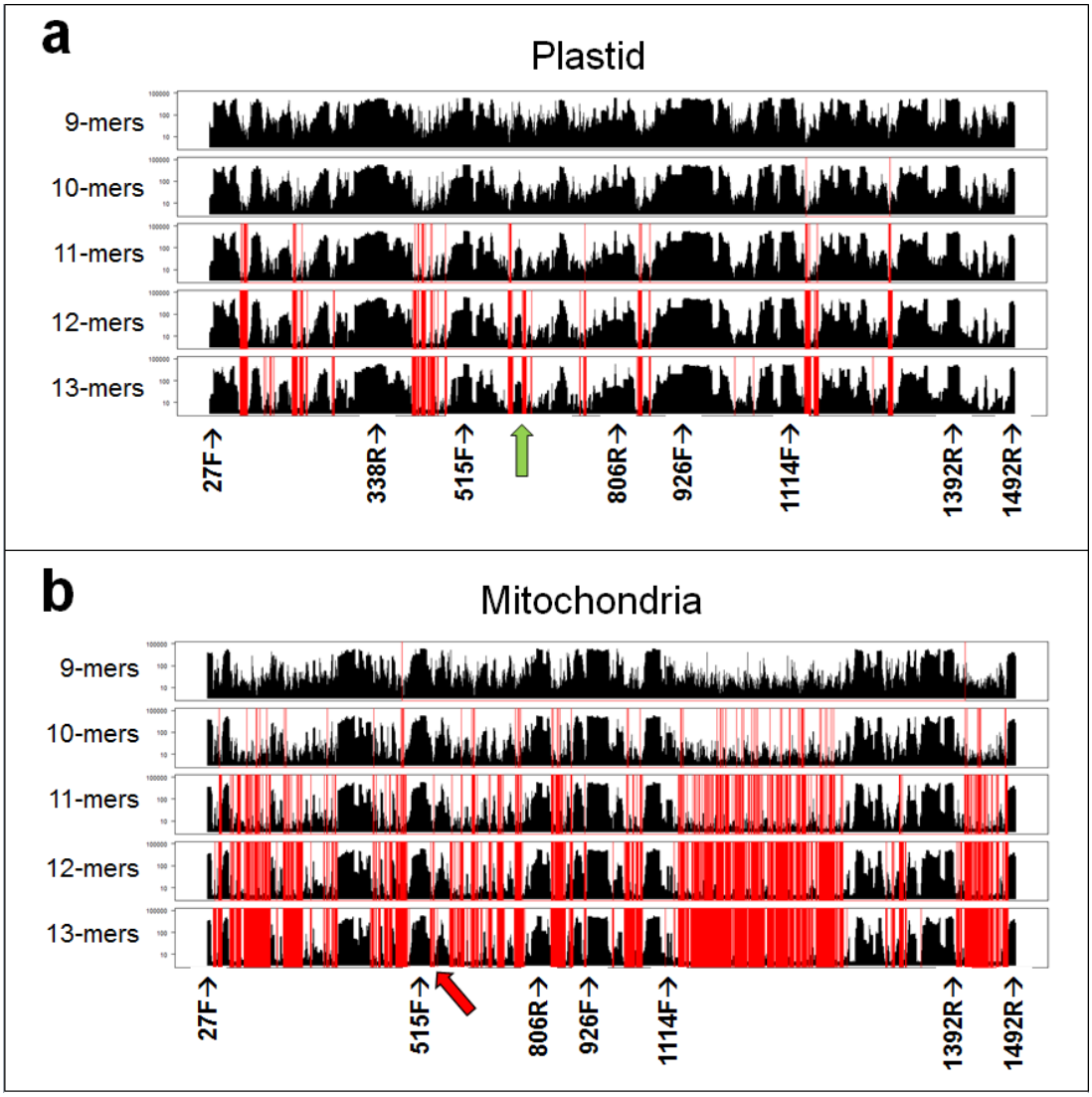


Figure 3.12. Exhaustive search for PNA oligo candidates.

(**a**) The full length chloroplast 16S rRNA sequence was split *in silico* into all possible 9-mers, 10-mers, 11-mers, 12-mers, and 13-mers. Each fragment was searched against the full length sequence for all sequences in the Greengenes 97% representatives microbial database, and the number of matches was graphed (black; log scale). Fragments of each length matching no sequences are marked with a red vertical line; these represent the best candidates for PCR clamping. The location of common 16S primers is shown beneath each histogram, and the location of the "pPNA" used in this study is shown with a green arrow.

(**b**) As above, but for the mitochondrial 16S sequence. The location of the "mPNA" used in this study is shown with the red arrow.

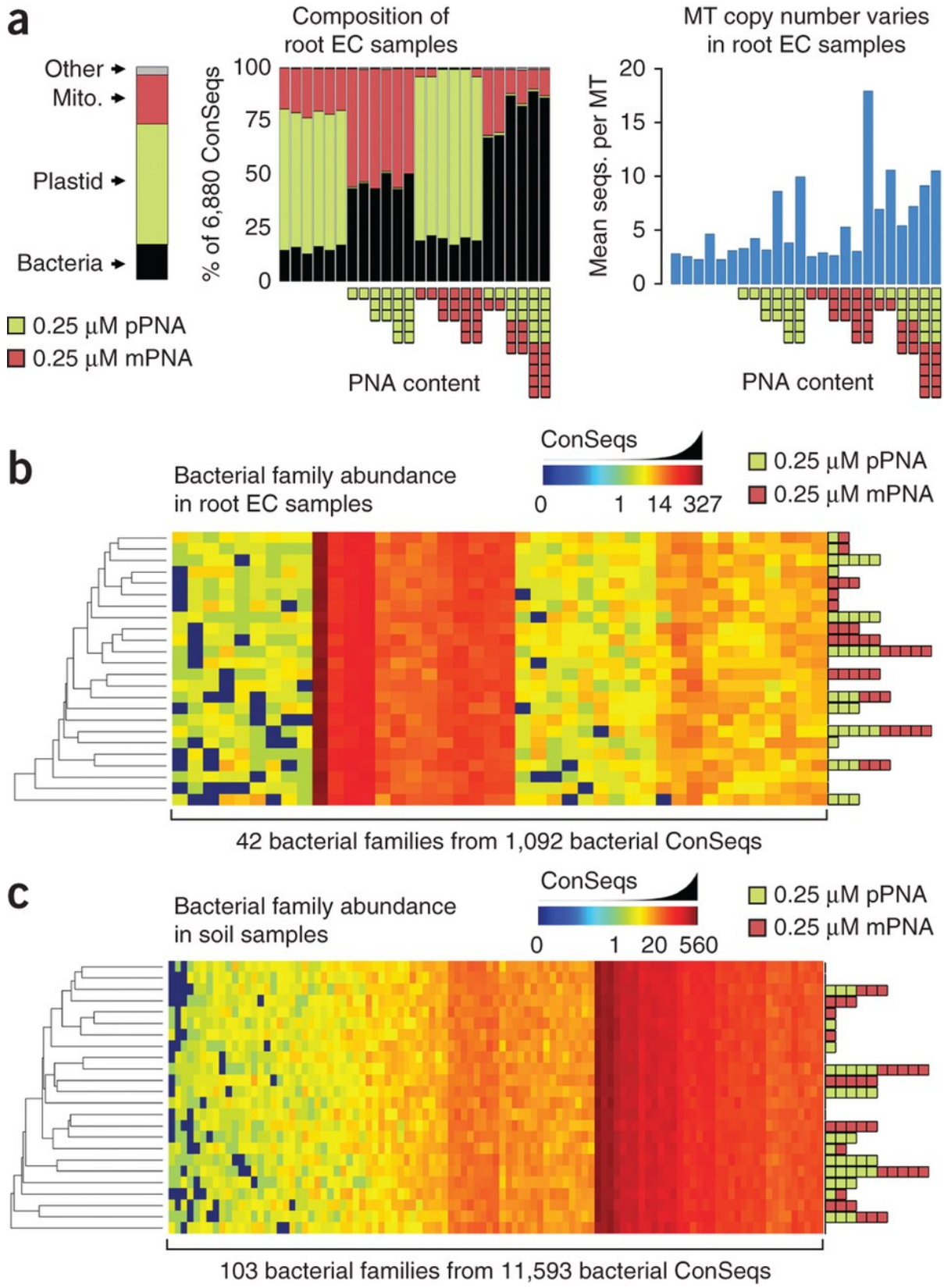

a
Composition of root EC samples
MT copy number varies in root EC samples
Other
Mito.
Plastid
Bacteria
% of 6,880 ConSeqs
100
75
50
25
0
Mean seqs. per MT
20
15
10
5
0
0.25 µM pPNA
0.25 µM mPNA
PNA content
PNA content
b
Bacterial family abundance in root EC samples
ConSeqs
0
1
14
327
0.25 µM pPNA
0.25 µM mPNA
42 bacterial families from 1,092 bacterial ConSeqs
c
Bacterial family abundance in soil samples
ConSeqs
0
1
20
560
0.25 µM pPNA
0.25 µM mPNA
103 bacterial families from 11,593 bacterial ConSeqs

Figure 3.13. PNA specifically blocks amplification of contaminant sequences.

(a) The stacked bar chart legend (left) schematizes the relative abundance of ConSeqs classified as bacteria, plastid, mitochondria (Mito.) and other. PNA was titrated into PCR reactions of root EC DNA. Each green or red block below the histogram represents 0.25 μM of pPNA or mPNA in the final reaction, respectively. The sequence copy number per MT, and thus the mean number of sequences (seqs.) in each alignment used to compute the ConSeqs (blue bars, right), is determined by the sequencing depth and the amplifiable template concentration.

(b) Root EC samples (rows) to which varying titrations of PNA had been applied (colored blocks) were clustered on the basis of the abundance of bacterial families (columns; family IDs not shown). The relative abundance of each bacterial family is displayed as a heat map.

(c) Clustering and abundance as in b but with soil samples. Note that there is no clustering by PNA treatment in **b** and **c**.

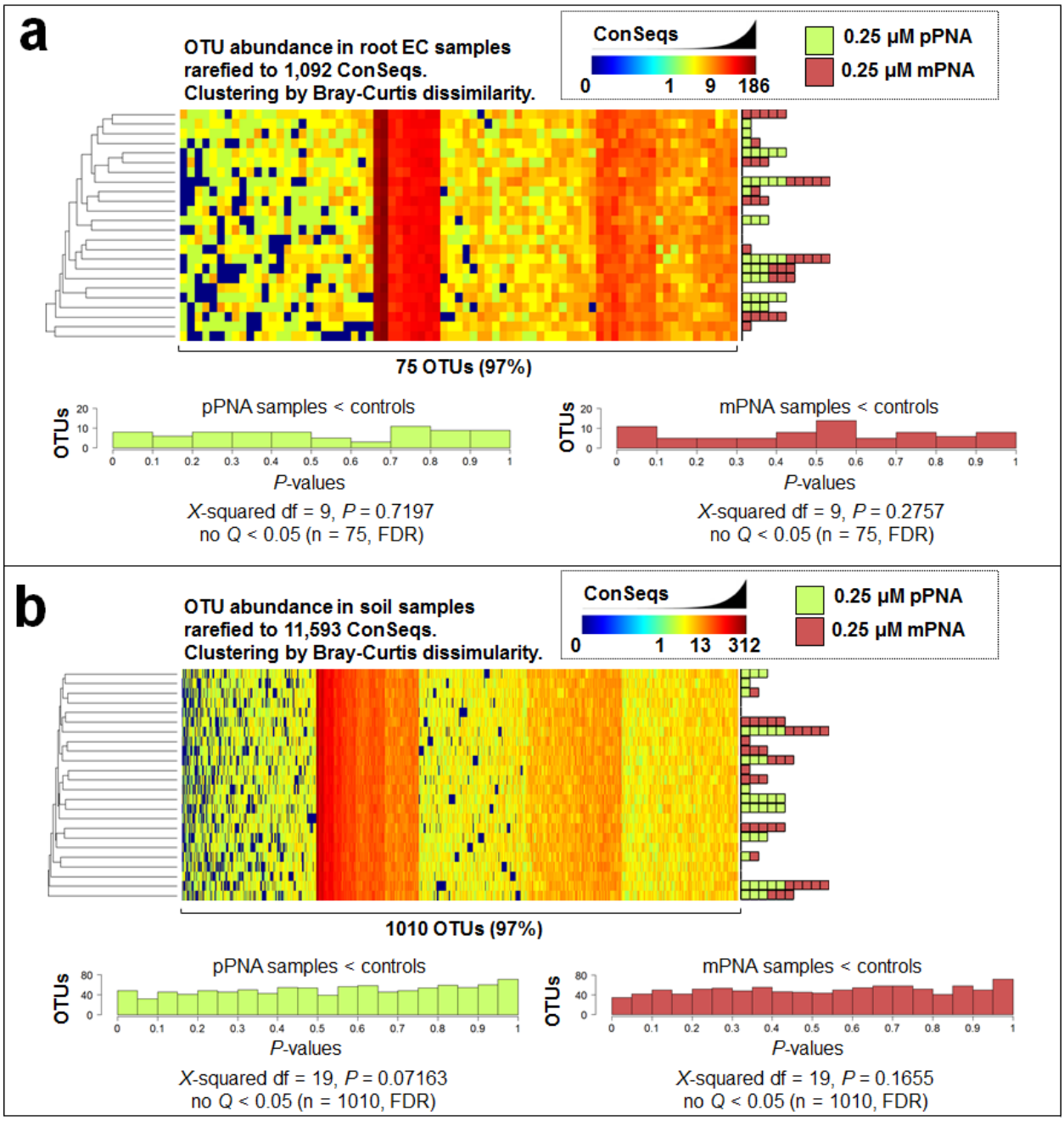

a
OTU abundance in root EC samples
rarefied to 1,092 ConSeqs.
Clustering by Bray-Curtis dissimilarity.
ConSeqs
0 1 9 186
0.25 µM pPNA
0.25 µM mPNA
75 OTUs (97%)
pPNA samples < controls
OTUs
P-values
X-squared df = 9, P = 0.7197
no Q < 0.05 (n = 75, FDR)
mPNA samples < controls
OTUs
P-values
X-squared df = 9, P = 0.2757
no Q < 0.05 (n = 75, FDR)
b
OTU abundance in soil samples
rarefied to 11,593 ConSeqs.
Clustering by Bray-Curtis dissimilarity.
ConSeqs
0 1 13 312
0.25 µM pPNA
0.25 µM mPNA
1010 OTUs (97%)
pPNA samples < controls
OTUs
P-values
X-squared df = 19, P = 0.07163
no Q < 0.05 (n = 1010, FDR)
mPNA samples < controls
OTUs
P-values
X-squared df = 19, P = 0.1655
no Q < 0.05 (n = 1010, FDR)

Figure 3.14. No bacterial OTU abundances are affected by pPNA or mPNA.

(**a**) Root EC samples were clustered by the abundance of the 75 bacterial OTUs with ≥5 ConSeqs in at least one of the 24 samples. The heatmap shows the relative abundance of each OTU (columns) for each of the samples (rows) with the PNA doses shown (colored blocks). For each OTU, the 12 samples containing pPNA were tested for lower abundance than the 12 samples containing no PNA or only mPNA (left; green). Similarly, the 12 samples containing mPNA were tested for lower abundance than the 12 samples containing no PNA or only pPNA (right; red). *P*-values were obtained with a permutation test on the means using 10,000 permutations, and the *P*-value distribution was plotted across 10 bins (histograms). *P*-values were corrected for multiple testing with the FDR method; no OTUs were found significant. Each *P*-value distribution was shown not to deviate from the null flat distribution with a Chi-squared test (*P*-values for Chi-squared below histograms).

(**b**) Same as in **a**, but for the 1,010 OTUs in soil samples with ≥5 ConSeqs in at least one of the 24 samples. Owing to the much greater number of OTUs the *P*-value distributions were plotted across 20 bins (histograms). The Chi-squared *P*-values, both for pPNA and mPNA comparisons, supported the null hypothesis of a flat distribution. *P*-values were corrected for multiple testing with the FDR method; limited OTUs in soil samples had significant *Q*-values (bold, red). Consistent with these statistics, there is no clustering (based on Bray-Curtis dissimilarity and group average linkage) by PNA treatment.

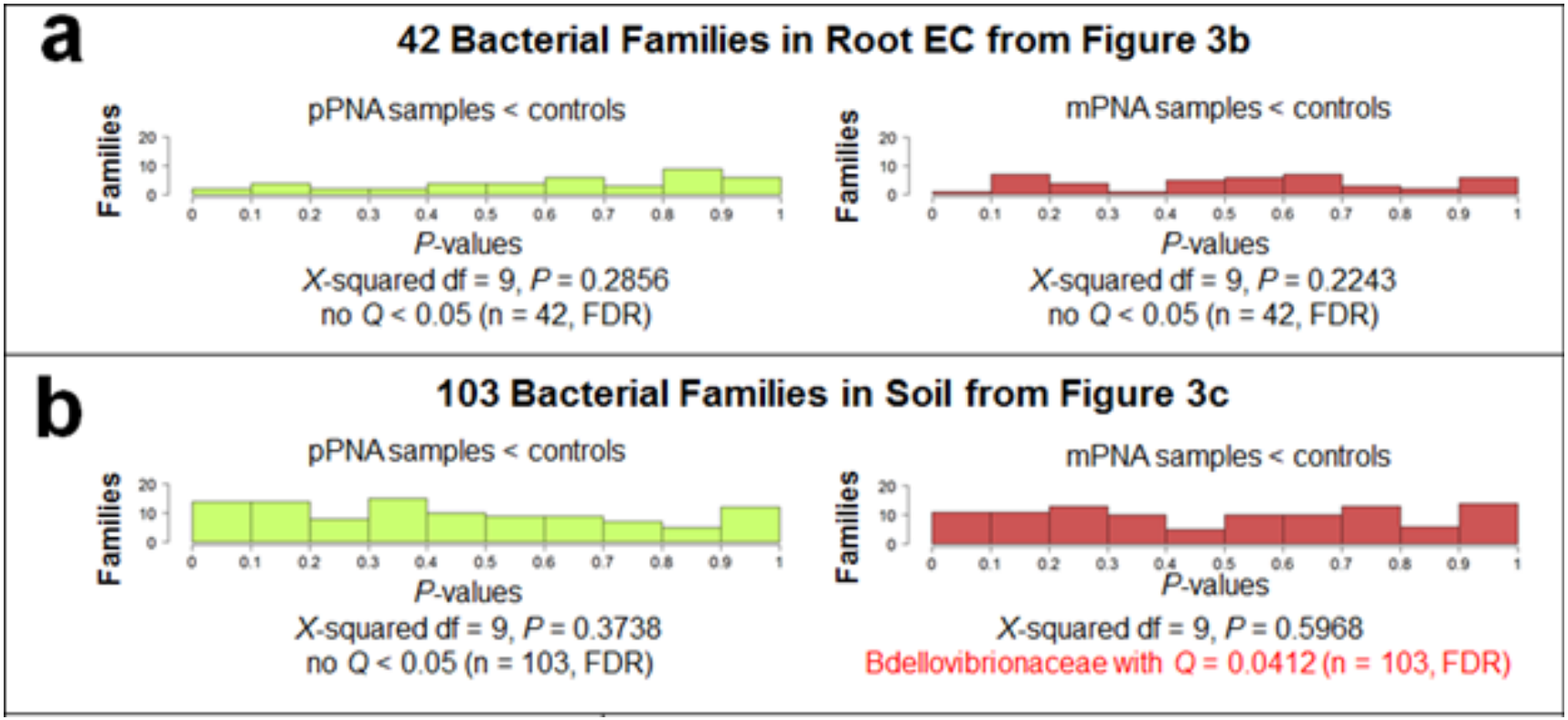
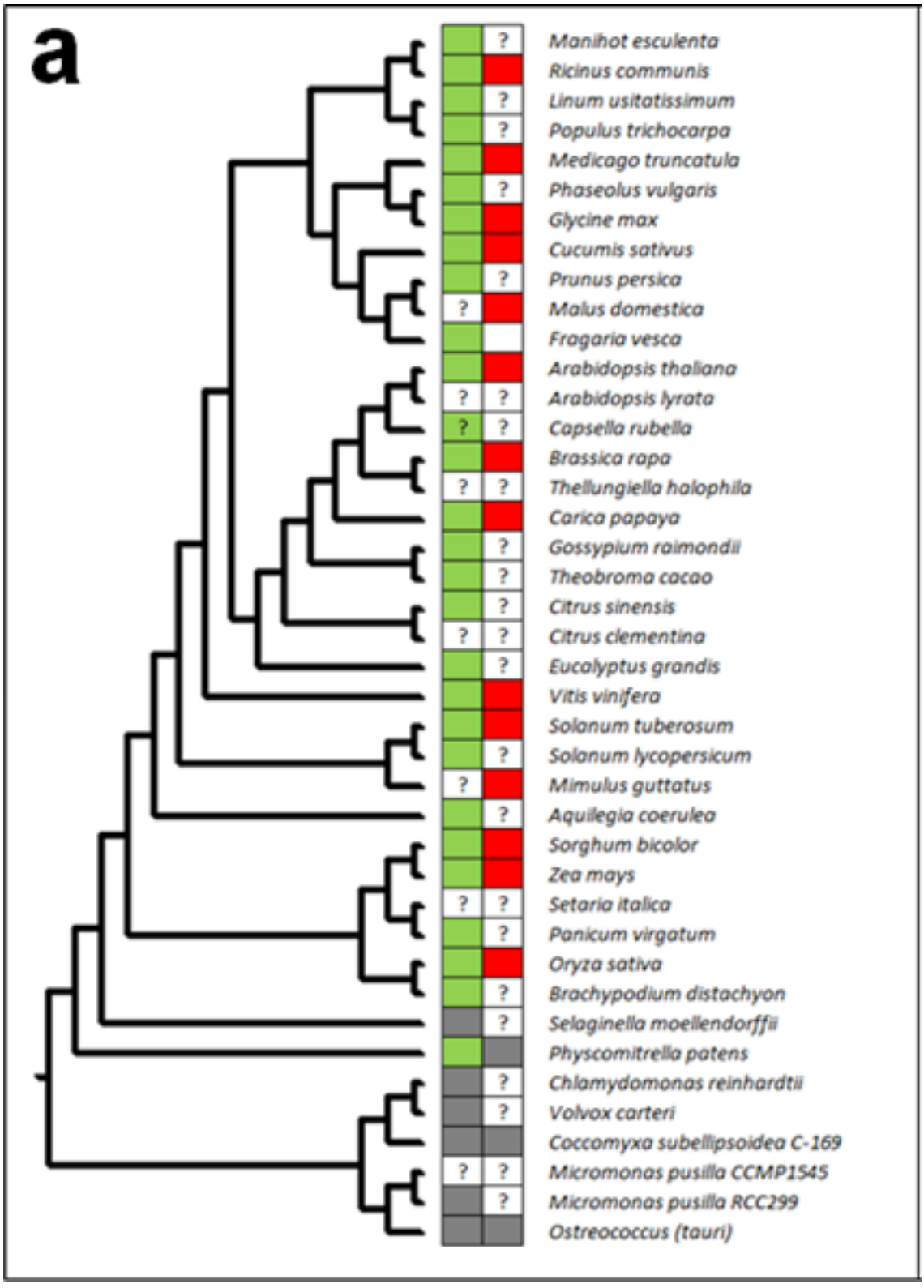

a
42 Bacterial Families in Root EC from Figure 3b
pPNA samples < controls
Families
P-values
X-squared df = 9, P = 0.2856
no Q < 0.05 (n = 42, FDR)
mPNA samples < controls
Families
P-values
X-squared df = 9, P = 0.2243
no Q < 0.05 (n = 42, FDR)
b
103 Bacterial Families in Soil from Figure 3c
pPNA samples < controls
Families
P-values
X-squared df = 9, P = 0.3738
no Q < 0.05 (n = 103, FDR)
mPNA samples < controls
Families
P-values
X-squared df = 9, P = 0.5968
Bdellovibrionaceae with Q = 0.0412 (n = 103, FDR)

Figure 3.15. No bacterial family abundances are affected by pPNA or mPNA.

(**a**) The abundance of each bacterial family with ≥5 ConSeqs in at least one of the 24 samples in different PNA treatments (**Figure 3.13b**) was compared for root EC. For each bacterial family, the 12 samples containing pPNA were tested for lower abundance than the 12 samples containing no PNA or only mPNA (left; green). Similarly, the 12 samples containing mPNA were tested for lower abundance than the 12 samples containing no PNA or only pPNA (right; red). *P*-values were obtained with a permutation test on the means using 10,000 permutations, and the *P*-value distribution was plotted across 10 bins (histograms). The *P*-values were corrected for multiple testing with the FDR method; none of the resulting corrected *Q*-values were significant. Each *P*-value distribution was shown not to deviate from the null flat distribution with a Chi-squared test (*P*-values for Chi-squared below histograms).

(**b**) Same as **a**, but analyzing bacterial families in soil (Figure 3.13c). One *Q*-value for the mPNA test, corresponding to the family Bdellovibrionaceae, was significant.


a
Manihot esculenta
Ricinus communis
Linum usitatissimum
Populus trichocarpa
Medicago truncatula
Phaseolus vulgaris
Glycine max
Cucumis sativus
Prunus persica
Malus domestica
Fragaria vesca
Arabidopsis thaliana
Arabidopsis lyrata
Capsella rubella
Brassica rapa
Thellungiella halophila
Carica papaya
Gossypium raimondii
Theobroma cacao
Citrus sinensis
Citrus clementina
Eucalyptus grandis
Vitis vinifera
Solanum tuberosum
Solanum lycopersicum
Mimulus guttatus
Aquilegia coerulea
Sorghum bicolor
Zea mays
Setaria italica
Panicum virgatum
Oryza sativa
Brachypodium distachyon
Selaginella moellendorffii
Physcomitrella patens
Chlamydomonas reinhardtii
Volvox carteri
Coccomyxa subellipsoidea C-169
Micromonas pusilla CCMP1545
Micromonas pusilla RCC299
Ostreococcus (tauri)

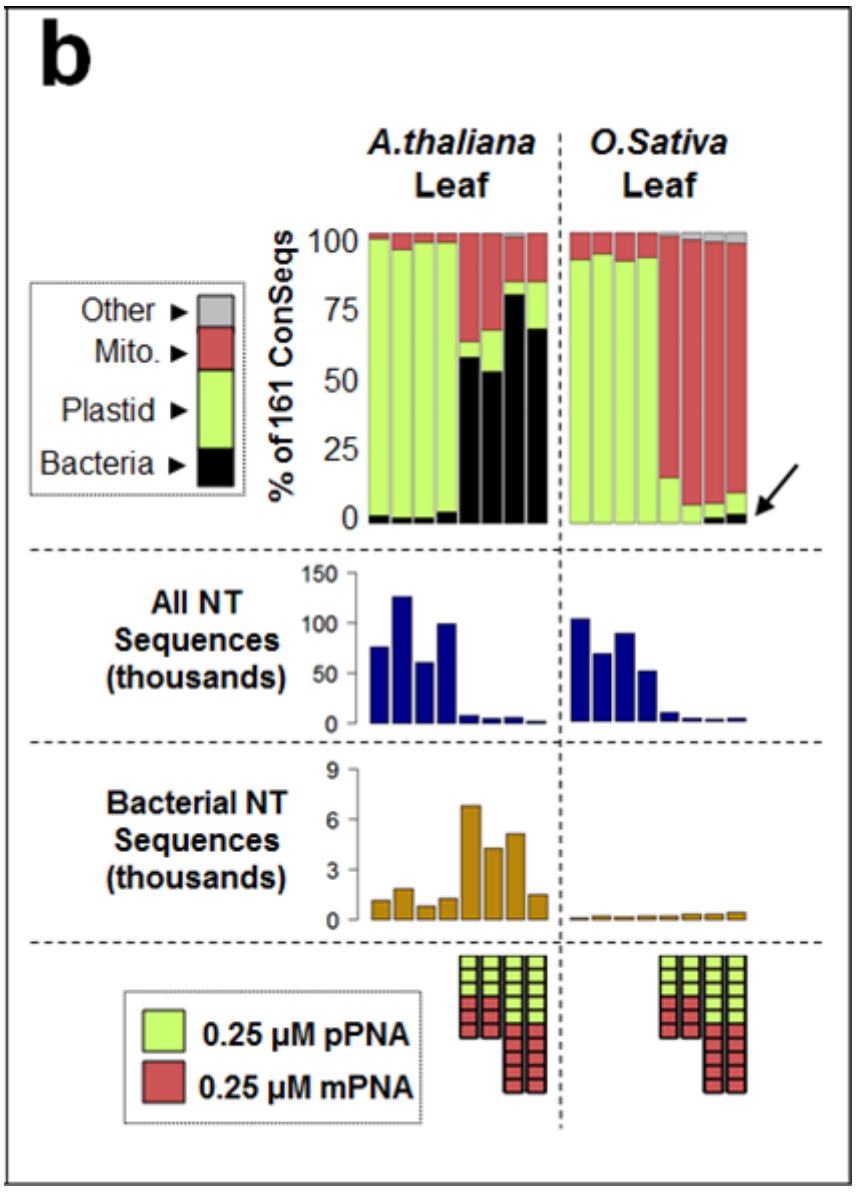


Figure 3.16. Diverse plant species for which the PNAs in this study should block organelle V4 16S amplification.

(**a**) Diverse plant species for which the PNAs in this study should block organelle 16S amplification based on an exact sequence match. Phylogenetic tree and choice of plant taxa adapted from Phytozome v9.1 (http://www.phytozome.net/). Branch lengths are not meaningful. Plastid and mitochondrial organelle sequences for each plant in the phylogeny, or a relative in the same genus if the Phytozome species was not available, were collected from NCBI GenBank. The pPNA and mPNA sequences were queried against all collected plastid and mitochondrial sequences, respectively. Green squares represent exact matches

of the pPNA to the plastid sequence; red squares represent exact matches of the mPNA to the mitochondrial sequence; grey squares represent a mismatch; white squares filled with "?" mean that the organelle sequence was not publicly available.

(**b**) Leaf samples from *A.thaliana* (left) and *O.sativa* (right) were amplified with or without a mix of both pPNA and mPNA. Despite the extreme host contamination present in DNA from ground leaves (98.3% and 99;8% for *A.thaliana* and *O.sativa* respectively), addition of PNA increased the relative abundance of bacterial reads (top). Although the effect appears modest for *O.sativa*, the use of 1.25 μM of both PNAs (arrow) represents a more than 20-fold increase in detectable templates. As with *A.thaliana* leaves, PNAs blocked the amplification of the majority of contaminant, and hence, template molecules of *O.sativa*, resulting in less sequenceable material (dark blue bars). However, more total bacterial sequences were nonetheless recovered (brown bars). These results are consistent with the PNAs functioning to block chloroplast and mitochondria, but not bacteria, in *O.sativa.*

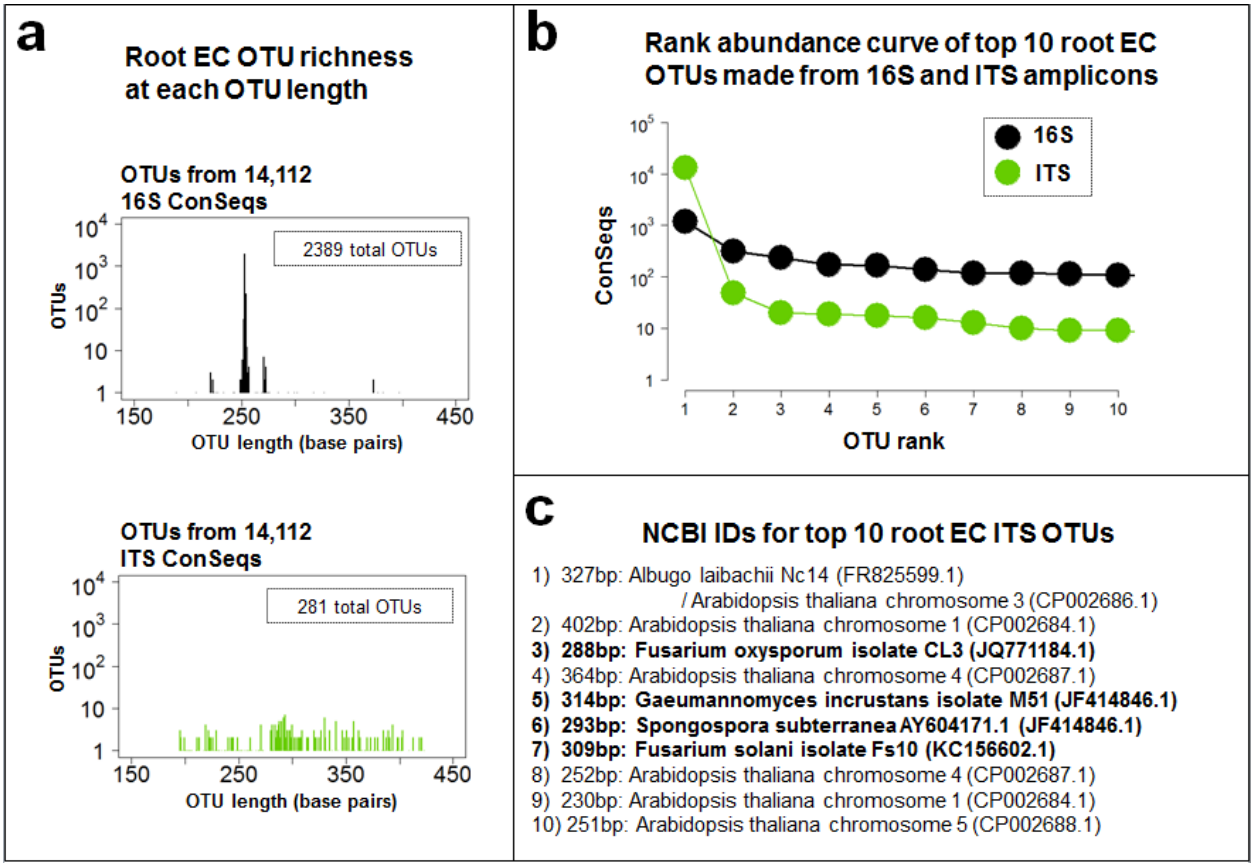


Figure 3.17. Universal PCR primers can be used to amplify and barcode other tagged templates.

(**a**) Root EC DNA was tagged with either V4 16S MT-FS primers or ITS2 MT primers. Tagged template was amplified with universal PCR primers, sequenced, and MTs were used to form ConSeqs. For 16S (top, black) and ITS (bottom, green), the OTUs present among 14,112 ConSeqs were classified by their sequence length (x-axis), and the number of OTUs present at each length was plotted (y-axis). The total number of OTUs for each amplicon is inlaid in each plot. Although there were more V4 16S OTUs, the distribution of amplicon lengths is much narrower than for ITS.

(**b**) The OTUs of 16S ConSeqs (black) and ITS ConSeqs (green) were ranked by their relative abundance and the number of sequences (log y-axis) is shown for the 10 most-abundant OTUs (x-axis). (**c**) The ITS OTUs shown in **b** were queried against the NCBI database using BLAST and the OTU length in base pairs and the best-scoring hit is shown. Several Arabidopsis OTUs demonstrate host contamination, but other eukaryotic and fungal OTUs are clearly present.

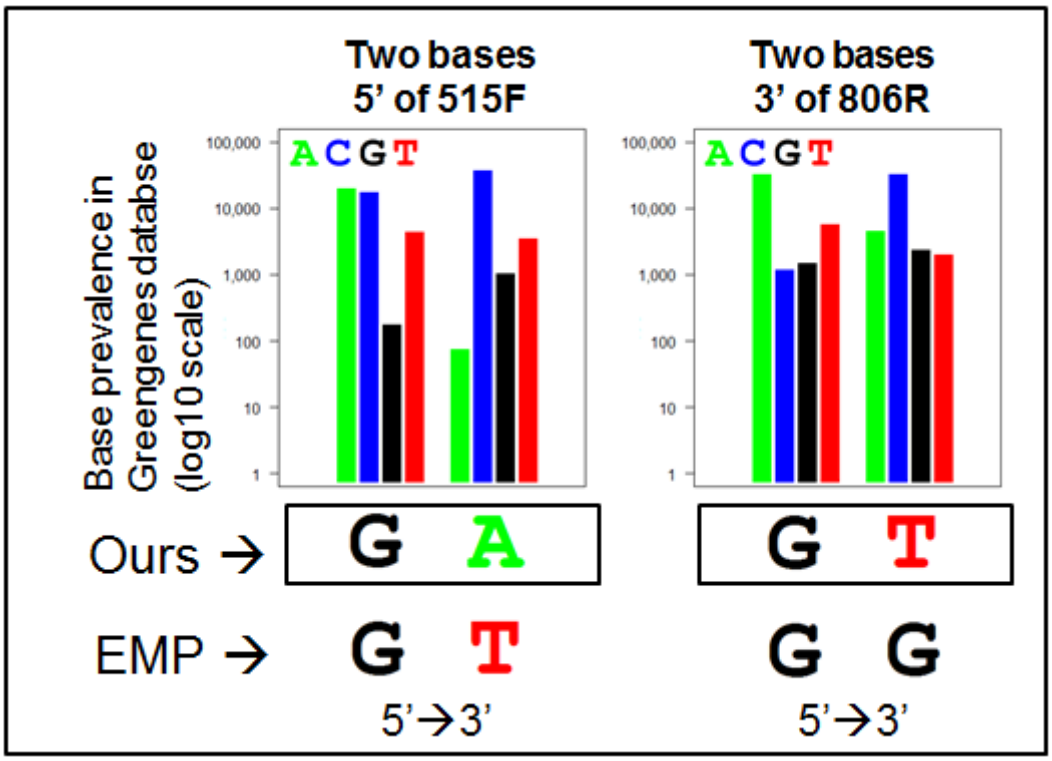


Figure 3.18. Primer linkers.

Our linkers differ from those used by the Earth Microbiome Project (Caporaso et al. 2012). Ideal linkers should lack identity to the majority of microbial sequences in order to buffer the other elements of the template-tagging primer from the template. Our choices are equally or more divergent from sequences in the Greengenes database than are the EMP primers.

REFERENCES

Benson AK, Kelly SA, Legge R, Ma F, Low SJ et al. (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proceedings of the National Academy of Sciences 107(44): 18933-18938.

Bulgarelli D, Rott M, Schlaeppi K, Ver Loren van Themaat E, Ahmadinejad N et al. (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature 488(7409): 91-95.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD et al. (2010) QIIME allows analysis of high-throughput community sequencing data. Nat Meth 7(5): 335-336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J et al. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6(8): 1621-1624.

Chakravorty S, Helb D, Burday M, Connell N, Alland D (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. Journal of Microbiological Methods 69(2): 330-339.

Chow S, Suzuki S, Matsunaga T, Lavery S, Jeffs A et al. (2011) Investigation on Natural Diets of Larval Marine Animals Using Peptide Nucleic Acid-Directed Polymerase Chain Reaction Clamping. Marine Biotechnology 13(2): 305-313.

Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H et al. (2013) The Long-Term Stability of the Human Gut Microbiota. Science 341(6141).

Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proceedings of the National Academy of Sciences 108(50): 20166-20171.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences 108(23): 9530-9535.

Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M et al. (2011) Counting absolute numbers of molecules using unique molecular identifiers. Nat Meth 9(1): 72-74.

Krueger F, Andrews SR, Osborne CS (2011) Large Scale Loss of Data in Low-Diversity Illumina Sequencing Libraries Can Be Recovered by Deferred Cluster Calling. PLoS ONE 6(1): e16607.

Larkin MA BG, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG (2007) ClustalW and ClustalX version 2. Bioinformatics 23(21): 2947-2948.

Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. Proceedings of the National Academy of Sciences 104(27): 11436-11440.

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J et al. (2012) Defining the core Arabidopsis thaliana root microbiome. Nature 488(7409): 86-90.

Magoč T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27(21): 2957-2963.

Paradis E, Claude J, Strimmer K (2004) APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289-290.

Patin N, Kunin V, Lidström U, Ashby M (2013) Effects of OTU Clustering and PCR Artifacts on Microbial Diversity Estimates. Microbial Ecology 65(3): 709-719.

Pei AY, Oberdorf WE, Nossa CW, Agarwal A, Chokshi P et al. (2010) Diversity of 16S rRNA Genes within Individual Prokaryotic Genomes. Applied and Environmental Microbiology 76(12): 3886-3897.

Ray A, Nordén B (2000) Peptide nucleic acid (PNA): its medical and biotechnical applications and promise for the future. The FASEB Journal 14(9): 1041-1060.

Sakai M, Ikenaga M (2013) Application of peptide nucleic acid (PNA)-PCR clamping technique to investigate the community structures of rhizobacteria associated with plant roots. Journal of Microbiological Methods 92(3): 281-288.

Sheward DJ, Murrell B, Williamson C (2012) Degenerate Primer IDs and the Birthday Problem. Proceedings of the National Academy of Sciences 109(21): E1330.

Sim K, Cox MJ, Wopereis H, Martin R, Knol J et al. (2012) Improved Detection of Bifidobacteria with Optimised 16S rRNA-Gene Based Pyrosequencing. PLoS ONE 7(3): e32543.

Tanaka T, Matsuoka M, Sutani A, Gemma A, Maemondo M et al. (2010) Frequency of and variables associated with the EGFR mutation and its subtypes. International Journal of Cancer 126(3): 651-655.

Team RC (2012) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Terahara T, Chow S, Kurogi H, Lee S-H, Tsukamoto K et al. (2011) Efficiency of Peptide Nucleic Acid-Directed PCR Clamping and Its Application in the Investigation of Natural Diets of the Japanese Eel Leptocephali. PLoS ONE 6(11): e25715.

Troedsson C, Lee RF, Walters T, Stokes V, Brinkley K et al. (2008) Detection and Discovery of Crustacean Parasites in Blue Crabs (Callinectes sapidus) by Using 18S rRNA Gene-Targeted Denaturing High-Performance Liquid Chromatography. Applied and Environmental Microbiology 74(14): 4346-4353.

von Wintzingerode F, Landt O, Ehrlich A, Göbel UB (2000) Peptide Nucleic Acid-Mediated PCR Clamping as a Useful Supplement in the Determination of Microbial Diversity. Applied and Environmental Microbiology 66(2): 549-557.

Warnes GR (2011) gplots: Various R programming tools for plotting data.

Wickham H (2009) ggplot2: elegant graphics for data analysis. Springer New York.

CHAPTER 4

CONCLUSIONS AND FUTURE DIRECTIONS

LIMITATIONS OF CURRENT WORK

One potential explanation of the failure of current DNA sequencing methods to identify numerous and robust plant-genotype dependent microbial associations is that sequencing restricted to ribosomal regions cannot readily infer microbial genome sequence or function. For example, the genus Pseudomonas contains strains ranging from plant growth promoting to pathogens, which, depending on the length and region of 16S ribosomal gene sequenced, may appear the same (Blakney and Patten 2011). Single genes in otherwise closely related microbial backgrounds (by ribotyping or whole genome content) may determine whether plant-associated microbes can colonize as pathogens or are instead recognized by the plant immune system and thus fail to grow (Jones and Dangl 2006). In a complex wild soil, 16S might suggest all plants associate with Pseudomonas, but hide the fact that each plant genotype was only compatible for colonization by certain Pseudomonads. An alternative possibility for the failure to attribute a larger fraction of microbiome community variance to host genotype is that the majority of microbes colonizing a plant species may be generalists with a wide host range. This is consistent with the concept of a core microbiome (Bulgarelli et al. 2012; Lundberg et al. 2012) that is largely stable across ~25-30 million years of separation in the Brassicaceae, including *A. thaliana* and two of its relatives (Schlaeppi et al. 2013). Whether root-associated microbes have host specificity that current methods do not detect, or whether they are generalists, will become

more clear as we learn more about plant microbiome function. And it may be that some plant-microbe interactions only occur under specific nutrient or stress conditions, consistent with the observations of soil-specific contributions to root endophytic compartment microbiomes (Bulgarelli et al. 2012; Lundberg et al. 2012).

Despite advances in high-throughput sequencing technologies, it is still difficult to answer mechanistic questions regarding microbiome function due to lack of experimental genetic or functional characterization. While in some cases it may be possible to reconstruct bacterial functional profiles from phylotyping (Langille et al. 2013), this is dependent on a high quality reference database, currently lacking for many plant-associated microbiota. In cases of simple microbial environments, including synthetic communities deployed in microcosm reconstitution experiments, the utility of ribosomal gene sequences is higher, because there is no ambiguity in mapping sequence reads to input organisms (Reyes et al. 2013). Because the synthetic community approach is limited to only culturable taxa, it is imperative that significant effort is given to the isolation and genome sequencing of plant-associated microbes to increase that fraction of the root- or leaf-enriched microbiota that can be cultured. Such collections will ultimately lead to synthetic communities of known and individually traceable microbial inputs, providing uniform experimental tools. From such experiments, quantitative information about one or a few marker genes can provide high confidence in the exact membership of the microbial community for each plant, and if the genomes of the community members are known, the full metagenome can be inferred from the set of genes detected (Faith et al. 2014; Faith et al. 2013; Faith et al. 2011; Goodman et al. 2009; McNulty et al. 2013; Rey et al. 2013). One hope is that synthetic recolonization experiments will reveal host-genotype dependent associations with specific microbial gene functional groups

Higher resolution techniques are needed to develop cocktails of microbial probiotics for plant health and to discover plant loci that enhance these functions for agronomic use. Because sequencing costs continue to drop, it is likely that shotgun metagenomics, metatranscriptomics, and, as the relevant technologies become more available, metaproteomics and metabolomics approaches (Ye et al. 2013) applied to large numbers of samples will increasingly describe communities in and on plants grown wild soils, ultimately impacting plant health (Mendes et al. 2013). Existing work looking across multiple samples with both 16S ribotyping, as well as whole metagenome sequencing, demonstrates that taxonomically diverse microbes may be functionally redundant (Burke et al. 2011; Lozupone et al. 2012). Thus, in some cases where organisms have the same ribosomal sequence, but different genomic content, ribosomal sequencing may mask functional diversity, whereas in other cases, microbes from different lineages may have similar functional roles. Further, approaches using conserved genes other than 16S are promising (Sunagawa et al. 2013), but will not overcome the fundamental limitations of single marker profiling. The power of re-colonization experiments to derive both the rules for the establishment and microbial interactions with hosts will expand rapidly. These developments, and their use in controlled environment conditions, are imperative if we are to successfully define and deploy mixtures of functionally redundant microbes as probiotics in uncharacterized environments where they will need to outperform indigenous communities.

FUTURE DIRECTIONS

Establish experimental system for "synthetic community" microcosm reconstitution experiments.

We have had an ongoing effort to culture single isolates of bacteria from Arabidopsis roots grown in Mason Farm or Clayton soil, and have collected to date about 600 strains. For each strain we made a freezer stock stored at -80 degrees C for future use, and we have sequenced the V8 and V4 regions of the 16S rRNA gene (Figure 3.1). Using pairwise sequence comparisons, we can match each isolate to the closest OTU from our deep-sequencing work of plants grown in those same soils. Bacteria that are close matches to OTUs (Chapter 2) are likely to be closely related to the bacteria that contributed to those OTUs.

This matching has revealed that we can readily culture more than 50% of the phylotypes that are enriched in the EC of *A. thaliana* based on the work in Chapter 2, which has encouraged us to continue culturing and to further investigate novel culturing media to find isolates that match currently unrepresented EC-enriched OTUs. We also keep and store root isolates that do not match OTUs enriched in the EC, because they nonetheless were associated with roots and may be important soil microorganisms that commonly interact with root-associated bacteria. We have not yet cultured fungi, because growth, quantification, and genetics are all more difficult than with bacteria, but this is an important future direction, especially as fungal databases and fungal quantification technologies improve, because fungi are in some cases growth promoting in Arabidopsis (Contreras-Cornejo et al. 2009; Qiang et al. 2012), and certainly are important pathogens (Foley et al. 2013).

Having frozen stocks available for experiments makes it possible to grow and mix them into specific cocktails, or "synthetic communites", with which to inoculate seeds or

plants. For highest reproducibility, numerous identical freezer stocks can be made from a single bacterial suspension; each of these can be thawed, cleaned of the glycerol cryoprotectant, and used directly in an experiment without an intermediate growth step.

Adding defined bacteria to an otherwise sterile substrate means that the organisms in the system are known. Sterilizing natural soils and potting soils is difficult to accomplish without changing the properties of the soil in unpredictable ways or introducing toxic compounds (Boyd 1971), but other more inert materials (sand, perlite, calcined clay) can be autoclaved without substantially changing their properties, and can be used as a plant growth substrate. I chose to use fine calcined clay (~ 1 mm particle size), a material used extensively in hydroponic gardening, one which has good water retention, and which allows one to easily extract clean roots (Eddy and Hahn 2008). If not fertilized in any way, the inert calcined clay barely supports growth of Arabidopsis, with rosettes only millimeters across at maturity (data not shown). However, buffered nutrient solutions such as MS media, or perturbations including nitrogen and phosphorus deficiencies, can be used to irrigate the calcined clay to either promote healthy growth or to create classic nutrient deficiency phenotypes. Arabidopsis grown in calcined clay irrigated with MS media reaches a similar size to plants grown in potting soil (Eddy and Hahn 2008).

For a proof of concept synthetic community experiment, we mixed a synthetic community of 42 bacterial members, where 42 represented the largest set of isolates from our collection for which each isolate differed from all other isolates by at least 3 SNPs in the V4 region of the 16S rRNA gene that we use for MiSeq phylotyping (Chapter 3). This threshold of 3 SNPs, which would require multiple mutations to convert the sequence into another of the 42, was chosen as a conservative first pass. In cases where multiple strains had V4 sequences less than 3 SNPs distant from each other, we prioritized for the community those isolates which had a V8 sequence matching one of the EC-enriched OTUs

in Chapter 2. We chose 28 EC-enriched isolates which fell into 18 families, with the *Bacillaceae* and *Streptomycetaceae* each represented by 4 strains. We then added additional individually-distinguishable bacteria from diverse phyla to the community to make the community as large as possible while satisfying the ≥ 3 SNPs criteria, including DH5α *E. coli* as a control

The 42 strains were resuspended in dilute MS media buffered with MES to pH 6.0. They were either mixed to a final concentration of 10^5 cfu / mL for each strain ("equal OD input", Figure 4.1), or each strain was assigned a final concentration of 10^3, 10^4, 10^5, or 10^6 cfu / mL. (perturbed OD input, Figure 4.1) The communities were irrigated onto autoclaved, dry calcined clay and sterile 1 week old Col-0 and Cvi-0 seedlings were transferred from agar plates to the inoculated clay. The plants were grown in a growth chamber using methods in Chapter 2 and watered as needed with sterile distilled water; roots were harvested at bolting and the root-associated microbes were quantified using methods in Chapter 3. In an accompanying experiment by Natalie Breakfield and Meghan Feltcher, the "equal OD" community was also mixed into phytoagar with MS media and 1 week old Col-0 seedlings were transferred to the agar. Agar plates were grown vertically, and after 2 weeks of growth seedlings were harvested and microbes in roots were quantified by 16S sequencing. The results of this early experiment show that while the agar and calcined clay systems show major differences, some of the bacteria from in inoculum do recolonize the root in both systems (Figure 4.1). Contamination from air and water in the calcined clay was minimal, around 5% of total sequences. To further create a gnotobiotic system, I also created growth containers modeled on a test tube using pots with sealed bottoms and modified Magenta™ jars (Figure 4.2). While we do not yet have sequence data from these, visually they show less contamination by ambient surface fungus that calcined clay pots grown in open air. In contrast to sealed containers, the gnotobiotic jars with calcined clay

allow gas exchange, lower humidity, healthier rosette development to maturity, and can be serviced to add sterile water or new treatments.

An experiment I consider an important next step, which is currently running, involves expanding upon the 42 member community experiment to include more isolates, more genotypes, and more replicates. Error rates were low enough in the 42-member experiment that we feel confident identifying and quantifying sequences differing by only 1 SNP, allowing us to independently quantify 62 strains. Sur Herrera Paredes and I inoculated calcined clay, both in open pots and in the gnotobiotic jars, with a mix of 62 bacterial isolates in buffered dilute MS media, and planted four different Arabidopsis accessions (Col-0, Shahdara, Cvi-0, and Oy-0) plus the related Brassicaceae *Capsella rubella* (Slotte et al.) and the model grass *Brachypodium distachyon* (Vogel et al. 2009). There are bulk soil controls as well as uninoculated plant controls for each genotype. The results from this experiment will reveal if the host genetic background alters the final relative abundance of these 62 bacteria; the question is the same as that asked in Chapter 2, but the precise tracking of individual strains made possible by defining the organisms allowed into the experiment will help resolve whether the root colonizers are generalists or whether there is preferential colonization of certain host genotypes. All of these 62 bacteria have or will have their genome sequenced, so we will be able to infer and computationally reconstruct the metagenomes enriched in each plant from the relative abundance of 16S rRNA genes, which will also let us see if there is consistency in enriched functional content (Burke et al. 2011).

Tag genomes of bacterial isolates

We are interested in understanding if closely related bacteria differ in their ability to colonize the root, which sets the stage for genetic dissection of bacterial colonization components. We are also interested in understanding how synthetic communities that we create in the lab, such as communities that confer benefits to the plant, may fare when unleashed into wild soil, where closely related organisms may already exist in abundance. Tracking a synthetic community among wild organisms, becomes a challenge, and the 16S rRNA gene loses utility in this context because closely related strains share the same 16S sequence. Whole metagenome sequencing also has limited utility for quantifying closely-related strains inside the root because 1) whole genome sequences for all bacteria of interest are required 2) it is difficult to overcome the abundance of host DNA and 3) when the strains of interest are closely-related, only a small fraction of the sequencing reads will actually be unique to each strain.

A strategy worth pursuing to overcome these problems is tagging the genomes of strains of interest with a unique sequence. This sequence can either be integrated via transposon into the highly-conserved Tn7 site (Choi et al. 2005), or put onto a stable single-copy plasmid. I designed a construct that drives a fluorescent protein from the broad host range promoter described in Choi et al. (Choi et al. 2005). The transcription terminator is linked to a unique 19 base pair sequence (added by priming with a degenerate primer) that serves as the strain ID, and the terminator and strain ID are flanked by two synthetic sequences that serve as annealing points for PCR primers (Figure 4.3). When this construct is integrated into the Tn7 site or a stable plasmid, the transformed bacteria will glow and can be imaged under a confocal microscrope, but importantly, each strain will also have a unique ID. Tagged strains can then be inoculated to plants growing on wild soil, and the terminator + ID can be amplified via PCR and sequenced using methods described in

Chapter 3. Each input strain can then be quantified by counting the abundance of its strain ID in the total pool of strain ID amplicons. We are well underway building this system.

Improve plant phenotyping

The ultimate practical goal of the research in this dissertation is to better understand how the genome of the host plant might influence the colonization of beneficial root bacteria, and eventually improve the health of the plant. To know if plant health is improving, a measure of plant health is needed. For example, rosette size is related to fecundity (Kawano 2012). Furthermore, plants that are lacking important nutrients and micronutrients, or that are affected by toxic levels of chemicals, often reveal this stress in their leaf color. For example, phosphate-deficient Arabidopsis is smaller and accumulates red-colored anthocyanins (Ticconi et al. 2001). Thus, images of rosettes can provide valuable information about overall plant health.

Manually measuring features of live plants is slow, but automated computational image analysis provides an opportunity to measure plant health quickly. Overhead images of plants growing in opaque substrates such as calcined clay and wild soil allow for analysis of rosette size, rosette color, and even more sophisticated phenotypes. Such remote phenotyping is an active research area for crop plants (Peng et al. 2011) and are used by biotech companies.

I have helped, along with Surge Biswas, to develop a high-throughput imaging system for Arabidopsis. Multiple plants in a flat can be imaged from overhead using a copy stand. Individual plants, or positions of the flat, have custom barcodes attached to them. We created software to recognize the barcodes in the image and segment the image into subimages of the individual plants (Figure 4.4); our goal is then to adopt and create software

to measure rosette size and rosette color attributes in each image, which can be converted into quantitative data that can be analyzed immediately. The speed and ease of such imaging makes it practical to image hundreds of plants every week, and search for links between growth rate and soil or endophytic microbial communities.

Of particular interest will be microbes or communities of microbes that result in larger and greener rosettes, microbes that can reverse the poor health of plants grown in nutrient-deficient soil, or microbes that can reverse the poor health of plants grown in the presence of pathogens.

FIGURES

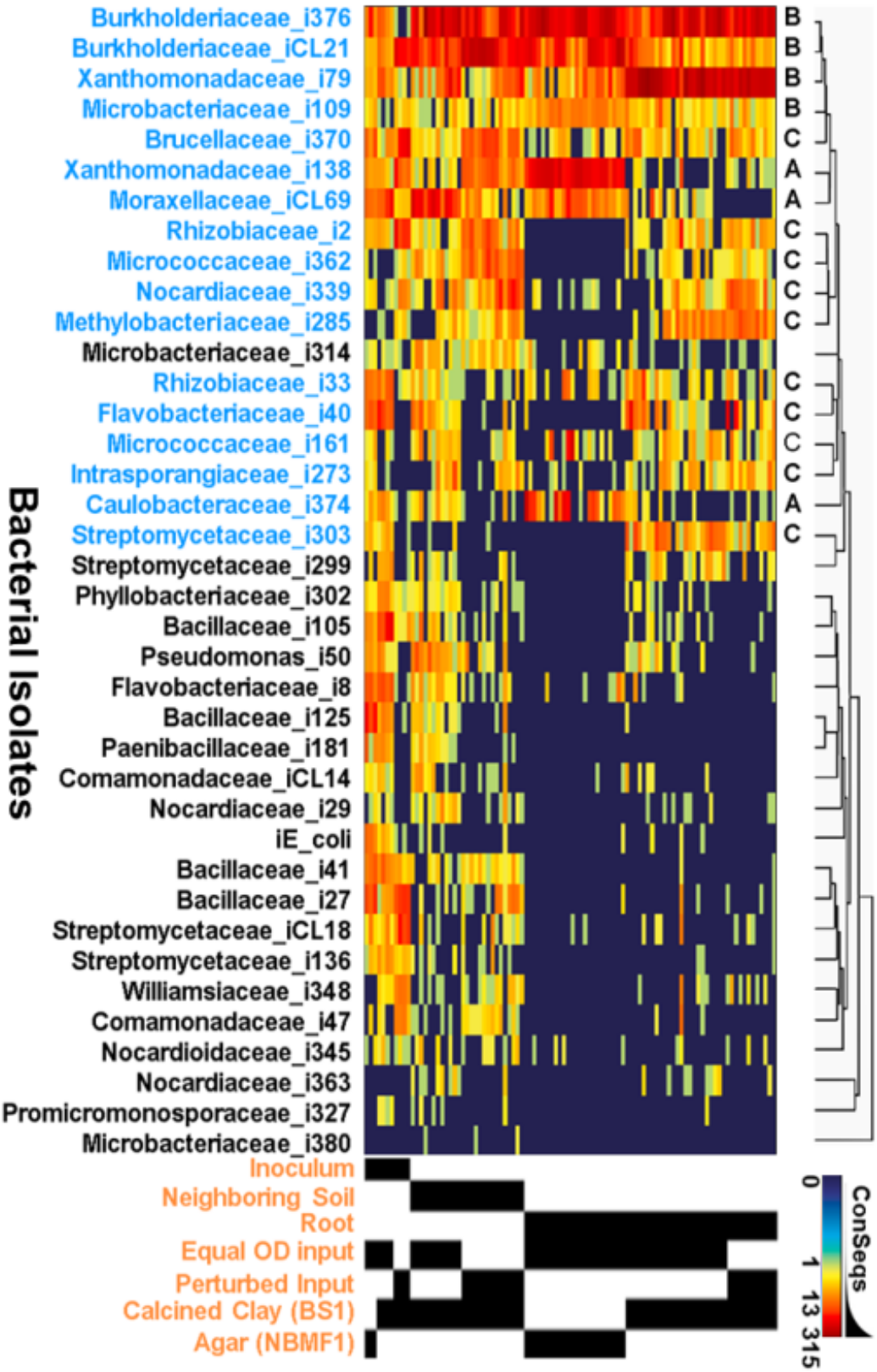

Bacterial Isolates
Burkholderiaceae_i376
Burkholderiaceae_iCL21
Xanthomonadaceae_i79
Microbacteriaceae_i109
Brucellaceae_i370
Xanthomonadaceae_i138
Moraxellaceae_iCL69
Rhizobiaceae_i2
Micrococcaceae_i362
Nocardiaceae_i339
Methylobacteriaceae_i285
Microbacteriaceae_i314
Rhizobiaceae_i33
Flavobacteriaceae_i40
Micrococcaceae_i161
Intrasporangiaceae_i273
Caulobacteraceae_i374
Streptomycetaceae_i303
Streptomycetaceae_i299
Phyllobacteriaceae_i302
Bacillaceae_i105
Pseudomonas_i50
Flavobacteriaceae_i8
Bacillaceae_i125
Paenibacillaceae_i181
Comamonadaceae_iCL14
Nocardiaceae_i29
iE_coli
Bacillaceae_i41
Bacillaceae_i27
Streptomycetaceae_iCL18
Streptomycetaceae_i136
Williamsiaceae_i348
Comamonadaceae_i47
Nocardioidaceae_i345
Nocardiaceae_i363
Promicromonosporaceae_i327
Microbacteriaceae_i380
B B B B C A A C C C C C C C C C A C
Inoculum
Neighboring Soil
Root
Equal OD input
Perturbed Input
Calcined Clay (BS1)
Agar (NBMF1)
ConSeqs
0
1
13
315

Figure 4.1: Re-colonization of Arabidopsis root endophytic compartments in two different experimental systems.

Heatmap illustrating the abundance of 38 out of 42 input isolates (rows) across individual samples (columns) from two synthetic community experiments, NBMF1 (sterile agar) and BS1 (sterilized calcined clay). Four isolates had fewer than one read on average in the inoculum and are therefore not included. The abundance of an isolate is measured as the number of molecule tag consensus sequences that exactly match its V4 reference Sanger sequence, and molecule tag consensus sequences were formed using the default parameters of MTToolbox (Lundberg et al. 2013). Covariate bars (black bars, bottom) describe each sample's fraction (inoculum, neighboring soil, root), inoculum evenness (Equal OD input; Perturbed Input), and surrounding medium (calcined clay, MS agar). Rows of the heatmap are sub-blocked according to each isolate's ability to colonize the root in both experiments (8 isolates), NBMF1 only (3 isolates), BS1 only (8 isolates), or neither experiment (19 isolates). An isolate is said to *colonize* the root if the probability of its presence in a sample statistically exceeds 0.5 (q-value < 0.05, binomial test). Names of Isolates that colonized, based on a greater than 50% probability of being observed in a root, are colored light blue. The experiment in which that isolate colonized is noted underneath the dendrogram (A, agar only; C, calcined clay only; B, both agar and calcined clay). Figure and analysis made with help from Surojit Biswas.

Figure 4.2: Gnotobiotic calcined clay system.

Left: pots are sealed on the bottom with silicone so that microbes and water cannot pass, and then filled with calcined clay. Pots are placed on a perforated surface that allows for airflow. The pots are covered with an upside down Magenta^{TM} GA-7 jar and the whole unit is autoclaved. Right: Calcined clay is irrigated with growth media at 40% its dry volume and sterile seeds or seedlings are transferred – pots are grown in a climate-controlled growth chamber and serviced as necessary in a sterile hood.

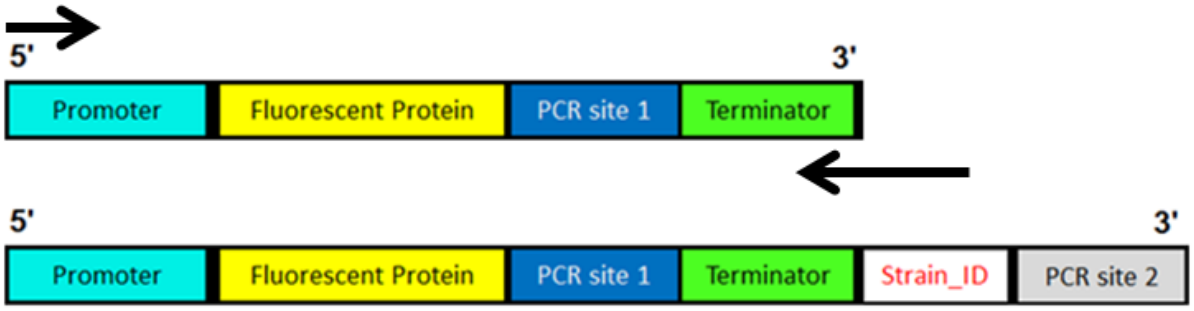


Figure 4.3: Tagging construct for bacterial isolates

The broad host range promoter $P_{A1/04/04}$ (a P_{lac}-derivative) (Lambertsen et al. 2004) drives a citrine gene or other fluorescent protein. A unique PCR-able site is 5' of the transcription terminator (top). To add a strain ID, the promoter through the terminator are PCR-amplified with a forward primer and a degenerate reverse primer that adds the strain ID and a second PCR-able site (arrows), generating millions of possible constructs with different strain IDs (bottom). Each unique construct can be used to transform one and only one isolate, giving each an independent strain ID. The relative abundance of all tagged isolates can be measured by sequencing the amplicon defined by PCR site 1 and PCR site 2.

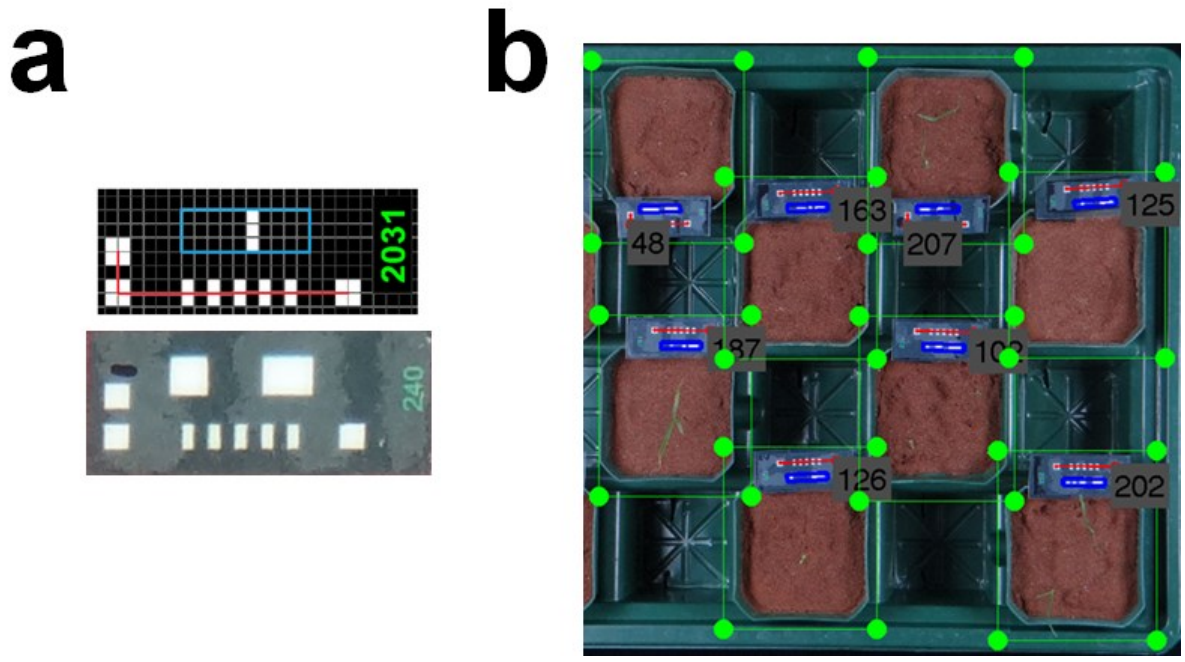


Figure 4.4: High throughput rosette imaging

a) Barcodes are produced using heatmap-type functions in R, printed, and laminated. Each barcode has three white squares in an L orientation (red line, top), where the long axis has a 'zebra' pattern that helps delimit the 11 horizontal units of the barcode (blue box). An actual image of a printed and laminated barcode is shown below.

b) Actual image of 8 pots, each with a barcode clipped to the side of the pot via a binder clip. The matlap script recognizes the "L" orientation for each barcode (red), boxes the barcode (blue), and reads the binary barcode (digits). Each pot is then cropped (green boxes) and the subimage is saved as a new file with a filename that incorporates the barcode digits.

REFERENCES

Blakney AJ, Patten CL (2011) A plant growth-promoting pseudomonad is closely related to the Pseudomonas syringae complex of plant pathogens. FEMS Microbiol Ecol 77(3): 546-557.

Boyd HW (1971) Manganese toxicity to peanuts in autoclaved soil. Plant and Soil 35(1-3): 133-144.

Bulgarelli D, Rott M, Schlaeppi K, Ver Loren van Themaat E, Ahmadinejad N et al. (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature 488(7409): 91-95.

Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T (2011) Bacterial community assembly based on functional genes rather than species. Proc Natl Acad Sci U S A 108(34): 14288-14293.

Choi K-H, Gaynor JB, White KG, Lopez C, Bosio CM et al. (2005) A Tn7-based broad-range bacterial cloning and expression system. Nat Meth 2(6): 443-448.

Contreras-Cornejo HA, Macías-Rodríguez L, Cortés-Penagos C, López-Bucio J (2009) Trichoderma virens, a Plant Beneficial Fungus, Enhances Biomass Production and Promotes Lateral Root Growth through an Auxin-Dependent Mechanism in Arabidopsis. Plant Physiology 149(3): 1579-1592.

Eddy R, Hahn DT (2008) 101 Ways to Try to Grow Arabidopsis: What Root Media Worked Best to Cleanly Remove Roots? Purdue Methods for Arabidopsis Growth: Paper 4.

Faith JJ, Ahern PP, Ridaura VK, Cheng J, Gordon JI (2014) Identifying gut microbe-host phenotype relationships using combinatorial communities in gnotobiotic mice. Science translational medicine 6(220): 220ra211.

Faith JJ, Guruge JL, Charbonneau M, Subramanian S, Seedorf H et al. (2013) The Long-Term Stability of the Human Gut Microbiota. Science 341(6141).

Faith JJ, McNulty NP, Rey FE, Gordon JI (2011) Predicting a human gut microbiota's response to diet in gnotobiotic mice. Science 333(6038): 101-104.

Foley RC, Gleason CA, Anderson JP, Hamann T, Singh KB (2013) Genetic and Genomic Analysis of <italic>Rhizoctonia solani</italic> Interactions with Arabidopsis; Evidence of Resistance Mediated through NADPH Oxidases. PLoS ONE 8(2): e56814.

Goodman AL, McNulty NP, Zhao Y, Leip D, Mitra RD et al. (2009) Identifying genetic determinants needed to establish a human gut symbiont in its habitat. Cell host & microbe 6(3): 279-289.

Jones JDG, Dangl JL (2006) The plant immune system. Nature 444(7117): 323-329.

Kawano S (2012) Biological Approaches and Evolutionary Trends in Plants: Elsevier Science.

Lambertsen L, Sternberg C, Molin S (2004) Mini-Tn7 transposons for site-specific tagging of bacteria with fluorescent proteins. Environmental Microbiology 6(7): 726-732.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D et al. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 31(9): 814-821.

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

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J et al. (2012) Defining the core Arabidopsis thaliana root microbiome. Nature 488(7409): 86-90.

Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL (2013) Practical innovations for high-throughput amplicon sequencing. Nat Meth 10(10): 999-1002.

McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK et al. (2013) Effects of diet on resource utilization by a model human gut microbiota containing Bacteroides cellulosilyticus WH2, a symbiont with an extensive glycobiome. PLoS biology 11(8): e1001637.

Mendes R, Garbeva P, Raaijmakers JM (2013) The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS microbiology reviews 37(5): 634-663.

Peng Y, Gitelson AA, Keydan G, Rundquist DC, Moses W (2011) Remote estimation of gross primary production in maize and support for a new paradigm based on total crop chlorophyll content. Remote Sensing of Environment 115(4): 978-989.

Qiang X, Zechmann B, Reitz MU, Kogel K-H, SchÃ¤fer P (2012) The Mutualistic Fungus Piriformospora indica Colonizes Arabidopsis Roots by Inducing an Endoplasmic Reticulum Stress-Triggered Caspase-Dependent Cell Death. The Plant Cell Online 24(2): 794-809.

Rey FE, Gonzalez MD, Cheng J, Wu M, Ahern PP et al. (2013) Metabolic niche of a prominent sulfate-reducing human gut bacterium. Proc Natl Acad Sci U S A 110(33): 13582-13587.

Reyes A, Wu M, McNulty NP, Rohwer FL, Gordon JI (2013) Gnotobiotic mouse model of phage-bacterial host dynamics in the human gut. Proceedings of the National Academy of Sciences 110(50): 20236-20241.

Schlaeppi K, Dombrowski N, Oter RG, Ver Loren van Themaat E, Schulze-Lefert P (2013) Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. Proc Natl Acad Sci U S A.

Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F et al. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet 45(7): 831-835.

Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA et al. (2013) Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 10(12): 1196-1199.

Ticconi CA, Delatorre CA, Abel S (2001) Attenuation of Phosphate Starvation Responses by Phosphite in Arabidopsis. Plant Physiology 127(3): 963-972.

Vogel JP, Garvin DF, Mockler TC, Schmutz J, Rokhsar D et al. (2009) Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature 463(7282): 763-768.

Ye H, Gemperline E, Venkateshwaran M, Chen R, Delaux PM et al. (2013) MALDI mass spectrometry-assisted molecular imaging of metabolites during nitrogen fixation in the Medicago truncatula-Sinorhizobium meliloti symbiosis. The Plant journal : for cell and molecular biology 75(1): 130-145.