Correction: 23 July 2015

www.sciencemag.org/cgi/content/full/science.aaa8764/DC1

Supplementary Material for

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

Sarah L. Lebeis, Sur Herrera Paredes, Derek S. Lundberg, Natalie Breakfield, Jase Gehring, Meredith McDonald, Stephanie Malfatti, Tijana Glavina del Rio, Corbin D. Jones, Susannah G. Tringe, Jeffery L. Dangl

Published 16 July 2015 on *Science* Express
DOI: 10.1126/science.aaa8764

This PDF file includes:

Materials and Methods
SupplementaryText
Figs. S1 to S14
Full Reference List

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/science.aaa8764/DC1)

Tables S1 to S10 as separate Excel files
Databases S1 to S4 as zipped archives

Correction: Several typographical errors were corrected. In lines 1 to 4, the Roman numerals were incorrectly started at II, instead of I. Line 247 and line 258 were missing fig. S10 call outs. Figure S10A incorrectly had cpr1 marked as significantly different. In the Fig. S10 legend, the wrong statistical test was written; additionally, it was clarified that C and D were plants grown on plates, not in soil. In the Fig. S11 legend, a Methods reference was added.

I. **List of Supplemental Tables**
II. **List of Supplemental Figures**
III. **List of Supplemental Datasets**
IV. **Methods**

I. List of Supplemental Tables:

Supplemental Table captions:

ST1 – Description of plant genotypes used

Each of the plant genotypes used. For each genotype the full genotype, the name of each single mutation contained in each mutant, the gene number for each mutation, the expression level in roots, the salicylic acid related phenotype, the reference from where each mutant was created and the origin of the seed stock.

ST2 – Primer sequences

Primer sequences utilized for the different library preparations. For the census experiments primers organized according to sequencing technology and region. For the synthetic community experiments Method 2b) we provide the molecule tagging and PCR step primers as well as the PNA sequences. The index primer sequences can be found in Lundberg et. al 2013 (*17*). The extra primers required for the microbial density (Method 2b) protocol are also provided.

ST3 – Sample numbers in each experiment

Top: Number of samples per genotype and fraction for each of the census experiments. For each genotype, the samples are divided in endophytic compartment (EC) and rhizosphere (R) with the percent of planted seedlings that survived to be harvested at the formation of an inflorescence meristem, which was when we harvested the samples. Bottom: Number of samples per genotype, fraction, inoculation, and fraction treatment for each of the synthetic community experiments. To account for experimental variability, every experiment contains wild-type Col-0 and each mutant genotype was present in at least two experiments.

ST4 – ZINB Model results for census experiments and SynCom

ZINB model results for each of the datasets used in this paper. **A)** The main survey done with Roche 454 at the family level; **B)** same as A but at the OTU level; **C)** subset re-sequencing of region V4 with Illumina MiSeq; **D)** subset re-sequencing of region V8 with Illumina MiSeq. **E)** SynCom. On each table, the results of parameter estimation and testing for each variable and taxon combination is given. Estimate is the coefficient value from the model; StdErr is the standard deviation of the Estimate; z.value is the z-score used for calculating a p-value (p.value); p-values are corrected by the Benjamini-Hochberg method to obtain q-values (q.value); and model indicates the type of model that was selected by the AIC criterion (Method 6b). Finally, the “Legend” tab provides a description for each variable on the models.

ST5 – Percent variance explained by each variable

Results of CAP estimation of the percent of total variation explained by each variable (Method 6e). The analysis was performed both on the census and on the SynCom datasets, and a 95% confidence interval was calculated with 1000 bootstrap pseudoreplicates.

ST6 – Genotype Differentially Abundant (DA) families

Bacterial families that are differentially abundant between Col-0 and each of the mutants in the census experiment according to the ZINB model. The table shows families as rows and genotypes as columns, where a value of 1 means enrichment and a value of -1 means depletion relative to Col-0, while 0 means no statistically significant difference. This table was generated by finding all the families with a significant q-value in table ST4a for any of the fracgenE variables, and the sign of the estimate on the same table was used to determine enrichment and depletions. The order of the rows matches the order of families in the heatmap found in fig. S6A.

ST7 – Genotype Differentially Abundant OTUs

Bacterial OTUs that are differentially abundant between Col-0 and each of the mutants in the census experiment according to the ZINB model. The table shows OTUs as rows and genotypes as columns, where a value of 1 means enrichment and a value of -1 means depletion relative to Col-0, while 0 means no statistically significant difference. This table was generated by finding all the OTUs with a significant q-value in table S4b for any of the fracgenE variables, and the sign of the estimate on the same table was used to determine enrichment and depletions. The order of the rows matches the order of families in the heatmap found in fig. S7A.

ST8 – Technically robust EC-enriched and EC-depleted families

List of families that have overlapping enrichment or depletion in EC with respect to bulk soil in the three census datasets: Roche 454 V8, Illumina MiSeq V8 and Illumina MiSeq V4. These families were obtained by finding all of the families that had a significant q-value for any of the fracgenE variables in table S4 while overlap in enrichment or depletion is defined as the value of "Estimate" in table S4 having the same sign.

ST9 – List of SynCom isolates

For each isolate this table provides its phylum and family. It also indicates whether the family of that isolate was enriched in the endophytic compartment (EC-enriched) or differentially abundant between wildtype and a mutant (genotype-DA) in the census experiment. It also shows the results from the SynCom experiments, indicating which isolates were robust colonizers, EC-enriched or EC depleted (Fig. 2C), as well as which isolates were genotype-DA (Fig. 3C) or SA-DA (Fig. 4). The column Category summarizes the consistency between census and SynCom experiments: EC-enriched confirmed indicates that an isolate is EC enriched both in the census and SynCom; similarly, DA-confirmed indicates that an isolate is genotype-DA in both the census and SynCom. Finally we provide genome assembly identifiers for the sequenced isolates and full isolate names, these two terms can be used to retrieve the genome sequence and annotations on the IMG/ER website.

ST10 –Salicylic acid metabolism in isolates

For isolates noted at left with sequenced genomes, we performed BLAST searches to identify salicylic acid degradation or biosynthesis (A) genes (Method 6g). For each isolate, we show the results of the ZINB model regarding salicylic acid (same information as table S4D), and we show the percent identity and percent query coverage (number in parenthesis) for each experimentally characterized gene involved in salicylic acid metabolism. Yellow background indicates a strong match, while a green background indicates that both the query and the subject had the same annotation, regardless of the quality of the match. Query genes were retrieved from MetaCyc and their sequences can be found in (B).

II. List of Supplemental Figures and Figure Legends

SF1 – Alpha and beta diversity for different 16S rRNA and ITS regions

SF2 – DEPS and JEN root microbiome communities contain a disproportionate number of oomycete mitochondria reads

SF3 – Processing pipeline for Roche 454 census experiments

SF4 – Sample fraction drives differences in alpha and beta diversity of root microbiome communities

SF5 – Differential abundance of Proteobacteria families in different sample fractions

SF6 – Genotype-DA family enrichments and depletions

SF7 – Genotype-DA OTU enrichments and depletions

SF8 - The absolute quantification of bacteria in samples grown in MF soil

SF9 – Zero-Inflated Negative Binomial (ZINB) model

SF10 – Salicylic acid production in MF soil and root morphology of defense phytohormone mutants

SF11 – Technical reproducibility between variable regions and sequencing platforms

SF12 – Induction of Runaway Cell Death in isd1 mutants grown in SynCom with salicylic acid treatment

SF13 – SynCom differentiates sample fractions

SF14 - Salicylic acid treatment affects SynCom composition, but not growth of isolate #40 in SynCom or in liquid growth curves

III. List of Supplemental Datasets

SD1 – Count tables

SD2 – Mapping files and design matrices

SD3 – Representative sequences

SD4 – Full 16S sequence for SynCom members without IMG#

IV. Materials and Methods:

1) **Census study experimental procedures**
    a. Soil collection and preparation
    b. Defense phytohormone mutant genotypes
    c. Seed sterilization, germination and plant growth
    d. Harvesting and DNA extraction
    e. Measuring root length and morphology
    f. Measuring salicylic acid production in leaves and roots
2) **Massive parallel sequencing library preparations**
    a. 454 16S library preparation and pyrotag sequencing
    b. Illumina library preparation and sequencing at JGI
    c. Illumina library preparation for SynCom experiments
3) **Processing of sequencing data**
    a. Sequence processing pipeline
    b. Pre-processing Roche 454 census experiments
    c. Pre-processing Illumina MiSeq census experiment
    d. Pre-processing Illumina MiSeq synthetic community experiments
    e. Clustering sequences into OTUs
    f. Mapping MT consensus to isolate 16S genes
    g. OTU and isolate annotation
    h. *In silico* pooling of samples in census experiments

4) **Microbial quantification procedures**
    a. CARD-FISH
    b. Differential eukaryotic 18S and prokaryotic 16S determination
5) **Synthetic community (SynCom) experimental procedures**
    a. Microbe isolation
    b. Synthetic community plant inoculation experiments
    c. Growth curves
6) **Statistical analysis**
    a. Diversity analysis for census experiments
    b. ZINB family and OTU-level analysis for census experiments
    c. Definition of a technically robust set of enrichments and depletions
    d. Comparison of enrichment and depletion profiles between genotypes
    e. PCA and CAP analysis of synthetic community experiments
    f. ZINB analysis of synthetic community experiments
    g. Genomic analysis of isolates in synthetic community experiments
    h. Defining robust colonizers in synthetic community experiments

1) Census study experimental procedures:

a. Soil collection and preparation: For each experiment, we collected the top 20cm of earth from Mason Farm (MF), which is managed by the North Carolina Botanical Garden. This site is free from pesticide and fertilizer use and has low human disturbance, providing a fairly stable soil source. Soil micronutrient analysis was used to define this as a loam soil with a variety of nutrients and a pH of 6 (*10*). Soil was dried and crushed using an aluminum mallet. After crushing, debris was removed by sifting, resulting in a very fine soil. To improve drainage, soil is mixed 2:1 with steamed and autoclaved sand. The resulting soil mixture is used to grow plants in 2 x 2 inch square pots for each experiment.

b. Defense phytohormone mutant genotypes: Plant-associated microbial communities promote plant productivity by improving accessibility to nutrients, producing plant growth stimulating factors, and inducing protection against pathogen infection and various abiotic stresses (*18, 19*). The plant immune system detects microbes using highly polymorphic external and internal receptors, which recognize both general microbe-associated molecular patterns and specific pathogen virulence molecules. Salicylic acid (SA) biosynthesis is induced by immune receptor-mediated recognition of microbial pathogens that require living host tissue (biotrophs) (*20*). By contrast, ethylene and jasmonic acid (JA) biosynthesis are induced by pathogens that cause and exploit host cell death (necrotrophs); the consequence of their action contributes to limiting necrotrophic infections (*20*).

In order to determine the role for the phytohormones salicylic acid, jasmonic acid, and ethylene production and signaling in controlling microbiome community composition, we examined the microbial communities of roots in a variety of *Arabidopsis thaliana* defense phytohormone mutants (table S1). We used four hyper-immune mutants (*cpr1*, *cpr5*, *cpr6*, and *snc1*) previously characterized to constitutively produce enhanced levels of salicylic acid and constitutive defense signaling through salicylic acid in leaves (*21-24*). We investigated two classes of immunocompromised mutants, which either lack pathogen-induced biosynthesis of salicylic acid (*sid2*; (25)), or produce salicylic acid, but lack sensitivity to it (*pad4*; (26)). We examined the role of salicylic acid biosynthesis or signaling in combination with a loss in jasmonic acid (JA) biosynthesis and ethylene sensitivity (with *dde2 ein2 pad4 sid2* (DEPS); (27)). In the second class of immunocompromised mutants, we examined the role of salicylic acid sensitivity in combination with jasmonic acid sensitivity (with *npr1 jar1* (NJ); (28)), ethylene sensitivity (with *ein2 npr1* (EN); (28)), or sensitivity to salicylic acid, jasmonic acid, and ethylene (with *jar1 ein2 npr1* (JEN); (28)). Root expression of each of these genes in wild type Col-0 roots was confirmed via Genevestigator's plant biology database (https://genevestigator.com/gv/doc/intro_plant.jsp) with the exception of *CPR6*, which was not in the database.

c. Seed sterilization, germination and plant growth: All seeds were surfaced-sterilized by treatment with 70% ethanol with 0.1% Triton-X100 for 1 minute, 12 minutes of treatment with freshly made bleach solution (10% household bleach and 0.1% Triton-X100), and 3 rinses with sterile distilled water. Seedlings grown from such seeds have previously been shown to not contain endophytic microbes, and this treatment eliminates any seed-borne microbes on the seed surface (10). Seeds were stratified at 4^{o}C in the dark for 3 days and germinated on 0.5% agar containing ½ strength Murashige and Skoog (MS) vitamins and 1% sucrose for 1 week at 24^{o}C under 18 hours of light. Healthy 1 week old seedlings were aseptically transplanted from the MS germinating plates into sterile 2.5 inch square pots filled with Mason Farm soil prepared as described above. We also included "bulk soil" controls, which were pots without plants added to them and were randomly interspersed among the planted pots. All pots, including bulk soil controls, were watered from the top with non-sterile distilled water to avoid chlorine and other tap water additives 2-3 times a week. In order to promote large rosette and root growth, plants were grown in growth chambers with short day, 8 hours of light (108-135 µEinsteins) at 21 and 16 hours of dark at 18^{o}C until the formation of an inflorescence meristem.

d. Harvesting and DNA extraction: Plants and bulk soil controls were harvested and their rhizosphere and endophytic compartment microbial communities isolated as previously described (*10*). At the formation of an inflorescence meristem, the above ground plant organs were aseptically removed and loose soil was physically removed until only soil within 1mm from the root surface remained. The roots were placed in a clean and sterile 50mL conical tube containing 25mL of phosphate buffer (6.33g of NaH2PO4*H20, 16.5g of Na2HPO4*H20, and 200uL Silwet L-77 in 1L of water). Rhizospheres (R) were separated from the roots by vortexing the root system in buffer at maximal speed for approximately 15 seconds. The resulting turbid solution was filtered through a sterile 100µm nylon mesh cell strainer (BD Biosciences) into another sterile 50mL conical tube to filter out plant material, sand, and other large debris. The filtrate was centrifuged in 2 steps to form tight pellets (averaging 250mg), defined as our "rhizosphere" (R) sample. Bulk soil samples were taken by discarding the top 1cm of soil from the pot, homogenizing the remaining soil on a sterile work surface, and scooping approximately 250mg of the mixed soil into a buffer tube and following the same protocol as rhizosphere samples. To isolate the endophytic compartment (EC) microbial community, roots were rinsed in sterile distilled water and debris was aseptically removed with tweezers. Roots were subsequently placed in new sterile phosphate buffer for sonication to remove soil or microbial aggregates remaining on the root surface using a Diagenode Bioruptor set on the low frequency for five minutes (five 30s bursts followed by five 30s rests). The clean sonicated roots constitute the EC samples. All Bulk soil, R, and EC samples were flash frozen and stored at -80^{o}C until DNA was extracted with the 96-well format MoBio PowerSoil kit. For the EC samples, we performed a pre-homogenization step by lyophilizing the root samples, putting them in a 2mL tube with 3 glass beads (4 mm), snap freezing again and running through a cycle on the MPBio FastPrep 24 for 20s at 4.0 m/s. This pre-homogenization allowed us to grind the tissue before adding lysis buffer and ensure that the kit was able to work efficiently.

e. Measuring root length and morphology: For both root length and root morphology measurements, surface sterilized mutant seeds and control seeds were grown vertically on plates containing ½ strength Murashige and Skoog (MS) salt mixture, 1% sucrose, 2.5 mM 2-(N-morpholino) ethanesulfonic acid (pH5.7), and 0.5% phytagar for 7 days. Root lengths were measured using ImageJ (29), and the Student's t-test was used to determine statistical significance (fig. S10C). Seedlings were stained in 10 mg/ml propidium iodide for 0.5 to 2 minutes and mounted in water. Imaging was on a Zeiss LSM710 confocal laser-scanning microscope using the 488-nm laser line for excitation and a 40x water objective (fig. S10D).

f. Measuring salicylic acid production in leaves and roots: We confirmed the salicylic acid hyper-accumulation phenotypes in leaves of *cpr1*, *cpr5*, *cpr6*, and *snc1*, and the absence of salicylic acid in biosynthetic *sid2* mutant leaves. However, we noted low salicylic acid levels in roots of all genotypes grown in wild Mason Farm soil (fig. S10A). Further, *cpr1, cpr5, cpr6, snc1, pad4, and sid2* seedlings were grown axenically for 18 days in vertical plates as described above for tissue to measure salicylic acid accumulation (fig. S10B). Finally, root morphological differences between genotypes grown on agar could not explain the observed overlap in microbiome differences from wildtype (fig. S10C and S10D). Previously, production of salicylic acid has been measured in the leaves of many of the mutants used in this study (*21-24*). For measuring salicylic acid production in the leaves and roots in MF soil, hyper-responsive mutants (*cpr1*, *cpr5*, *cpr6*, and *snc1*) as well as negative control (*sid2*) and isogenic wild type (Col-0) were grown in Mason Farm soil as described above with the exception that 4-5 seedlings of each genotype were grown in a 4.5" pot together to increase the amount of plant material harvested for each sample. When the inflorescence meristem formed, plants were harvested and four 100mg samples of leaves and roots were taken. Samples were snap frozen and stored at -80^{o}C until SAG levels were assessed biochemically. Briefly, the levels of total salicylic acid and salicylic acid glucoside (SAG) were determined for each genotype using the Acinetobacter sp. ADPWH_lux biosensor (30).

2) Massive parallel sequencing library preparations:

a. 454 16S library preparation and pyrotag sequencing: 454 pyrosequencing libraries were created in triplicate using the same protocol as in (*10*) and sequencing was performed at the Joint Genome Institute and Roche. The raw data from the 454 survey experiment is available in the Short Read Archive (ERP010780), and the processed OTU representative sequences are in Supplementary Dataset 3.

b. Illumina library preparation and sequencing at JGI: Three sets of primers were used to amplify the V4 region of the 16S rRNA gene (515F-806R), V8 region of the 16S (1114F-1392R), and ITS intergenic sequence (ITS4 and ITS9) (table S2). In each case, the reverse primer had a unique molecular barcode for each sample. This allowed multiplexing of 92 samples for V4, 48 samples for V8, and 92 for ITS. PCR reactions with ~20ng template were performed with 5 Prime Hot Master Mix in triplicate along with a positive and negative control to reveal contamination. The PCR program used was 94^{o}C for 3 min followed by (94^{o}C for 45 sec, 50^{o}C for 1min, 72^{o}C for 1.5 min) x 35 cycles, followed by 72^{o}C for 10 min and then cool down to 4^{o}C. Reactions were purified using 1.2X volume of AMPureXP magnetic bead and quantified with Qubit HS assay. Amplicons were pooled in equal amounts following qualitative analysis with a Bioanalyzer. Pooled amplicons were then diluted to 10uM and submitted for qPCR for quality control. For family-level microbiome comparisons, samples were sequenced on an Illumina MiSeq machine at the Joint Genome Institute with a target cluster density of 500K/mm^2. Each sample was spiked with approximately 25% PhiX control to increase sequence diversity. The data from the Illumina re-sequencing on the JGI portal. We will need to provide the following url in the appropriate section (http://genome.jgi.doe.gov/Immunesamples/Immunesamples.info.html) and the processed OTU representative sequences are in Supplementary Dataset 3.

c. Illumina library preparation for SynCom experiment: Illumina libraries for the SynCom experiments were created using the same protocol as in (*17*), which allows counting of original template molecules, and sequencing was performed at UNC. The raw data for the SynCom experiments is available in the Short Read Archive (ERP010863).

3) **Processing of sequencing data:**

a. Sequence processing pipeline: Sequences from each platform, library preparation method and experimental design were first pre-processed as described below into a fasta file containing high quality sequences matched to a given sample on the fasta headers. The resulting sequences were then converted into a count table either by clustering into Operational Taxonomic Units (OTUs) or mapping to known isolates' 16S rRNA gene. Representative sequences from OTUs were further taxonomically annotated (see OTU and isolate annotation below). Samples that had less than 1000 usable reads in the census were pooled *in silico* with samples of the same fraction, developmental stage, genotype and experiment that also had less than 1000 usable reads to provide enough depth for statistical analyses (Methods 3h). A number of off-the-shelf tools were used, and in-house Perl scripts filled the gaps (see details below).

b. Pre-processing Roche 454 census experiments: As each 454 plate was sequenced, raw reads from individual plates were immediately run through Pyrotagger (31) to diagnose plate quality (based on the number of reads passing quality checks) and determine if a plate needed to be re-sequenced. Plates with a reasonable number of long, high quality raw reads with matching barcodes were processed and quality controlled following the pipeline defined in (*10*). Briefly, reads were trimmed to 220bp and short reads removed, low quality reads were removed using default quality control settings in QIIME-1.3.0 (32) with the split_libraries.py script, and individual reads were matched to sequence barcodes.

c. Pre-processing Illumina MiSeq census experiment: MiSeq lanes with a high number of sequence pairs matching barcodes and successful merging of paired-ends were used for downstream analysis. An in-house pipeline was implemented in Perl to process these sequences with the following steps: i) sequence pairs were identified and unpaired sequences were discarded; ii) reads were trimmed to 165bp and merged using FLASH (33) (options: -m 30 –M 165 –x 0.25 –r 165 -f 282 –s 20), any read pair that did not merge was discarded; iii) expected primer sequences were matched to the merged sequences using standard regular expression techniques, primer sequences were removed and the resulting "*in silico* amplicons" were kept; any sequences without primer matching were discarded; iv) for the V4 region only, sequences shorter than 240 bp were removed because the primers used for this region also amplify oomycete mitochondrial genes; v) sequences were de-multiplexed.

d. Pre-processing Illumina MiSeq synthetic community experiments: Libraries were prepared following the protocol from (*1*). MiSeq reads were processed with MT-Toolbox (*17, 34*). Briefly, sequence pairs were merged with FLASH (*33*) and merged sequences were binned by molecule tag (MT). The resulting bins were used to correct for PCR and sequencing errors and biases. Only MTs with at least 3 merged sequences were kept for downstream analysis.

e. Clustering sequences into OTUs: For the census experiments (both from 454 and MiSeq), the high quality sequences were clustered into Operational Taxonomic Units (OTUs) using custom made implementation of OTUpipe (http://www.drive5.com/usearch/manual/otupipe.html) with USEARCH6 (35). Our implementation performs the following steps: i) de-replicate sequences; ii) de-noising by clustering at 99% identity; iii) cluster de-noised sequences at 97% to define OTUs; iv) identify chimeric sequences using both a reference-based and a de-novo chimera detection step. Sequences from Roche 454 were further scanned for chimeric OTUs using ChimeraSlayer (36) as implemented in QIIME (32). The number of reads matching a given OTU were counted for each sample and a count table was generated for each set of libraries (454, Illumina V4 with PNA, Illumina V4 without PNA, Illumina V8 and Illumina ITS2). Comparison of the Illumina V4 with PNA and V4 without PNA showed a very high degree of reproducibility (fig. S11) and thus the resulting count tables were combined to generate a single Illumina V4 count table.

f. Mapping MT consensus to isolate 16S genes: For the synthetic community experiments, every high quality consensus sequence produced by MT-Toolbox (34) was mapped with BWA version 0.7.10-r78 (37) to a reference set of sequences made up of the Sanger 16S sequence from the 38 isolates in the synthetic community, as well as to known plant nuclear and organellar rRNA genes. Up to 3 mismatches were allowed during mapping and the number of consensus sequences matching to each isolate or host sequences were used to create a count table for downstream analysis.

g. OTU and isolate annotation: We profiled the bacterial and the fungal communities by high-throughput sequencing of segments of the 16S rRNA gene and intergenic transcribed spacer (fig. S11). For each prokaryotic dataset (454 V8, Illumina MiSeq V8 an Illumina MiSeq V4), representative sequences from each OTU were given a taxonomic annotation using the RDP classifier (38) as implemented in QIIME 1.3.0. The 2011/-2/04 Greengenes database was used as a training set. OTU representative sequences were also BLASTed (39) against: i) a modified Greengenes database that includes plant and oomycete-derived sequences, and ii) the GOLD database (http://drive5.com/uchime/gold.fa). Any OTU annotated as plant, archaea or oomycete-derived (nuclear or organellar) by any of the three methods was removed from downstream analysis. For the fungal ITS dataset, OTUs were classified by BLAST against the UNITE database (https://unite.ut.ee/) which was modified to contain the *A. thaliana* nuclear and organellar ITS region.

Profiles of the strongly immunocompromised *jar1 ein2 npr1* (JEN) triple mutant and the *dde2 ein2 pad4 sid2* (DEPS; table S1) quadruple mutant contained a disproportionate abundance of sequences not classified as bacteria, despite our use of bacteria-specific 16S rRNA gene primers (table S2). The vast majority of these sequences corresponded to two Operational Taxonomic Units (OTUs) that were ~20bp shorter than bacterial amplicons, and matched mitochondrial sequences from the oomycete genera *Phythophthora* and *Pythium* (fig. S2). Our identification of oomycete sequences closely related to known plant pathogens is consistent with increased susceptibility of these mutant lines to infection (27, 28). Presumably as a consequence, both the JEN and DEPS mutants survived poorly on wild soil over the experimental time course, resulting in a lower number of replicates (table S3). Because oomycete prevalence and abundance were otherwise rare across samples (fig. S2), we removed these sequences during sequence processing (fig. S3) in order to focus on the alterations in the respective bacterial communities.

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

h. *In silico* pooling of samples in census experiments: In the 454 dataset, some DNA samples were barcoded and sequenced on multiple plates in an effort to achieve adequate depth. The resulting OTU counts from barcodes corresponding to the same original DNA sample were pooled (added) *in silico* after processing, but prior to any statistical analysis. Any barcode with 50 or less total reads was discarded, but samples that had between 50 and 1000 usable reads were matched with samples from the same experiment, fraction and genotype and, when possible, pooled to obtain samples with at least 1000 reads that were amenable to rarefaction. To allow for direct comparison between the Illumina and 454 datasets, samples that were pooled in the 454 dataset were also pooled in all the Illumina datasets regardless of their depth.

4) Microbial quantification procedures:

a. **CARD-FISH**: We used CARD-FISH (40) to show that the relative abundance decrease in Actinobacteria in the salicylic acid signaling mutant *pad4* EC samples compared to wildtype Col-0 EC controls was due to a decrease in the absolute number of metabolically active Actinobacteria in *pad4* EC tissue (fig. S8A). On the other hand, the relative abundance increase of Proteobacteria in *cpr5* roots was due to a lower total number of other types of metabolically active Eubacteria (fig. S8B).

We applied a previously described protocol (*10, 40*). Briefly, several root systems from bolting plants grown in Mason Farm soil were fixed using 4% formaldehyde in PBS at 4 °C for 3 h, washed twice in PBS and stored in 1:1 PBS:molecular-grade ethanol at −20 °C. Bulk MF soil, rhizosphere, and ground EC samples from 3 sets of Col-0, *cpr5*, or *pad4* samples were pooled and harvested as described above. Samples were made equal by mass and probe sonicated for 5 minutes in 30 sec bursts. The sample suspension was diluted 1:500 in water and applied to a 25-mm polycarbonate filter with a pore size of 0.2 mm (Millipore) using a vacuum microfiltration assembly. Filters were embedded in 0.2% low-melting point agarose and dried. Prepared filters were treated with lysozyme solution (1 h at 37 °C, 10 mg ml^{-1}; Fluka) and achromopeptidase (30 min at 37 °C, 60 U ml^{-1}; Sigma) and subsequently washed. Endogenous peroxidases were inactivated with methanol treatment amended by 0.15% H_2O_2 at room temperature for 30 min and washed again. Probes targeting either the 16S or the 23S rRNA of eubacteria (EUB338 (5′-GCTGCCTCCCGTAGGAGT-3′, 35% formamide), actinobacteria (HGC69a (5′-TATAGTTACCACCGCCGT-3′, 25% formamide), proteobacteria (1:1:1, ALF968 (5′ –GGTAAGGTTCTGCGCGTT- 3′, 20% formamide), (5′ -Bet42a (5′ –GCCTTCCCACTTCGTTT- 3′, 35% formamide), and Gam42a (5′ –GCCTTCCCACATCGTTT- 3′, 35% formamide)) and the negative control (NON338 (5′-ACTCCTACGGGAGGCAGC-3′, 30% formamide) were defined using probeBase (41), labeled with enzyme horseradish peroxidase on the 5′ end (Invitrogen), diluted in hybridization buffer (final concentration of 0.19 ng ml^{-1}) with each probe's optimum formamide concentration, and hybridized at 35 °C for 2 h. Unbound probes were washed away from samples in wash buffer (NaCl content adjusted according to the formamide concentration in the hybridization buffer) at 37 °C for 30 min. Fluorescently labeled tyramide was used for signal amplification, and samples were washed before mounting on glass slides. For double CARD–FISH, samples went through a second round of the protocol, starting at the peroxidase inhibition with a second variety of fluorescently labeled tyramide used to be able to distinguish the signals from each probe. Filter sections were mounted on glass slides using Vectashield with DAPI (Vector Laboratories, catalogue no. H-1200) for mounting solution, and sealed with nail polish for storage. All microscopy images were made on a Nikon Eclipse E800 epifluorescence microscope.

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

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

Reaction

5uL	Kapa Enhancer
5uL	Kapa Buffer A
0.4uL	5uM 338F
0.375uL	mixed PNAs (1:1 mix of 100uM pPNA and 100uM mPNA)
0.5uL	Kapa dNTPs
0.25	Kapa Robust Taq
8uL	dH2O
5uL	DNA

Temperature cycling
95 for 45 seconds
50 cycles of
- 95 for 15 seconds
- 78 (PNA) for 5 seconds
- 60 (338F) for 30 seconds
- 72 for 30 seconds

Bead clean

Following linear amplification, we performed the molecular tagging protocol as described previously (17), but substituting the tagged primer 806R (806R_f1-806R_f6, ST2) for tagged 926R (926R_f1-926R_f4, ST2). Primer 926R is universal (while 806R is bacteria specific) thus allowing to amplify nuclear 18S templates. For the forward primer we used the bc1 modification suggested by Lundberg et al 2013 (515F_bc1_f1-515_bc1_f6, ST2, 17). The molecular tagging protocol with this modification is:

Reaction

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

Temperature cycling
95 for 60 seconds
78 (PNA) for 5 seconds
60 (515F) for 60 seconds
72 for 60 seconds

Remove reaction from PCR block and add the following mix while on ice:

Reaction

0.4uL	5uM 926R TAGGED
1.6uL	dH2O

Temperature cycling
95 for 60 seconds
78 (PNA) for 5 seconds
50 (926R) for 60 seconds
72 for 60 seconds

Bead clean and proceed to exponential PCR.

Reaction

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

Temperature cycling
95 for 45 seconds
35 cycles of
    95 for 15 seconds
    78 (PNA) for 5 seconds
    60 (index primer) for 30 seconds
    72 for 30 seconds
    4 forever

We chose 192 samples covering all mutants in different experiments on MF soil, and we applied this protocol. There are only 96 index primers but we used combinations in the frameshift length of the molecule tagging to multiplex all 192 samples in one sequencing run, while keeping the average size the same:

Plate1: 515F_bc1_f1, 515F_bc1_f3, 515F_bc1_f5, 926R_f2, 926R_f4

Plate2: 515F_bc1_f2, 515F_bc1_f4, 515F_bc1_f6, 926R_f1, 926R_f3

After applying the dPCR protocol to these samples, we ran each reaction on an agarose gel to confirm the presence of two bands of the right sizes (one for the 16S and a larger one for the 18S). Then we pooled 3uL of each reaction into a master mix and bead cleaned twice eluting in 200 uL. This library mix was run on an agarose gel to confirm the presence of two bands of the right size and the absence of primer dimer. This library master mix was quantified with pico green (Quant-IT) and loaded into an Illumina MiSeq instrument (following the manufacturer's protocol) using a 50-cycle V2 chemistry kit.

Resulting sequences were demultiplexed and quality controlled with Sickle (42) by removing any sequence that had at least one base with a Q-score < 30. The remaining sequences were matched to a reference set that included the Arabidopsis 18S, Arabidopsis organellar 16S and the 17 most abundant bacterial sequences in the Greengenes database. No mismatches were allowed during this phase. After mapping the sequences, a ratio of bacterial 16S to plant 18S was calculated (Bactratio) and the results were analyzed with ANOVA and a post-hoc Tukey test using the "aov" and "tukeyHSD" functions in R. Results are presented in fig. S8C.

5) Synthetic community (SynCom) experimental procedures

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

b. Synthetic community experiments: In order to validate our sequencing results and associations found by the ZINB analysis of our census data, we performed three independent microcosm reconstitution experiments (table S3). Each experiment consisted of *A. thaliana* plants inoculated with a simplified synthetic mix of bacteria. *A. thaliana* seeds were surface-sterilized and germinated the same way as we did for the wild soil experiment (Methods 1c). Seedlings on MS plates were transferred to 2.5 inch square plastic pots (Kord Products Ltd.) containing (~100 mL) sterilized (autoclaved) calcined clay (Diamond Pro Calcined Clay Drying Agent, (http://www.diamondpro.com/Products/CalcinedClayDryingAgent) pots supplemented with 40% volume (~ 40mL) of ¼ MS (no sugar source), and inoculated with a mix of 38 bacterial strains (table S9) that could each be readily differentiated by 16S amplicon sequencing; these isolates were isolated from surface sterilized *A. thaliana* roots grown in either MF soil, or another previously characterized wild soil from Clayton, North Carolina, plus laboratory *E. coli* (table S9 (10)). Strains were selected from a set of isolates in order to maximize the number of strains with differentiable 16S genes so that they could be accurately quantified via 16S amplicon sequencing.

We applied exogenous salicilyc acid (0.5 mM) every 3 days to leaves and soil of additional plants as part of our 8-week synthetic community experiment. Roughly half of the samples for each experiment were sprayed with 0.5mM salicylic acid every 3 days, which is above physiological levels (43), but can induce systemic acquired resistance (20). This treatment can also induce runaway cell death in a mutant that is hyper-responsive to salicylic acid via activation of an immune receptor, *lsd1* (44). Under our synthetic community experiment control, this treatment elicited runaway cell death in control *lsd1* plants (45), but not Col-0 leaves 96 hours after spraying (fig. S12). Plant roots and bulk soil controls were harvested when an inflorescence meristem formed. Unlike the Mason Farm soil experiments, only EC and bulk soil fractions were collected due to the granular texture of the calcined clay that made rhizosphere harvest difficult. DNA from both bulk soil and EC samples was extracted using the MoBio PowerSoil kit. We utilized a recently published improvement of Illumina library preparation, which takes advantage of molecular tags (MT) to allow direct counting of original DNA templates in the sample, thus reducing PCR and sequencing errors and biases, as well as peptide nucleic acid to block amplification of host DNA (*17*). We sequenced the V4 region of the bacterial input (inoculum) as well as EC and bulk soil samples, with primers 515F and 806R (table S2) from three independent biological replicates.

c. Growth curves: Growth curves were performed in 1/10 LB with 0.1M phosphate buffer containing 0.01% yeast extract (46) and either 0, 0.125, 0.25, or 0.5mM salicylic acid added. 200 -well of media was inocula
plate and grown at 28°C shaking at 150 rpm. Optical density at 600 nm was measured every 2 hours for 50 hours of growth using a Synergy 2 multi-detection microplate reader (BioTek). Supernatants were harvested from liquid cultures of #273 or #303 grown in ½ and 1/10 LB with either 0, 0.125, 0.25, or 0.5 mM salicylic acid added after 0, 24, or 48 hours of growth and total salicylic acid was measured as described in Method 1f. No loss of total salicylic acid signal was detected for either culture in any media conditions (data not shown). For #303 growth on agar plate, minimal salts media ((NH4)2SO4 2g, K2HPO4 14g, KH2PO4 6g, sodium citrate 1g, MgSO4 0.2g per L) was supplemented with 0.5 mM salicylic acid in phosphate buffer, or phosphate buffer alone. #303 colonies were evident after 4 days of growth on this media and after 2 days of growth on LB.

6) Statistical analysis:

a. Diversity analysis for census experiments: Alpha and beta diversity were calculated on count tables that were rarefied to 1000 reads. Samples with less than this number of usable reads after pooling (see processing section) were discarded. Alpha diversity (Shannon index, richness, Simpson index) metrics were calculated using vegan (47), and differences between groups were tested with ANOVA. Beta diversity metrics were calculated with QIIME (UniFrac) or vegan (Bray-Curtis), and Principal Coordinate Analysis (PCoA) was performed with labdsv (48).

b. ZINB family and OTU-level analysis for census experiments: To determine which taxonomic groups associate differentially with each variable of interest, we took a linear modeling approach. We first collapsed OTUs assigned to the same bacterial family (see processing section), by aggregating their counts into a family-level count table. We decided to focus mainly on family-level abundances, because most of the data (i.e. the Roche 454 census experiments) is based on fragments of only 220bp, and it has been previously shown that only a small portion of sequences can be accurately given a genus level assignment (49), and it has been suggested that genus level assignments should only be performed with at least 250bp sequences (50). We also prefer family-level over OTU based analysis, because taxonomic families likely represent monophyletic groups while OTUs can be (and many are) paraphyletic (51).

Despite all of these drawbacks, we analyzed the OTU-level count table as well using exactly the same model specification that we used in the family-level analysis, and we observed similar trends (fig. S7). We previously found that only OTUs with at least 25 reads in at least each of 5 different samples, produce reproducible abundances, and we defined these as "measurable taxa" (*10*). We restricted our analysis to these measurable taxa, and applied a Zero-Inflated Negative Binomial (ZINB) model (fig. S9). A ZINB model (52) acknowledges that some proportion of the observed zeroes in the count tables might not be biologically meaningful, but rather experimental error (fig. S9, upper branch) and was therefore appropriate to use on our sparse family tables. At the same time, a ZINB model can focus on the variability associated with the variables of interest (fig. S9, lower branch). A ZINB model achieves its purpose by combining a classic count GLM with a "bad zero" generating process, and it links the two processes via a single parameter (p) that indicates the proportion (i.e. the probability) that a given zero is a "bad zero" (fig. S9; equation 1).

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

- π=probability of a bad zero (comtrols for sparsity)
- f_{nb}=Probability under Negative Binomial process, depends on β and α:
  - β = effects of interest and confounders
  - α = Controls for overdispersion

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

The implemented ZINB model depends on three parameters: i) p is the probability of a "bad zero", ii) k is an over-dispersion parameter that quantifies the deviation of the count process from the standard Poisson assumption of equality between mean and variance, and iii) a vector of coefficients β that quantifies the association of counts with each variable of interest. Each of these parameters has to be estimated in a full ZINB model, but p and k can be fixed to a set value by making extra assumptions and simplifying the model (fig. S9). It is impossible to say *a priori* whether the extra assumptions made by simple models are justified, so we fit each of the four models from fig. S9 on each family for each dataset, and then compared the model fits by means of the Akaike Information Criterion (AIC), which is a measure that combines the quality of the model fit while penalizing more complex models, so that extra parameters are only included when justified (53).

Each of the four models was fit with the same design matrix plus the natural logarithm of the number of usable reads per sample (*i.e.* depth) (see legend sheet on table S4 for details on the variables, and supplementary dataset SD2 for the design matrices), and the best model was chosen for each family on each dataset based on the AIC. The design matrix was constructed in a way that the genotype coefficients represent the difference with respect to Col-0 wildtype, and the fraction coefficients represent the difference with respect to bulk soil samples. The resulting coefficents (β) were tested for significance with z-tests and corrected for multiple testing with the Benjamini-Hochberg method (54). Model fits were performed with the stats (55) MASS (56) and pscl (57) packages in R.

c. Definition of a technically robust set of enrichments and depletions: We re-sequenced a subset of samples from a single experiment (3 soil samples, 3 Col-0 R samples, and 90 EC samples from nine genotypes, table S3) using the Illumina MiSeq platform and two different hypervariable regions of the 16S rRNA gene (fig. S11a). Four libraries were prepared (V8, V4 with peptide nucleic acid (PNA), V4 without PNA and ITS2 (see above and (*17*)). Each MiSeq lane was multiplexed to 48 samples. Sequences from each lane were run through the DOE JGI iTags pipeline at the DOE JGI for basic quality control. We used the same ZINB model approach on each dataset to identify family-level enrichments with respect to soil while controlling for batch effects within each platforms. We noted that the two V4 libraries (with and without PNA) gave identical results so we combined them into one abundance table. We also noted that the Illumina MiSeq gave much more consistent results, regardless of the variable region, than the Roche 454 instrument. Nevertheless, both platforms and all variable regions recapitulated the differences between EC and R samples and the alpha-diversity patterns (fig. S11). Further, we found that even when different 16S rRNA gene regions are assessed across sequencing platforms (MiSeq V4 vs. 454 V8), the correlation between taxonomic profiles is ~80% (fig. S11b). Finally, bacterial families that were enriched or depleted consistently in all three bacterial datasets (Illumina V4, Illumina V8 and Roche 454 V8) according to the ZINB model (Table S4) were considered to be technically robust (fig S11c-d), and represent a core set of enrichments and depletions that are insensitive to technical variation and thus are likely to represent true biological differences.

d. Comparison of enrichment profiles between genotypes: The ZINB model allowed us to identify the bacterial families that are enriched or depleted in the EC of specific plant genotypes with respect to Col-0. In order to compare the enrichment/depletion profiles between genotypes, each genotype is given a profile, which is a vector of numbers defined as following: each enriched family gets a value of 1, depleted families get a value of -1 and families that are not significantly different from Col-0 are given a value of 0. In this manner, each genotype gets an ordered vector of numbers, and such a vector can be compared directly to vectors of other genotypes. We chose the Manhattan distance, because given our definition of enrichment depletion profiles, only families that are different contribute to the distance metric, and families that have opposite effects between two genotypes (i.e. families enriched in one genotype and depleted in another) contribute more than families where the difference is between effect and no effect. Notably, DEPS EC samples had 52 DA families, nearly all of which were depletions (Fig. 1C; fig. S6). The decrease in alpha-Diversity observed in DEPS EC samples (Fig. 1B) likely reflects these depletions. This large number of depletions and low diversity in DEPS roots cannot be explained by their oomycete burden, since the equally oomycete-laden JEN EC samples exhibited only four DA families (fig. S6A), and only one of these was shared with DEPS.

Finally, to test whether the observed distances between genotypes are significant, we used a Monte Carlo procedure, by randomly permuting the order of the enrichment/depletion profiles 1000 times and re-calculating the Manhattan distance in each instance. This approach provides an empirical null hypothesis that can be compared to the value observed on the original data, and an empirical p-value can be calculated as the proportion of cases in the simulation that have distance values at least as extreme as the distance from the real data. The table of p-values is provided in figure S6c for the family level analysis, and figure S7b for the OTU level analysis.

e. PCA and CAP analysis of synthetic community experiments: For the synthetic community, the count table was rarefied to 400 consensus, and Principal Component Analysis was performed with the "princomp" function of R. Canonical Analysis of Principal Coordinates (CAP) (14) was performed using the "capscale" function of the vegan package (47) in R. CAP was performed on the full table of both the survey and the SynCom data and the constrained variation of fraction (fig. S13B) and salicylic acid (fig. S14A) was obtained after conditioning for every other technical and covariate. The proportion of variance explained by each variable (table S5), was estimated as the proportion of the total variation explained by the constrained axis of CAP, and confidence intervals were obtained by bootstrapping the taxa of the count tables for 1000 pseudoreplicates. For all of the CAP analysis, the CY Index, sometimes referred as Cao Index (58) was used as implemented in the "vegdist" function of the vegan package.

f. ZINB analysis of synthetic community experiments: For the synthetic community experiments, we repeated the ZINB analysis performed on the census datasets, but at the isolate level since we chose the isolates to have easily differentiable 16S sequences on the basis of Sanger sequencing of their 16S gene. The same four model structures were used, and AIC was used to decide on the best model. Hypothesis testing and multiple testing correction were done in the same manner as described above. The same software was utilized. A different design matrix (corresponding to the experimental design differences) was used, and the variables included are described in the legend sheet of table S4.

g. Genomic analysis of isolates in synthetic community experiments: Experimentally verified pathways that involve salicylic acid (SA, salicylate) were first obtained from MetaCyc (http://www.metacyc.org/). Five pathways were identified for SA degradation: salicylate degradation I, salicylate degradation II, salicylate degradation III, salicylate degradation IV, and enzyme salicylate 1,2-dioxygenase (accession number G-12243 MetaCyc). Of the two salicylic acid biosynthesis pathways, only one has evidence in bacteria (salicylate biosynthesis I) and so it was the only one used in our analyses. Among three other Streptomyces strains in the SynCom inoculum (#136, #299; table S9) and two additional Actinobacteria that were significantly associated with salicylic acid treatment prior to multiple testing correction (#29 and #362), the only obvious salicylic acid metabolism gene was an salicylic acid dioxygenase found in Arthrobacter sp. #362 (59, 60). The amino acid sequence of all the characterized genes in this reaction were retrieved from the databases linked by MetaCyc (table S10c) and were used to perform a BLAST searches against the predicted ORFs of the isolates' genomes. BLAST searches were performed on the IMG/ER webserver with default parameters. The results of the best hit (identity percent, and query coverage) are given in table S10a-b. Yellow color in table S10 indicates a good homolog hit while green indicates matching annotations between query and subject regardless of the hit quality.

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

References:

10. D. S. Lundberg *et al.*, Defining the core Arabidopsis thaliana root microbiome. *Nature* **488**, 86-90 (2012).
14. M. J. Anderson, T. J. Willis, Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology* **84**, 511-525 (2003).
17. D. S. Lundberg, S. Yourstone, P. Mieczkowski, C. D. Jones, J. L. Dangl, Practical innovations for high-throughput amplicon sequencing. *Nature methods* **10**, 999-1002 (2013).
18. D. Bulgarelli, K. Schlaeppi, S. Spaepen, E. Ver Loren van Themaat, P. Schulze-Lefert, Structure and functions of the bacterial microbiota of plants. *Annual review of plant biology* **64**, 807-838 (2013).
19. J. A. Vorholt, Microbial life in the phyllosphere. *Nature reviews. Microbiology* **10**, 828-840 (2012).
20. S. H. Spoel, X. Dong, Making sense of hormone crosstalk during plant immune responses. *Cell host & microbe* **3**, 348-351 (2008).
21. S. A. Bowling *et al.*, A mutation in Arabidopsis that leads to constitutive expression of systemic acquired resistance. *The Plant cell* **6**, 1845-1857 (1994).
22. J. D. Clarke, Y. Liu, D. F. Klessig, X. Dong, Uncoupling PR gene expression from NPR1 and bacterial resistance: characterization of the dominant Arabidopsis cpr6-1 mutant. *The Plant cell* **10**, 557-569 (1998)
23. V. Kirik *et al.*, CPR5 is involved in cell proliferation and cell death control and encodes a novel transmembrane protein. *Current biology : CB* **11**, 1891-1895 (2001).
24. Y. Zhang, S. Goritschnig, X. Dong, X. Li, A gain-of-function mutation in a plant disease resistance gene leads to constitutive activation of downstream signal transduction pathways in suppressor of npr1-1, constitutive 1. *The Plant cell* **15**, 2636-2646 (2003).
25. J. Dewdney *et al.*, Three unique mutants of Arabidopsis identify eds loci required for limiting growth of a biotrophic fungal pathogen. *The Plant journal : for cell and molecular biology* **24**, 205-218 (2000).
26. J. Glazebrook, E. E. Rogers, F. M. Ausubel, Isolation of Arabidopsis mutants with enhanced disease susceptibility by direct screening. *Genetics* **143**, 973-982 (1996).
27. K. Tsuda, M. Sato, T. Stoddard, J. Glazebrook, F. Katagiri, Network properties of robust immunity in plants. *PLoS genetics* **5**, e1000772 (2009).
28. J. D. Clarke, S. M. Volko, H. Ledford, F. M. Ausubel, X. Dong, Roles of salicylic acid, jasmonic acid, and ethylene in cpr-induced resistance in arabidopsis. *The Plant cell* **12**, 2175-2190 (2000).
29. M. D. Abramoff, Magalhaes, P.J., Ram, S.J., Image Processing with ImageJ. *Biophotonics International*, 3-42 (2004).
30. C. T. Defraia, E. A. Schmelz, Z. Mou, A rapid biosensor-based method for quantification of free and glucose-conjugated salicylic acid. *Plant methods* **4**, 28 (2008).
31. V. Kunin, A. Engelbrektson, H. Ochman, P. Hugenholtz, Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental microbiology* **12**, 118-123 (2010).
32. J. G. Caporaso *et al.*, QIIME allows analysis of high-throughput community sequencing data. *Nature methods* **7**, 335-336 (2010).
33. T. Magoc, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).

34. S. M. Yourstone, D. S. Lundberg, J. L. Dangl, C. D. Jones, MT-Toolbox: improved amplicon sequencing using molecule tags. *BMC bioinformatics* **15**, 284 (2014).

35. R. C. Edgar, Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).

36. B. J. Haas *et al.*, Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome research* **21**, 494-504 (2011).

37. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760 (2009).

38. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and environmental microbiology* **73**, 5261-5267 (2007).

39. C. Camacho *et al.*, BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421 (2009).

40. T. Eickhorst, R. Tippkotter, Improved detection of soil microorganisms using fluorescence in situ hybridization (FISH) and catalyzed reporter deposition (CARD-FISH). *Soil Biol Biochem* **40**, 1883-1891 (2008).

41. A. Loy, F. Maixner, M. Wagner, M. Horn, probeBase--an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic acids research* **35**, D800-804 (2007).

42. N. A. Joshi, Fass, J.N., Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). (2011).

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

44. V. Bonardi *et al.*, Expanded functions for a family of plant intracellular immune receptors beyond specific recognition of pathogen effectors. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 16463-16468 (2011).

45. D. H. Aviv *et. al.*, Runaway cell death, but not basal disease resistance, in lsd1 is SA- and NIM1/NPR1-dependent. *The Plant journal : for cell and molecular biology* **29**, 381-391 (2002).

46. T. R. Silva, E. Valdman, B. Valdman, S. G. Leite, Salicylic acid degradation from aqueous solutions using *Pseudomonas fluorescens* HK44: Parameters studies and application tools. *Brazilian Journal of Microbiology* **38**, 39-44 (2007).

47. J. Oksanen *et al.*, The vegan package. *Community ecology package*, (2013).

48. D. W. Roberts, labdsv: Ordination and Multivariate Analysis for Ecology (2013); available at https://cran.r-project.org/web/packages/labdsv/index.html.

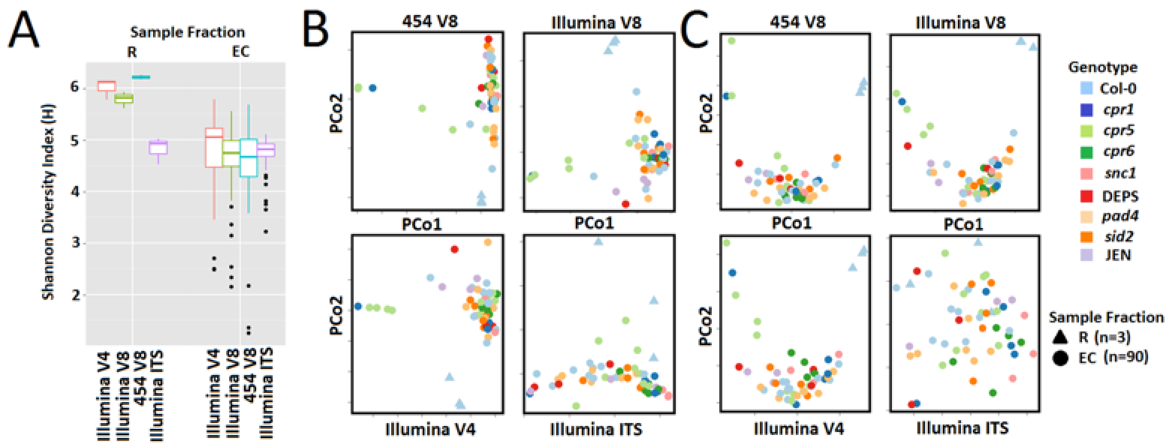
49. F. Guo, F. Ju, L. Cai, and T. Zhang, Taxonomic precision of different hypervariable regions of 16S rRNA gene and annotation methods for functional bacterial groups in biological wastewater treatment. *PloS One* **8**, e76185 (2013).

50. Z. Liu, T. Z. DeSantis, G. L. Andersen, R. Knight, Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic acids research* **36**, e120-e120 (2008).

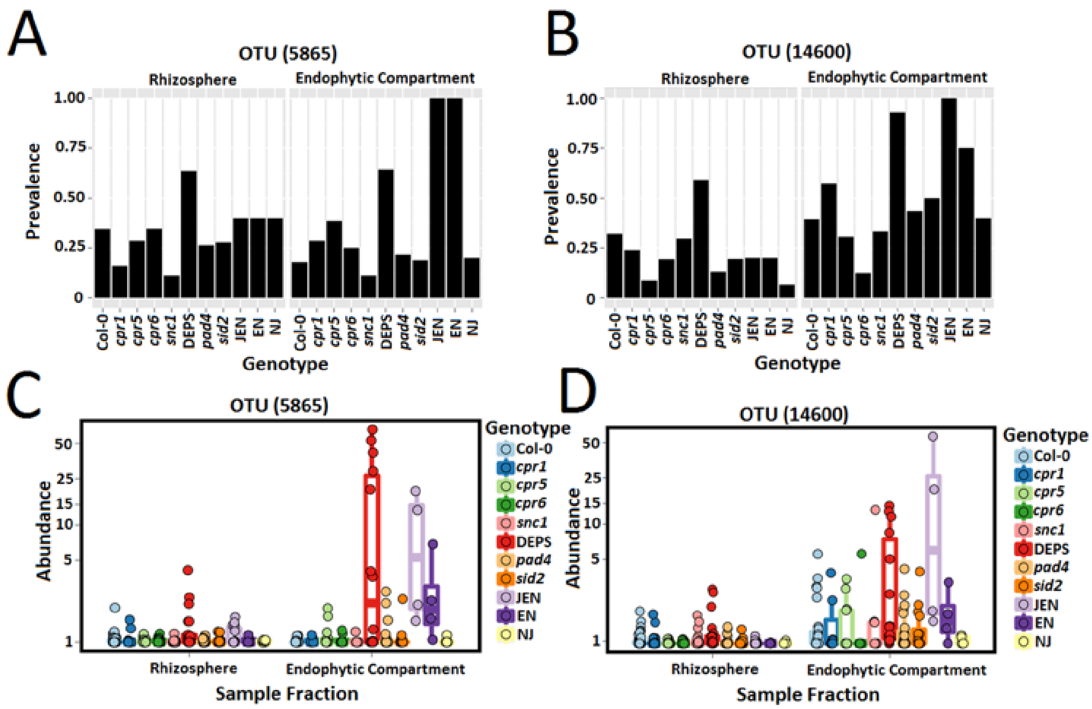
51. A. F. Koeppel, M. Wu, Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic acids research* **41**, 5175-5188 (2013).

52. A. F. Zuur, E. N. Ieno, N. J. Walker, A. A. Saveliev, G. M. Smith, in *Mixed effects models and extensions in ecology with R*. (Springer, 2009), pp. 261-293.

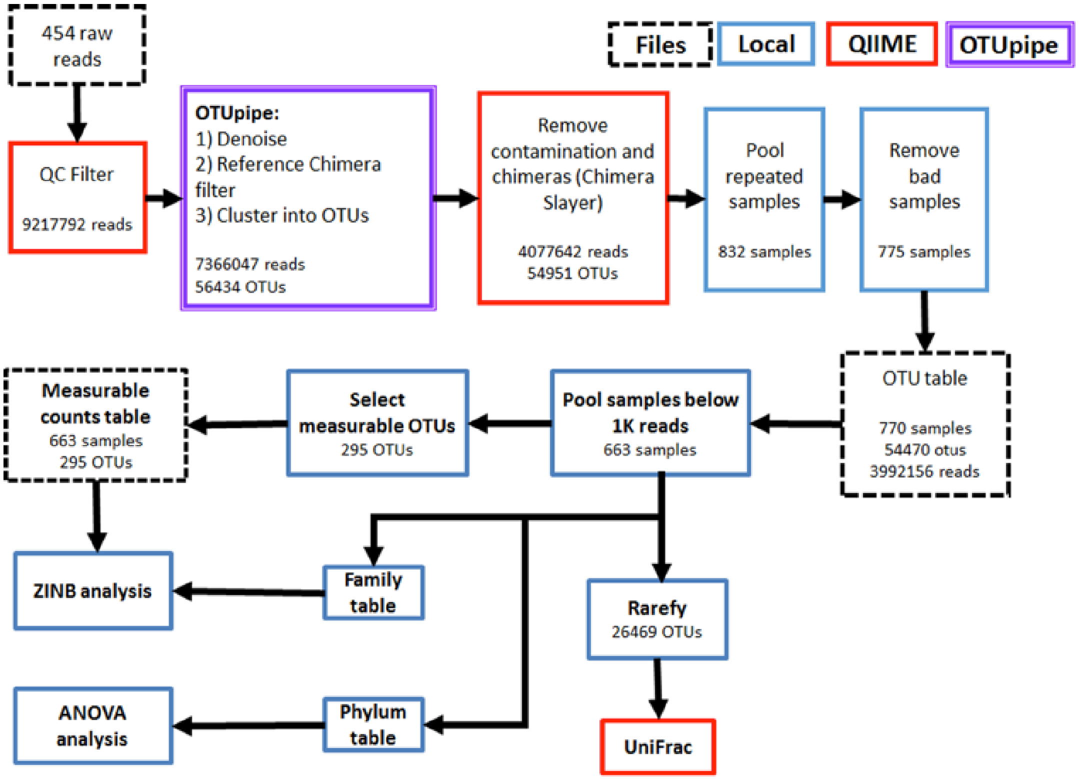
53. H. Akaike, A new look at the statistical model identification. *Automatic Control, IEEE Transactions on* **19**, 716-723 (1974).
54. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300 (1995).
55. R Core Team, R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, (2014).
56. W. N. Venables, B. D. Ripley, *Modern applied statistics with S*. (Springer, 2002).
57. A. Zeileis, M. A. Wiel, K. Hornik, T. Hothorn, Implementing a class of permutation tests: The coin package. *Journal of Statistical Software* **28**, 1-23 (2008).
58. Y. Cao, A. W. Bark, W. P. Williams, Analysing benthic macroinvertebrate community changes along a pollution gradient: a framework for the development of biotic indices. *Water Research* **31**, 884-892 (1997).
59. M. Ferraroni *et al.*, The salicylate 1,2-dioxygenase as a model for a conventional gentisate 1,2-dioxygenase: crystal structures of the G106A mutant and its adducts with gentisate and salicylate. *The FEBS journal* **280**, 1643-1652 (2013).
60. J. P. Hintner *et al.*, Direct ring fission of salicylate by a salicylate 1,2-dioxygenase activity from Pseudaminobacter salicylatoxidans. *Journal of bacteriology* **183**, 6936-6942 (2001).
61. H. Cao, J. Glazebrook, J. D. Clarke, S. Volko, X. Dong, The Arabidopsis NPR1 gene that controls systemic acquired resistance encodes a novel protein containing ankryin repeats. *Cell* **88**, 57-63 (1997).
62. J. M. Alonso *et al.*, Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*. **301**, 653-657 (2003).



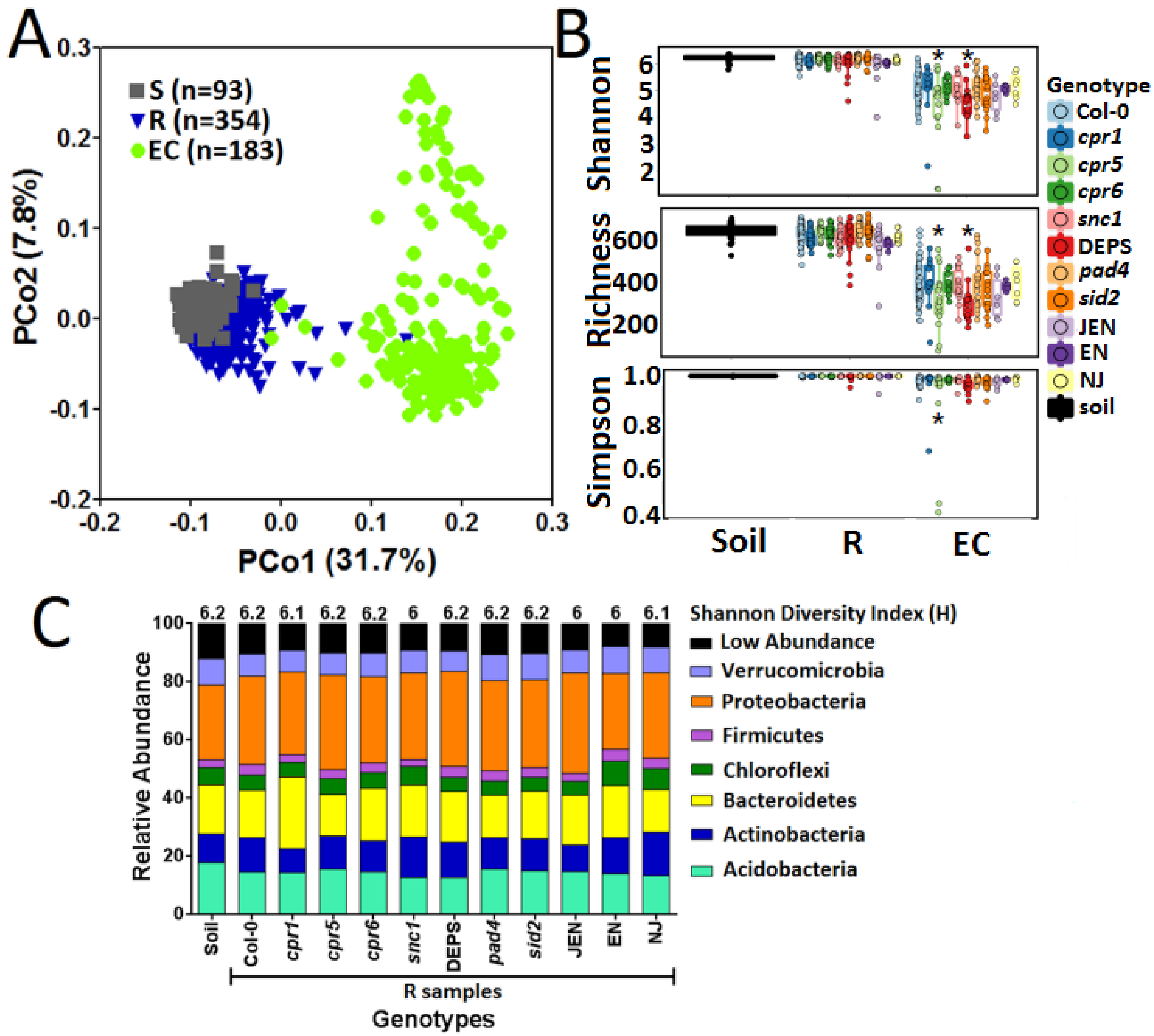
Supplementary Figure S1: Alpha and beta diversity for different 16S rRNA and ITS regions. (A) Shannon diversity index (H) for Illumina V4, Illumina V8, 454 V8, and Illumina ITS in both R and EC samples. Principal Coordinate Analysis of weighted UniFrac **(B)** and unweighted UniFrac **(C)** R (triangles) and EC (circles) samples sequenced by Illumina V4, llumina V8, 454 V4, and Illumina ITS demonstrates that bacterial profiles differ between R and EC samples regardless of sequencing platform and variable region. In contrast ITS profiles are remarkably similar (both in alpha and beta diversity) between R and EC samples.



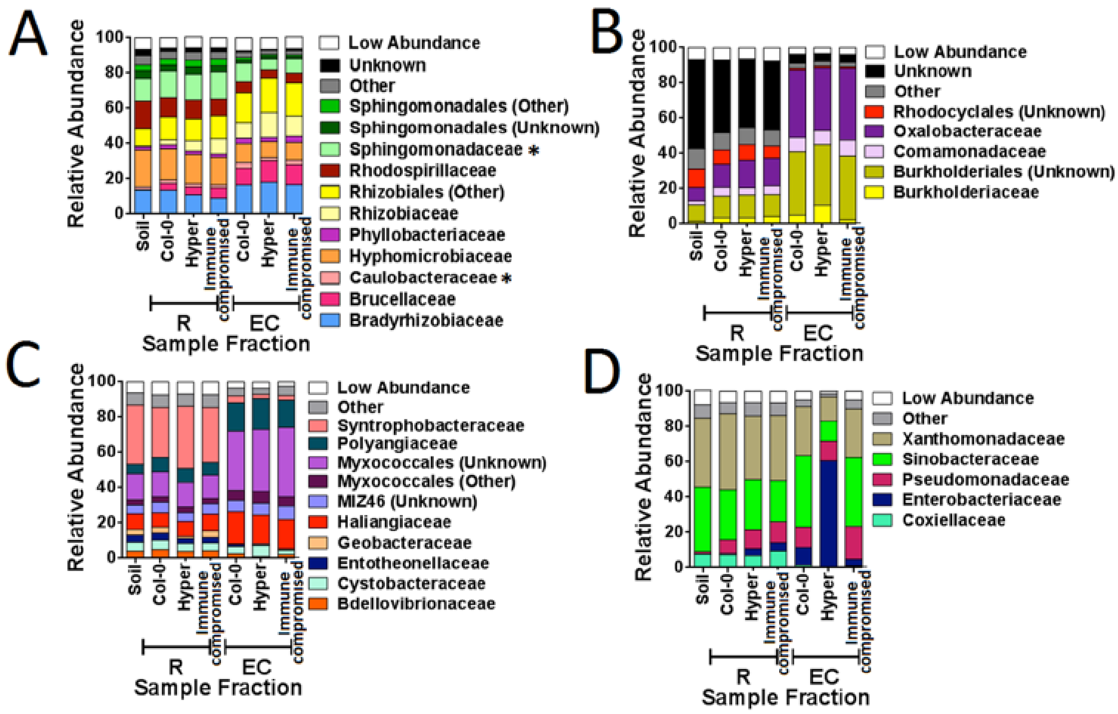
Supplementary Figure S2: DEPS and JEN root microbiome communities contain a disproportionate number of oomycete mitochondria reads. The prevalence of the top two OTUs matching oomycete mitochondria, OTU 5865 **(A)** and OTU 14600 **(B)** in each genotype. The percent abundance (over total non-plant reads) of OTU 5865 **(C)** and OTU 14600 **(D)** in Rhizosphere or Endophytic Compartment samples of each genotype is shown.



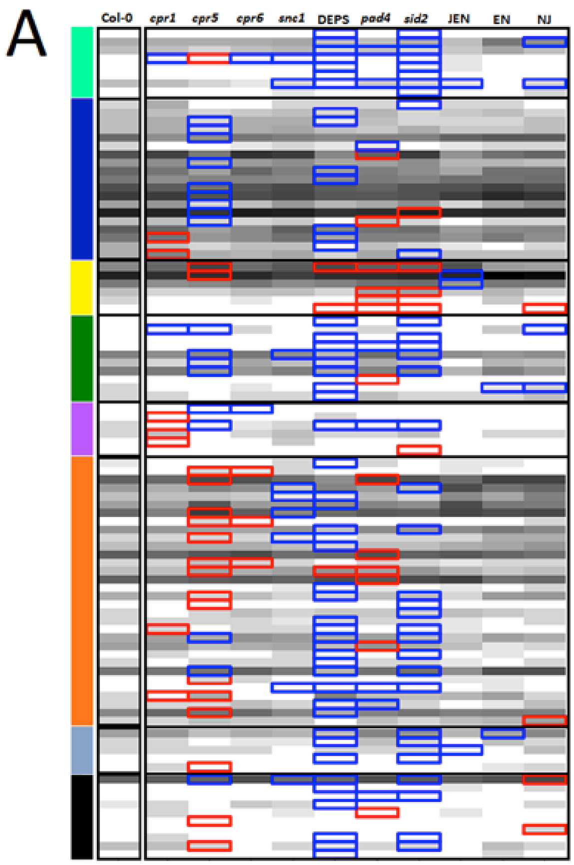
Supplementary Figure S3: Processing pipeline for Roche 454 census experiments. This flowchart is order of events that occur in processing the sequencing data. Boxes with dashed black lines represent files. Boxes with blue lines describe events that occur locally using custom scripts. Boxes with red lines describe steps that occur through QIIME. Boxes with double purple lines describe events that occur using OTUpipe. For full details see supplementary information (Method 3).

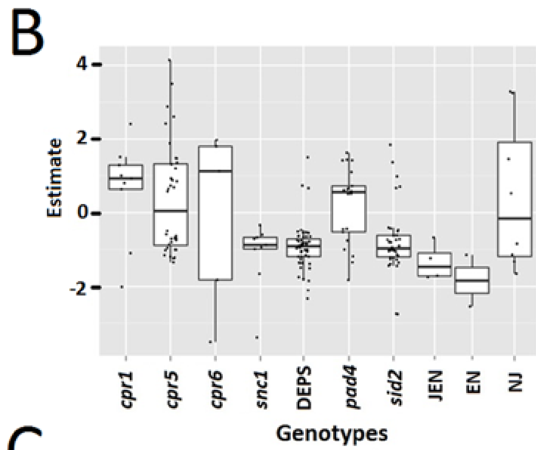


Supplementary Figure S4: Sample fraction drives differences in alpha and beta diversity of root microbiome communities. (A) Principal Coordinate Analysis (PCoA) of pairwise normalized, weighted UniFrac distances between the samples considering rarified to 1000 abundance of all OTUs. **(B)** Shannon diversity index, richness, and Simpson index for bulk soil, rhizosphere (R), and endophytic compartment (EC) samples for each genotype with the median represented by the bar and the 25th and 75th percentiles represented by the box. * indicates significantly lower than Col-0 EC samples at p<0.001 by ANOVA test with *post hoc* Tukey test. **(C)** Phyla distributions were separated into sample fractions (Soil or Rhizosphere) and plant genotypes. Shannon Diversity indices are listed above each bar. There were no significant differences in the Shannon Diversity or phyla abundances.



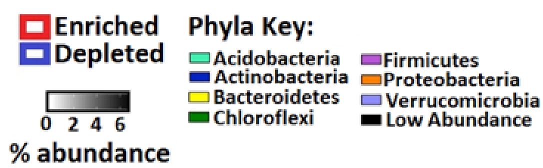
Supplementary Figure S5: Differential abundance of Proteobacteria families in different sample fractions. Relative abundance of Proteobacteria families in the alpha **(A),** beta **(B)**, delta **(C)** and gamma **(D)** orders in bulk soil, rhizosphere (R), and endophytic compartment (EC) sample fractions.* in **(A)** indicates that these families are significantly less abundant in EC-hyper samples compared to EC-Col-0 samples by ANOVA and *post hoc* Tukey's test, $p<0.05$.



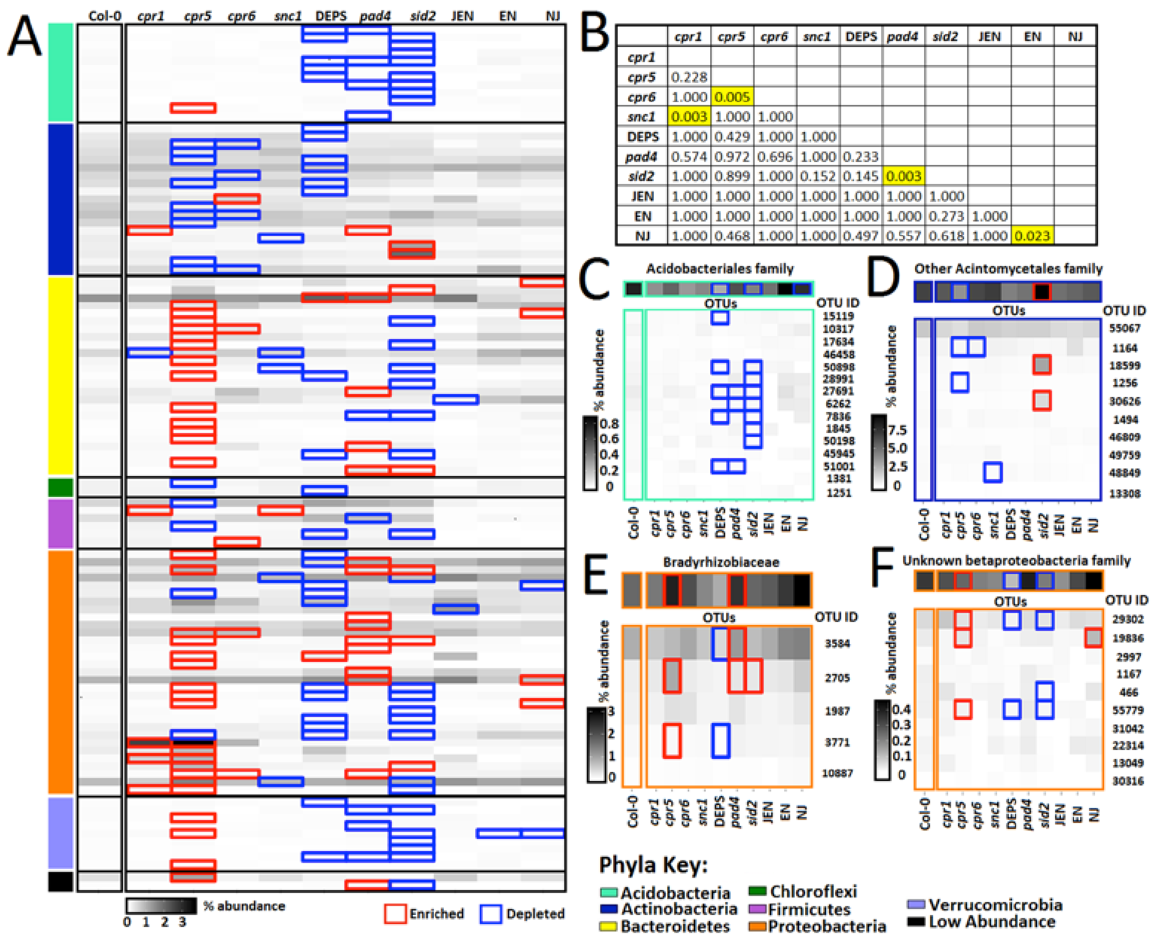


C

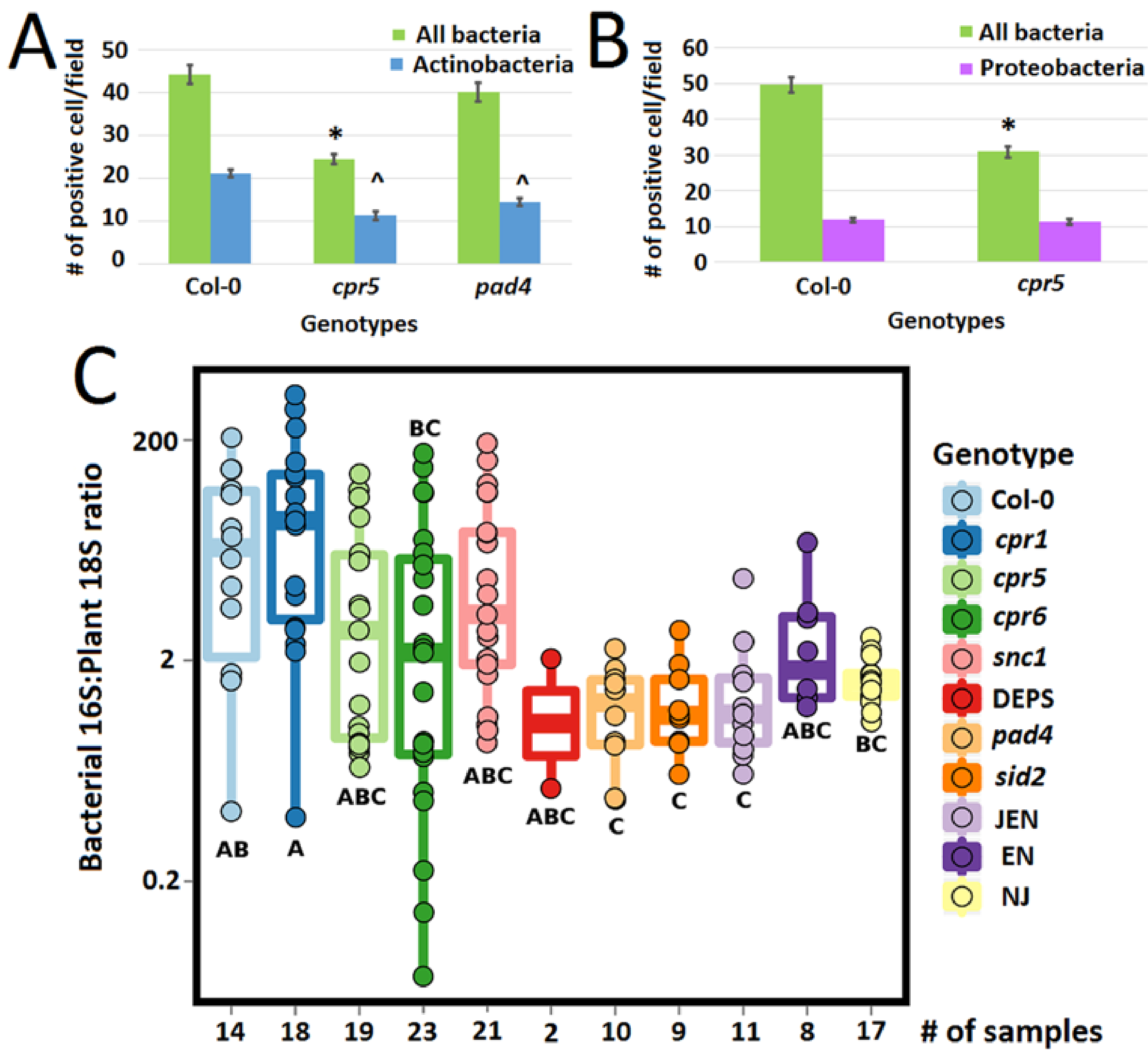
	cpr1	*cpr5*	*cpr6*	*snc1*	DEPS	*pad4*	*sid2*	JEN	EN	NJ
cpr1										
cpr5	0.544									
cpr6	0.249	0.003								
snc1	0.176	0.547	0.194							
DEPS	0.816	0.542	0.764	0.274						
pad4	1.000	0.553	1.000	0.158	0.002					
sid2	0.709	0.610	0.648	0.043	<0.001	<0.001				
JEN	1.000	1.000	1.000	0.335	0.948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



Supplementary Figure S6: Genotype-DA family enrichments and depletions: (A) Grid depicting the abundances for each family (grey scale), illustrating the overlap of differentiating families that are either enriched (red outline) or depleted (blue outline) in each mutant compared to the Col-0 abundances organized by phyla (indicated on the left side). **(B)** Dots represent families that are predicted to be significant by the ZINB model for each genotype compared to the Col-0 abundances. The magnitude of these predictions is represented by the estimate on the y-axis with enriched family represented by positive numbers and depletions represented by negative numbers. **(C)** Table of the p-values of the Monte Carlo testing of Manhattan distances between the enrichment and depletion profiles for each genotype pairing. The significance level for the pale yellow cell is $p<0.05$, while the significance level for the dark yellow cells are $p<0.003$.

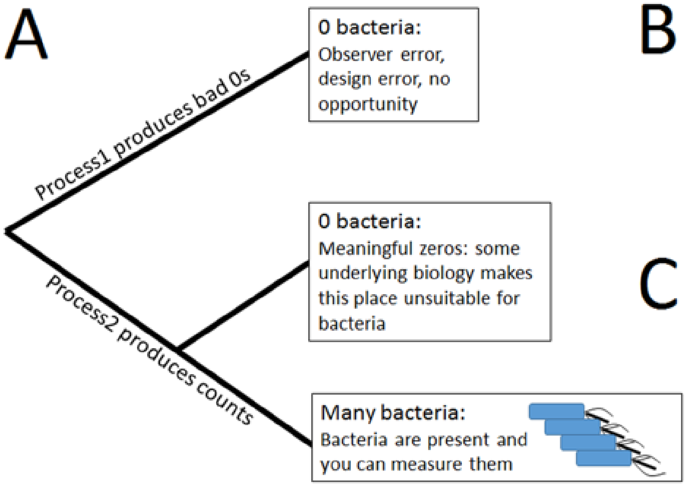


Supplementary Figure S7 - Genotype-DA OTU enrichments and depletions: (A) Grid depicting the abundances for each OTU (grey scale) illustrating the overlap of differentiating OTUs that are either enriched (red outline) or depleted (blue outline) in each mutant compared to the Col-0 abundances organized by phyla (indicated on the left side, color-coded to Figure 1). **(B)** Table of the p-values of the Monte Carlo testing of Manhattan distances between the enrichment and depletion profiles for each genotype pairing. The significance level for the yellow cell is p<0.05. The majority of families defined above are represented by only ‘measurable OTU’ so we focused on families with at least five measurable OTUs to address consistency between OTU and family level analyses. **(C-F)** Grids depicting the abundances of individual OTUs (grey scale), illustrating the overlap and consistency of differentiating OTUs that are either enriched (red outline) or depleted (blue outline) in each mutant compared to the Col-0 abundances within four families (grey scale above each grid) from figure S6: Acidobacteriales family **(C)**, other Acintomycetales family **(D)**, Bradyrhizobiaceae **(E)**, and unknown Beta-proteobacteria family **(F)**.



Supplementary Figure S8: The absolute quantification of bacteria in samples grown in MF soil. CARD-FISH results from EC samples applied to filters for counts (Method 4a), and were probed for metabolically active Eubacteria (green) bacteria and Actinobacteria (blue) **(A)** or Proteobacteria (purple) **(B)**. 20 fields were counted for each genotype with mean and standard error of the mean (s.e.m.) shown. * indicates significantly lower than Col-0 All bacteria counts ($p<0.001$). ^ indicates significantly lower than Col-0 Actinobacteria counts ($p<0.001$). **(C)** The ratio of bacteria 16S to plant 18S sequences in EC samples (Method 4b). A, B, and C labels denote results from a Tukey's HSD test. Genotypes that do not share any letters are statistically different.

A



B

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

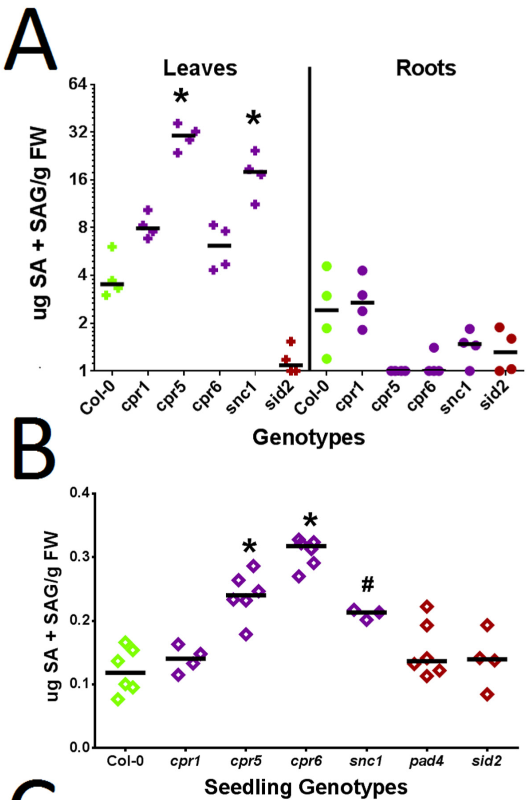
- π=probability of a bad zero (comtrols for sparsity)
- f_{nb}=Probability under Negative Binomial process, depends on β and α:
  - β = effects of interest and confounders
  - α = Controls for overdispersion

C

π	α	β	model
Estimate	Estimate	Estimate	ZINB
Estimate	0	Estimate	ZIP
0	Estimate	Estimate	GLM NB
0	0	Estimate	GLM Poisson

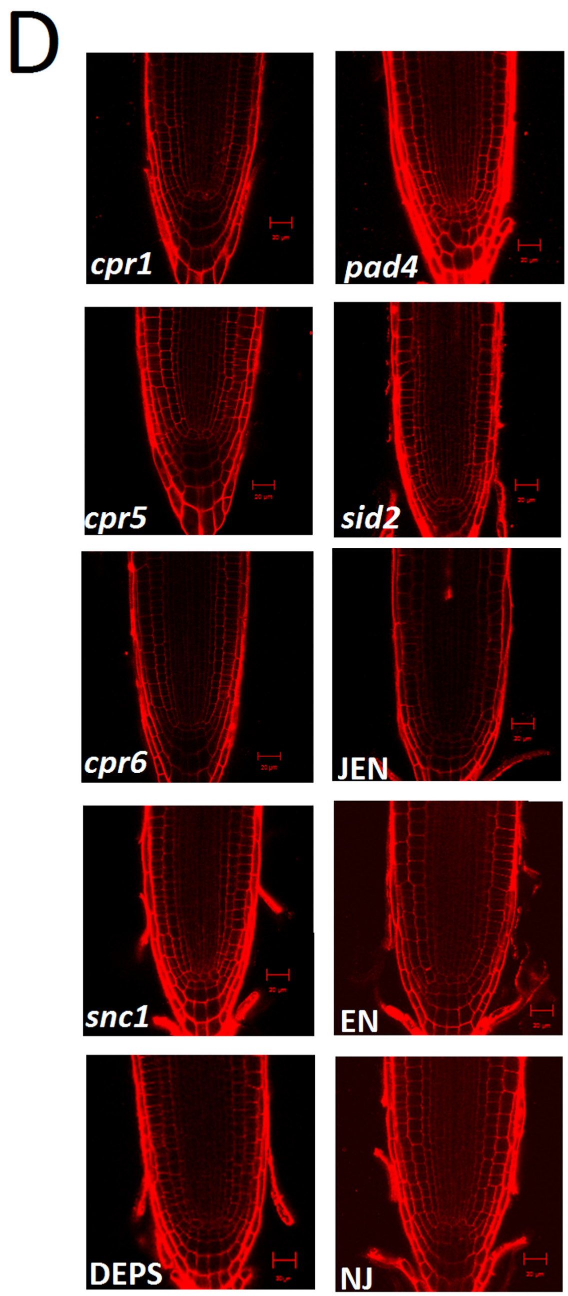
GLM NB is chosen ~70% of time and ZINB the remaining times

Supplementary Figure S9: Zero-Inflated Negative Binomial model. The rationale **(A)** and formula **(B)** for the ZINB model is shown. **(C)** The models tested which were tested with this data set.

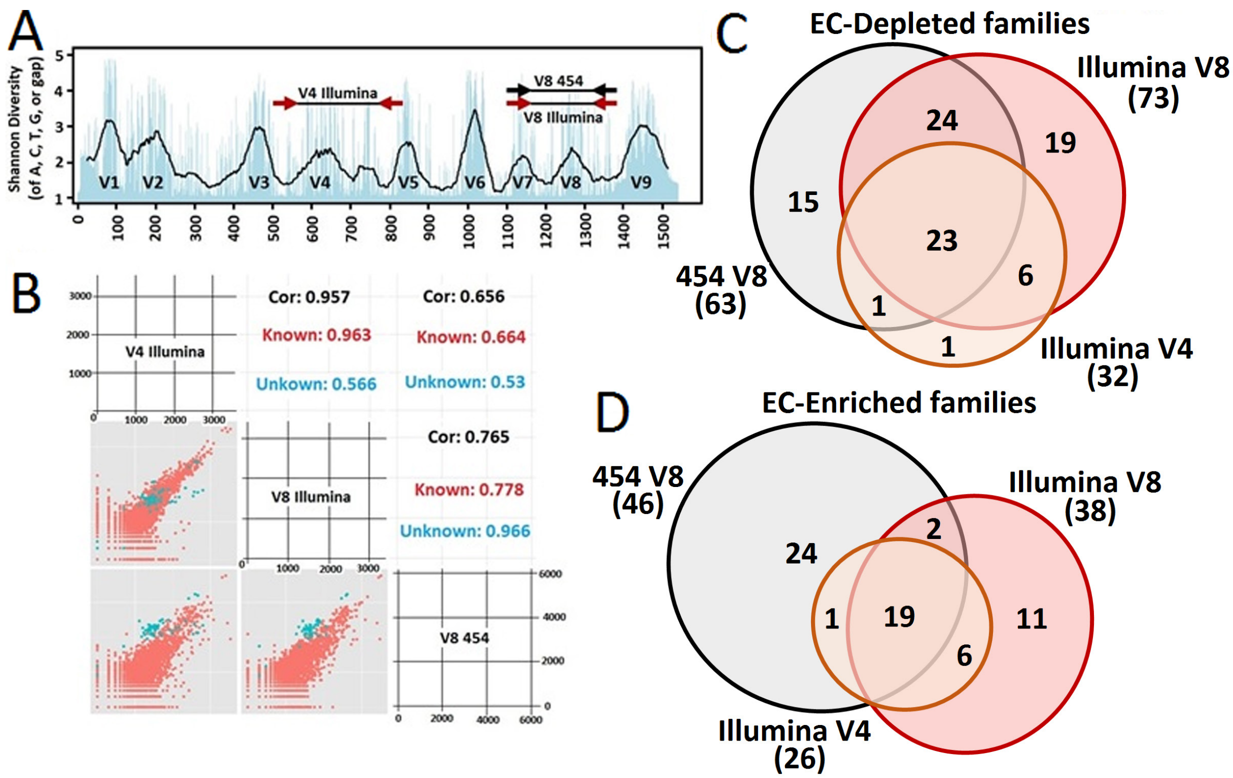


C

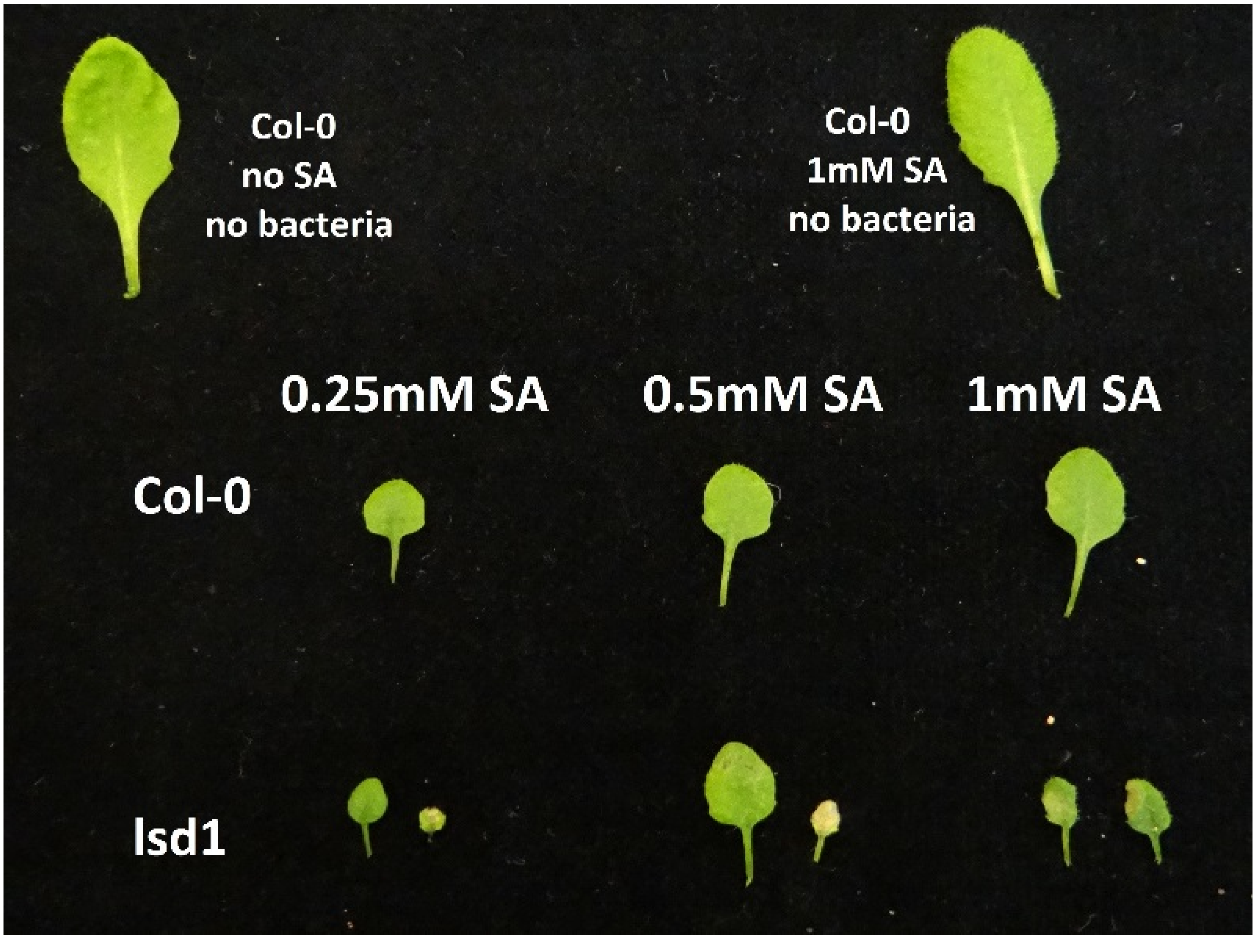
Genotype	Normalized root length	Root Morphology
Col-0	1 ± 0.1	Normal
cpr1	1 ± 0.1	Normal
cpr5	1 ± 0.1	Normal
cpr6	0.9 ± 0.2	QC & columella disorganization
snc1	1 ± 0.2	QC divisions
DEPS	1 ± 0.2	QC defects, possible early endo divisions
pad4	Not measured	QC & columella defects
sid2	1 ± 0.1	Normal
JEN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



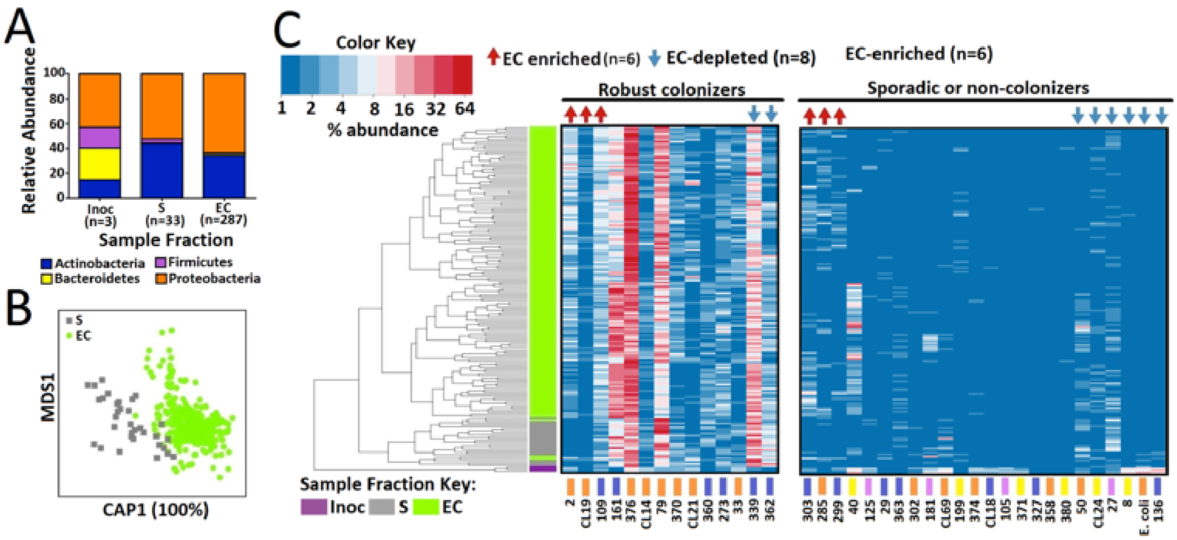
Supplementary Figure S10: Salicylic acid production in MF soil and root morphology of defense phytohormone mutants. (A) Representative of salicylic acid (SA) measurements performed three times in leaves and roots grown in MF soil (n= 4 for each type of sample, Method 1F). * indicates statistically higher than Col-0 ($p<0.0001$) by ANOVA with Bonferroni multiple test correction. **(B)** Representative of salicylic acid (SA) measurements performed on sterilely grown 18-day-old seedlings on ½ MS agar plates (n=3-6 for each type of sample, Method 1F). * indicates statistically higher than Col-0 ($p<0.0001$) and # indicates statistically higher than Col-0 ($p<0.005$) by ANOVA with Bonferroni multiple test correction. **(C)** Overview of root morphology at the root tip of each defense phytohormone mutant frown on ½ MS agar plates with representative images **(D)**, which are root tips stained with propidium iodide and observed with a 40x water objective (Method 1E).



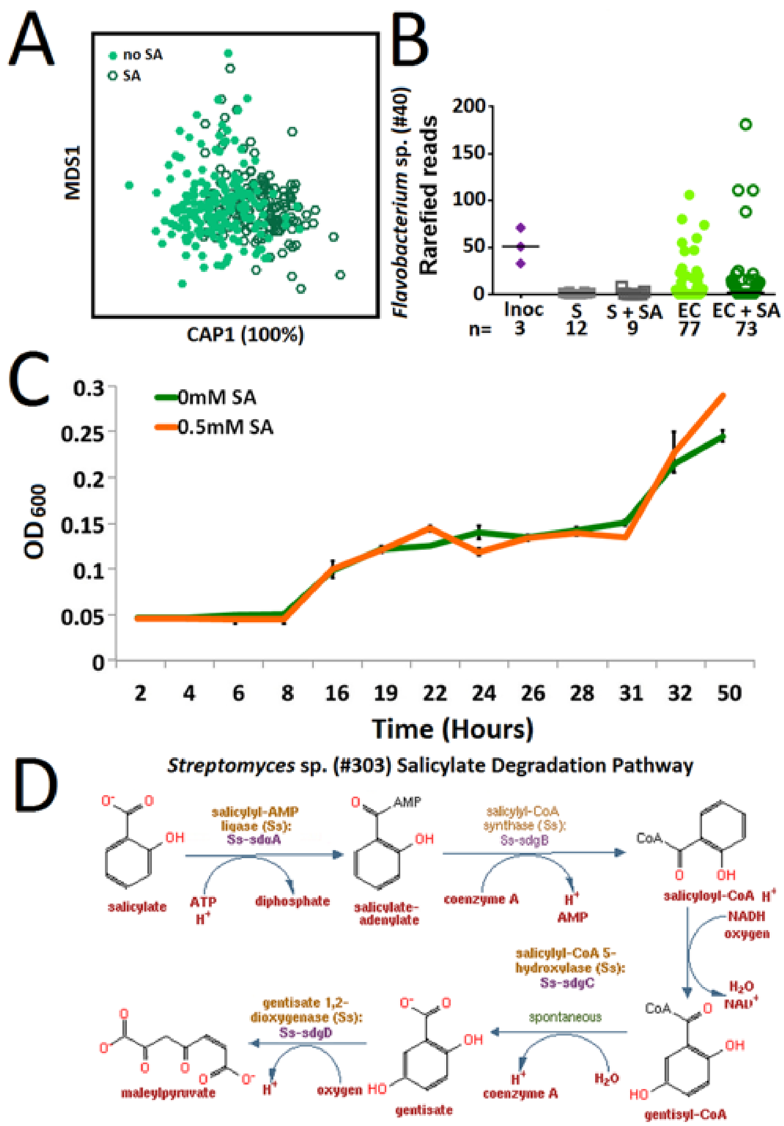
Supplementary Figure S11: Technical reproducibility between variable regions and sequencing platforms. (A) A schematic of the three 16S rRNA gene sequencing strategies used. **(B)** The reproducibility of family-level abundances between each sequencing strategy pairwise comparison for both taxonomically known (red dots) and unknown (blue dots) families with the calculated correlation. Venn diagrams showing the overlap of EC-depleted **(C)** and EC-enriched **(D)** families. The 19 EC-enriched and 23 EC-depleted families in all sequencing strategies are listed in table S8.



Supplementary Figure S12: Induction of Runaway Cell Death (RCD) in *lsd1* mutants grown in the SynCom with salicylic acid treatment of leaves. Col-0 and *lsd1* were grown in SynCom. 0, 0.25mM, 0.5mM, or 1mM salicylic acid (SA) was applied to their leaves. 96 hours later RCD was assessed (Method 5B).



Supplementary Figure S13 – Synthetic community differentiates sample fractions. (A) Phyla distributions in the synthetic community (SynCom) inoculum, soil, or EC fraction samples from all genotypes. **(B)** CAP analysis to showing the contribution of sample fraction to overall community composition. **(C)** Hierarchical clustering and heat map showing percent abundance (log_2 scale) of selected isolates. Sample clustering split by fraction (left), with EC samples grouping by biological replicate. Isolates are grouped by their presence in the majority of Col-0 EC samples (Robust colonizers) or absence in the majority of Col-0 EC samples (Sporadic or non-colonizers). Isolates color-coded to phyla as in Fig. a. Isolates that were significantly more abundant (red arrows) or less abundant (blue arrows) in EC with respect to bulk soil are denoted along the top.



Supplemental Figure S14: salicylic acid treatment affects SynCom composition, but did not affect growth of Flavobacterium #40 in SynCom or in liquid growth curves. (A) CAP analysis of the full count matrix to identify the contribution of salicylic acid (SA) treatment to community composition. **(B)** Dot plot of 400 rarefied consensus sequences from isolate #40 from synthetic community inoculum (purple diamonds), soil (grey squares), and EC samples (light/dark green circles) for both salicylic acid (SA) treated (open symbols) and untreated (closed symbols). No groups of samples were significantly different from any others. **(C)** Optical density of **isolate** #40 grown in phosphate buffered 1/10 LB with either 0 (green line) or 0.5mM (orange line) salicylic acid (SA) added. **(D)** Salicylate degradation pathway (MetaCyc) present in Streptomyces sp. (#303) genome contains all 4 genes in this pathway (% identities to each: sdgA-98%, sdgB-98%, sdgC-96%, and sdgD-94%).