
Functional capacities drive recruitment of bacteria into plant root microbiota

In the format provided by the
authors and unedited

Supplemental information - overview

Overview supplemental data and tables.....	3
Glossary.....	5
Supplementary Methods.....	6
Data availability.....	23
Methods – references.....	25
Supplemental figures and results.....	28
S1 – KO intravariability in SSC and simulated SynComs	
S2 – Experimental photos of inoculations	
S3 – Shoot weights	
S4 – <i>Lotus japonicus</i> nodule photos, counts, and profiles	
S5 – Rarefaction curves of SynCom member-pseudoaligned reads	
S6 – Microbiome profiles – inoculum	
S7 – Alpha diversity plot – Shannon diversity isolates and KOs – SynCom and host comparisons	
S8 – Microbiome profiles – bacterial genus	
<u>Supplemental results 1 – Nutrient condition effect on diversity analyses</u>	
S9 – Alpha diversity plot – nutrient condition subset	
S10 – Beta diversity plot – distance to centroids	
S11 – Bacterial family pangenomes and functional diversity correlation – family	
S12 – Functional diversity correlation – genus	
S13 – sPLS-DA PCoAs and box plots of contributing KOs to compositional differences	
S14 – DESeq2 number of significant KOs – overlap across hosts and inocula in dataset excluding nodulators or <i>Rhizobacter</i> sp. P2_G4	
<u>Supplemental results 2 – Host-specific functions</u>	
S15 – Lotus-specific nodulator functions – <i>nif</i>	
S16 – Lotus-specific nodulator functions – <i>nod</i>	
S17 – Lotus-specific functions – <i>ery</i>	
<u>Supplemental results 3 – Host-specific functions – ABC transporters</u>	
S18 – Host-specificity in ABC transporters	
S19 – DESeq2 – Root microbiome general pathway enrichment	
S20 – Fold change isolates with 266 KO pathways	
S21 – Niche replacement plant-specific and core pathways	
S22 – Bacterial colonization numbers LjSC Family drop out experiment	

Supplemental data and tables - overview

Data

01-Data related to the isolates' origin, taxonomy, and functions

- S1 – Source Data overview
- S2 – Phylogenetic tree of the SSC isolates
- S3 – AtSC (Stringlis *et al.* (2018)) & HvSC Microbiome profiles with SynCom member matches
- S4 – SSC isolate taxonomy, growth speed, and cultivation media
- S5 – SSC isolate genome lengths, completeness, and contamination (CheckM & BUSCO)
- S6 – 1,000 simulation SynComs – isolate composition
- S7 – KO profiles of bacterial isolates
- S8 – KO intravariability in SSC KOs
- S9 – KO intravariability in 1,000 simulation SynComs
- S10 – KO presence/absence in natural cucumber, wheat, and Arabidopsis soils and rhizospheres – Ofek-Ialzar *et al.* (2015), PRJNA435676, and PRJNA1108916.

02-Data related to the experimental set-up

- S11 – Overview of SSC samples and sequencing output
- S12 – Nutrient media composition used in SSC experiment
- S13 – Phenotypic data – shoot weights and *Lotus japonicus* nodule numbers
- S14 – SSC metadata
- S15 – SSC microbiome – isolate total abundances
- S16 – SSC microbiome – KO total abundances
- S17 – KO annotations, pathways, and pathway annotations

03-AtSC 7x200 SynCom experiment

- S18 – Metadata and SynCom composition
- S19 – Mapped reads overview
- S20 – Isolate abundance table
- S21 – Rank data functional diversity vs relative abundance

04-Data related to Figure 4

- S22 – Family impact on host and SynCom R^2 (including number of isolates and intrafamily functional variation)
- S23 – Raw sPLS-DA results

05-Data related to Figure 5

- S24 – DESeq2 output – root vs input
- S25 – KO fold change root vs input – General KOs (852 and 266 sets)

S26 – KO fold change root vs input – Host-specific KOs (3,075 KOs with stricter selection)
 S27 – KO fold change root vs input – non-nodulator dataset and Lotus-specific KOs
 S28 – Host-specific pathways – Frequency of the top 10 pathways
 S29 – Host-specificity of ABC transporters
 S30 – KO profiles of Arabidopsis-derived rhizobacteria grown in Cologne soil (Bai *et al.* (2015)
 S31 – Pathway fold change root vs input – General KOs (266 set)
 S32 – Top 70% and 90% cumulative relative abundance SynCom members per sample group
 S33 – Barnard statistic of intrafamily 266 KO overrepresentation top vs non-top

06-LjSC Family drop out SynCom experiment

S34 – Metadata and SynCom composition
 S35 – Mapped reads overview
 S36 – Isolate abundance table
 S37 – KO abundance table
 S38 – DESeq2 output – root vs input
 S39 – Fold change pathways - Plant-specific pathways
 S40 – Fold change pathways - Core pathways

07-Data related to publicly available data

S41 – Isolates used in Levy *et al.* (2018)
 S42 – KOs that were significantly plant-associated in Levy *et al.* (2018)
 S43 – Overview of the 266 KOs in Levy *et al.* (2018)
 S44 – Enriched KOs in natural rhizospheres in Lopez *et al.* (2023)
 S45 – Enriched KOs in natural Arabidopsis rhizospheres (PRJNA435676 and PRJNA1108916)

08-Shiny app

S46 – Shiny app – original dataset
 S47 – Shiny app – dataset without dominators

Tables

S1 – Physicochemical composition of soils from which the bacterial isolates derive
 S2 – KOs in natural rhizospheres vs soils.
 S3 – Abundance distribution of computationally simulated microbiome samples
 S4 – Statistics of deviation associated with Extended Data 2
 S5 – Taxonomic and functional R^2 values across datasets also with SynCom subsets, host drop-outs, and across taxonomic levels
 S6 – Statistical randomization test functional diversity vs relative abundance
 S7 – Rank test - functional diversity vs relative abundance – AtSC 7x200 SynComs
 S8 – Number and percentages of double selected KOs between inocula
 S9 – KO's contributed and lost – LjSC Family drop out experiment
 S10 – 266 and 852 KOs in natural rhizospheres and soils.

Glossary

Culture collection – Collections of pure bacterial cultures, commonly preserved in glycerol stocks. The culture collections are referred to as AtSC, LjSC and HvSC for Arabidopsis, Lotus and Barley, respectively.

Family representativeness – An indication on how well the isolate represents the family in terms of functionality, reflected by the **family KO proportion**: the percentage of unique KOs that a bacterial isolate has from the total amount of unique KOs that can be found in the family (pangenome).

Functional convergence/divergence – The process in which the composition of bacterial functions in the root microbiome between hosts becomes more similar (convergence) or dissimilar (divergence).

Functional diversity – An indication of a bacterial isolate's functional capacity and versatility. A high number of unique functions indicates a high functional diversity.

Genome completeness – An overview of the presence and absence of conserved bacterial markers in the genome, indicating how complete the genome is assembled from the short reads.

KOs – Functional gene annotations from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

KO diversity – The number of unique KOs that an inoculum or bacterial isolate has, which is indicative of functional capabilities.

KO intravariability/gene diversity – The diversity that exists within a KO. A high KO intravariability indicates the occurrence of multiple genes that have the same KO annotation but diverging nucleotide sequences, which could indicate multiple different functions within the KO. A low KO intravariability might indicate fewer genes in a KO or the presence of (near-)identical gene copies, suggesting a similar functionality.

KO/gene branch length – The cumulative branch length from a dendrogram of all genes within a KO, constructed based on nucleotide sequence identity. A high cumulative branch length indicates high intravariability within the KO, reflecting a greater number of distinct gene sequences.

Microbial context – An observation made in the microbial context; means it can only be observed in this specific microbiome (inoculum) composition and not in others.

PERMANOVA R^2 value – The proportion of compositional differences explained by a metadata variable. When investigating the microbiome or functional composition of samples, i.e. within-sample counts of bacterial isolates or bacterial functions, we evaluate how similar or distinct samples are by looking at the compositional differences. When samples derive from different hosts or are inoculated with different inocula, for example, we can calculate how much the composition is affected by these variables (the hosts or inocula), by calculating the R^2 value for each variable.

SynCom reconstitution experiment – Experiment in which bacterial isolates are individually cultivated and assembled into a “synthetic” community in similar proportions. The SynCom or inoculum members are most commonly representing the natural community. Inoculum and SynCom are synonyms.

Methods

Source Data overview for all main figures and extended data can be found in Data S1.

Culture collection establishment and genome assembly - AtSC

The *Arabidopsis thaliana* (hereafter Arabidopsis) culture collection was isolated from Arabidopsis Col-0 plants and *myb72* mutant plants grown in iron-deficient natural soil (Reijerscamp soil, the Netherlands) (Berendsen et al., 2018; Selten et al., 2024; Stringlis et al., 2018) (Table S1, Data S3). The rhizospheres of the Arabidopsis plants were harvested at three days and at seven days post-transplantation to the soil system by removing the roots from the soil and shaking vigorously to only keep tightly adhering root-associated bacterial isolates. The roots were subsequently washed in 10 mM MgSO₄, diluted 10⁶ times and plated on 1/10th tryptic soy agar, Reasoner's 2A agar, yeast extract medium, minimal medium with 2 mM coumarin, potato dextrose agar and King's B medium (Data S4, S12). Single colonies were picked three, five, seven, ten, fourteen and twenty-one days after plating and individually cultivated in liquid tryptic soy broth.

Bacteria that were cultivated at three days post-transplantation were subjected to high-throughput two-step barcoded amplicon sequencing of the V3-V4 region of the *16S rRNA* gene following a triple indexing strategy as described in (de Muinck et al., 2017). In the first PCR, the V3-V4 region was amplified using primers 341F and 785R (Klindworth et al., 2013) alongside an indexing sequence and a partial Illumina adapter. During the second PCR a third indexing sequencing is appended in addition to completion of the Illumina sequencing adapter:

PCR1-Forward: 5'-TCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-[i7]-GTGYCAGCMGCCGCGGTAA-3'

PCR1-Reverse: 5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-[i5]-GGACTACNVGGGTWTCTAAT-3'

PCR2-Forward: 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGAC-3'

PCR2-Reverse: 5'-CAAGCAGAAGACGGCATACGAGAT-[i5]-GTGACTGGAGTTCAGACGTG-3'

After both PCR steps, amplicons were purified using Agencourt AMPure XP beads, and pooled and sequenced on an Illumina MiSeq platform. The number of bacterial isolates was narrowed down according to uniqueness in the V3-V4 *16S rRNA* sequence using the DADA2 plugin of Qiime2 (Qiime2 version 2019.1) (Bolyen et al., 2019). Additional bacterial isolates that were isolated from plants at seven days post-transplantation were inspected for colony morphological characteristics such as size, color, shape, opacity, elevation, stickiness and shininess, and morphologically novel ones were added to those from three-day soil-grown plants to build up the Arabidopsis culture collection.

Subsequently, DNA was extracted from pure bacterial cultures using the MagAttract Microbial DNA kit and genomic fragments were amplified using the degenerate Nextera barcoded i5 and i7 primers following the Hackflex protocol (Gaio et al., 2022):

Forward: 5'-AATGATACGGCGACCACCGAGATCTACAC-[i5]-TCGTCGGCAGCGTC-3'

Reverse: 5'-CAAGCAGAAGACGGCATACGAGAT-[i7]-GTCTCGTGGGCTCGG-3'

Amplified genomic fragments were purified using the AMPure XP beads, pooled and sequenced on an Illumina NovaSeq platform. Genomic reads were demultiplexed and cleaned with cutadapt (version 2.8) (Martin, 2011) and assembled into genomes using A5 (A5-miseq version 20160825) (Coil et al., 2015). Genome completeness, contamination, and heterogeneity (collectively, purity) was checked with CheckM (version 1.1.3) (Parks et al., 2015) and BUSCO (Nextflow version 25.10.3) (Seppey et al., 2019) (Data S5) and any genomes with multiple single copy gene occurrences were subjected to a metagenomic binning procedure (MaxBin version 2.2.7) (Wu et al., 2014) to disentangle genomes from these apparent mixed bacterial cultures. Non-bacterial contigs in the genome assemblies were removed using MMSeqs2 (version 13.45111) (Steinegger & Söding, 2017). Next, dRep (version 3.4.1) (Olm et al., 2017) was run on the 563 genomes to dereplicate the genomes to a non-redundant set of 447 unique genomes at 99.99% sequence identity. Open reading frames were identified and annotated by PROKKA (version 1.14.6) (Seemann, 2014) and EggNOG (version 2.1.4-2) (Cantalapiedra et al., 2021), respectively (Data S7). The isolates' taxonomy was inferred using the Genome Taxonomy DataBase (GTDB, version v2.3.0) (Chaumeil et al., 2020) (Data S4).

Culture collection establishment and genome assembly - HvSC

The *Hordeum vulgare* (hereafter Barley) culture collection is a cereal culture collection isolated from both Barley (Golden Promise) and *Zea mays* (hereafter Maize) (W22) roots grown in Askov soil (Denmark) (Overgaard et al., 2022) (Table S1, Data S3). Bacteria associated with Barley and Maize roots were isolated at seven weeks post germination. The roots were sampled, excess soil particles were removed, and the remaining roots were shaken vigorously in sterile water. The bacteria-soil suspension was diluted 2×10^6 in $1/10^{\text{th}}$ tryptic soy broth or in yeast extract medium and distributed across 96-well plates (Data S12). Pure bacterial cultures were isolated at seven- and fourteen-days post-sampling.

From these bacterial suspensions, the DNA was extracted and subjected to high-throughput two-step barcoded amplicon sequencing on the V5-V7 region of the *16S rRNA* gene. High throughput DNA extraction was performed by adding a basic buffer to the bacterial suspension (100 mM NaOH and 100 mM EDTA at pH = 12) followed by a thermal lysis of the bacterial cells and the addition of a pH buffer (40 mM Tris-HCl at pH = 7.46). The V5-V7 region of the *16S rRNA* gene was amplified in both PCRs by degenerate 799F and 1192R primers (Bulgarelli et al., 2012; Wippel et al., 2021):

Forward: 5'-AATGATACGGCGACCGAGATCTACAC-[i5]-GACTGCGACTGGCG-ACMGGATTAGATACCCCKG-3'

Reverse: 5'-CAAGCAGAAGACGGCATACGAGAT-[i7]-CAGCCATTTAGTGTC-ACGTCATCCCCACCTTCC-3'

The amplified amplicons were cleaned by Agencourt AMPure XP beads after the first and second PCR, pooled, and sequenced on an Illumina HiSeq platform. The number of bacterial isolates was narrowed down according to uniqueness in the V5-V7 *16S rRNA* sequence using the USearch plugin of Qiime1 (Qiime1 version 1.9.1) (Caporaso et al., 2010).

Thereafter, the bacterial isolates were recultivated on tryptic soy agar and yeast extract medium plates to isolate pure bacterial colonies and remove any contamination. Next, the DNA was extracted from these liquid cultures for whole-genome sequencing according to Gonzalez-y-Merchand et al. (1996). DNA library preparation, whole genome sequencing, genomic read filtering and clean-up, genome assembly, genome assembly quality check, functional annotation and taxonomy inference were performed as described for the Arabidopsis culture collection (Data S4, S5, S7). The final version of the

Barley culture collection consists of 392 isolates dereplicated to 364 unique genomes at 99.99% sequence identity.

Culture collection establishment and genome assembly - LjSC

The *Lotus japonicus* (hereafter Lotus) culture collection derives from Lotus Gifu plants cultivated in Cologne Agricultural Soil (CAS) (Germany) (Bulgarelli et al., 2012) (Table S1). An extensive description of the isolation, sequencing, and genome assembly of the Lotus culture collection can be found in Wipfel et al. (2021). The Lotus culture collection is publicly available under the accession number [PRJEB37696](#). Additionally, we used MMSeqs2 (version 13.45111) (Steinegger & Söding, 2017) from any non-bacterial contigs, similarly to the Arabidopsis and Barley culture collections, and used dRep (version 3.4.1) (Olm et al., 2017) to dereplicate the isolate set of 290 isolates to 177 unique genomes at 99.99% sequence identity. We also adjusted the taxonomy of the bacterial isolates using GTDB (version v2.3.0) (Chaumeil et al., 2020), assessed genome assembly quality, and functionally profiled the isolates (Data S4, S5, S7).

SynCom preparation

All bacterial isolates from the Arabidopsis (563), Lotus (290) and Barley (392) culture collections were cultivated at nine, seven and five days prior to inoculation in 1/10th tryptic soy broth or yeast extract medium broth depending on their growth speed (Data S12). Deep-well cultivation plates (2 mL) were filled with 1 mL of bacterial growth media and 10 µL bacterial suspension, sealed with breathable membranes, and incubated on a shaker at 21°C.

On the day of the inoculation, bacterial isolates were washed twice by spinning the deep-well plates at 4,700 rpm for 8 min, removing the supernatant, adding 600 µL of low nutrient Long Ashton medium (SLA) (Data S12) and resuspending the bacterial culture. The bacterial cultures were subjected to OD₆₀₀ measurement and pooled per host-specific inoculum in similar proportions at an OD₆₀₀ of 0.1. The pooled AtSC, HvSC and LjSC were subsequently subjected to another OD₆₀₀ measurement, extraction of a small aliquot to facilitate analysis of the input samples, pooling, and finally dilution to an OD₆₀₀ of 0.02 using high nutrient SLA, low nutrient SLA or Hoagland. The AtSC, HvSC and LjSC were pooled in equal volume and concentration ratios to create the SSC. Equalising hundreds of isolates by OD₆₀₀ is inherently imprecise, leading to uneven starting abundances; biases during DNA extraction, library preparation, and sequencing depth can cause low-abundance strains to fall below the detection threshold, hence we observed that the number of strains in the input deviates slightly from the number of strains included in the inoculum (Figure 2b-c).

Seed surface-sterilization and germination

Arabidopsis seeds were surface sterilized in 70% ethanol for two minutes, washed four times in sterile water, submerged in sterile water and stored in 4 °C in the dark for three days to promote uniform germination through seed imbibition and stratification. Germination was induced by placing the imbibed and stratified seeds on a Murashige & Skoog agar (MS) plate (Data S12), which was then placed vertically at 20°C with a 16/8 hour light/dark cycle nine days prior to inoculation.

Barley seeds were surface sterilized in 1:20 bleach (NaClO) solution for 15 minutes, followed by five washes with sterile water and an hour incubation in sterile water on a moving tray for seed imbibition.

Subsequently, Barley seeds were placed in between sterile filter paper sheets that were soaked with sterile water previously and stored in sealed-off petri dishes horizontally at 20°C with a 16/8 hour light/dark cycle five days prior to inoculation.

Lotus seeds were scarified and surface sterilized by sandpaper scrubbing and submerging them in 1:20 bleach (NaClO) solution for 20 minutes respectively. Next, Lotus seeds were washed with sterile water and imbibed in sterile water at 4°C overnight. Lotus seeds were then placed on sterile filter paper sheets that were soaked with sterile water into sealed-off petri dishes and placed vertically at 20°C with a 16/8 hour light/dark cycle six days prior to inoculation.

Mycorrhizal germination

Mycorrhizal spores (Symbiom, MycoGrow™ batch from 2020, Czech Republic) were incubated at a concentration of 500 spores per mL at 28°C in low nutrient SLA media three days before inoculation in a dark environment (Data S12). The SLA was supplemented with strigolactone-like hormone GR24 (Chiralix, UK) to a concentration of 100 nM to induce hyphal germination. On the day of the inoculation, the fungal spores were washed three times with long nutrient SLA before addition to the experimental system.

Experimental set-up

UV-light sterilized black plastic vase-shaped square pots with a height of 15 cm and 7 x 7 cm width and length at the top were filled with filter paper and three different particle layers respectively: a 3-4 cm leca (lightweight expanded clay aggregate) grain layer (particle size between 0.5 and 1 cm), a 10 cm layer containing a mix of leca grains (leca particle size < 0.5 cm), vermiculite, gray sand and sterile water mix layer in a 3:3:1:1 ratio and a 1-2 cm vermiculite layer. These substrate layers were autoclaved for one hour at 120°C before filling the pots.

Two days prior to the inoculation, pots were water-saturated by the addition of 25 mL sterile water, and five Arabidopsis, one Barley or ten Lotus seedlings were transplanted into the pots with their roots facing downward. Each treatment consisted of a different combination of inoculum, host, and nutrient condition and included four replicate pots (Figure 2a, Data S11). The following control treatments were used: 1. hosts grown without inoculum and mycorrhizae, 2. substrate-soil control with inoculum and mycorrhizae, 3. substrate-soil control without inoculum but with mycorrhizae and 4. substrate-soil control without inoculum and mycorrhizae.

SynCom inoculation

The inocula were added to the pots by pouring 25 mL of inoculum (OD₆₀₀ equal to 0.02) diligently from the top into the pots. Additionally, 1 mL of SLA-washed mycorrhizal spores were added from the top into the pots. Controls without inoculum received 25 mL of sterile nutrient solution.

Growth and harvest

Plants were grown in open pots as indicated. The pots were placed individually in boxes to maintain individual watering. These were, in turn, placed into larger boxes to separate host-inoculum combinations. The large boxes were covered with plastic to maintain humidity in the pots during plant

establishment. These plastic covers were removed for Lotus and Barley at one week post inoculation and were kept on the boxes for Arabidopsis until the harvest.

The experiment was conducted twice, with small alterations in the growth conditions between the two experiments. In the first experiment (R1), the plants were only watered with low nutrient SLA media and grown under three different conditions when inoculated with the SSC or grown under one host-specific condition when inoculated with the host-specific inocula:

1. Arabidopsis when inoculated with any inoculum as well as Barley and Lotus when inoculated with the SSC were cultivated in short-day conditions at elevated temperatures (10 h/14 h light/dark, 75%/75% humidity, 22/18 °C, 200/0 lux).
2. Barley when inoculated with any inoculum as well as Arabidopsis and Lotus when inoculated with the SSC were cultivated in long-day conditions at lower temperatures (16 h/8 h light/dark, 80%/80% humidity, 18/16 °C, 200/0 lux).
3. Lotus when inoculated with any inoculum as well as Arabidopsis and Barley when inoculated with the SSC were cultivated in long-day conditions at elevated temperatures (16 h/8 h light/dark, 80%/80% humidity, 22/18 °C, 200/0 lux).

In the second experiment (R2), the three plant species were grown in only host-specific growth conditions irrespective of inoculum.

1. Arabidopsis plants (day/night) were grown in short-day conditions at elevated temperatures (10 h/14 h light/dark, 75%/75% humidity, 22/18 °C, 200/0 lux).
2. Lotus and Barley plants (day/night) were grown in long-day conditions at elevated temperatures (16 h/8 h light/dark, 75%/75% humidity, 22/18 °C, 200/0 lux).

Plants were regularly bottom-watered with low nutrient SLA (R1 and R2), high nutrient SLA (R2) or Hoagland (R2 – only Arabidopsis), depending on the nutrient condition that these plants were inoculated with. Three weeks post-inoculation, plants were removed from the pots, shoots were cut off from the roots and the roots were isolated from the middle clay, vermiculite, and gray sand layer using sterile tweezers. The shoots were weighted (Data S13) and representative pictures of each host and treatment combination (inoculum and nutrient condition) were taken (before removal from the pots) (Figure S2). Excess soil was removed from the roots, and roots were washed five times in sterile water by vigorous shaking. The wash-off of the first wash was spun down for 15 minutes at 4,700 rpm, excess water was removed, and the pellet's resuspension was taken as the rhizosphere fraction. Finally, the firmly washed roots were dried on filter paper and, together with the rhizosphere fraction, stored at -80 °C. For Lotus, nodules were separated from the roots, counted, and separately stored at -80°C (Data S13). The bulk soil samples consisted of 300 mg of substrate taken from the middle clay grain, vermiculite, gray sand layer and stored at -80°C.

DNA library preparation and community profiling

DNA was extracted from input, nodule, rhizosphere, root (endosphere and rhizoplane), and soil samples using the FastDNA SPIN Kit for Soil according to manufacturer's instructions. Metagenomic DNA was diluted to 5 ng/μL and amplified using the Hackflex protocol (Gaio et al., 2022). Amplified metagenomic

fragments were purified using the AMPure XP beads, pooled and sequenced on an Illumina NovaSeq 6000 platform (PE150, 2x150bp).

An overview of the sequencing data is presented in Data S11. Metagenomic samples were cleaned from host reads, depending on the host used in each case, using Bowtie2 (2.2.5) (Langmead & Salzberg, 2012). Once the samples were cleaned, the metagenomic reads were pseudoaligned against the relevant reference index composed of the genome sequences of the inoculated strains using Salmon (1.9.0), except for analysis of within- and between-sample diversity in which case we used the same reference index for all samples based on the genome sequences of all SSC isolates (Patro et al., 2017). Default settings were used for Salmon except for a minimum score for pseudoaligning reads (--minScoreFraction) set at 0.95, allowing for the distinction of bacterial isolates with high genomic similarity. The Salmon output was assembled into a microbiome count table and normalized for the bacterial isolates' genome lengths (Data S15). An overview of the proportion of pseudoaligned reads to the plant genome and SynCom indices is shown in Data S11. The PICRUST2 (2.5.1) algorithm was used to extrapolate the bacterial abundances to KO abundances to investigate the functional or KO composition of the samples (Douglas et al., 2020) (Data S16). The KO profiles from Data S7 were used for this extrapolation step. For further analysis, the isolate and KO abundances were normalized to produce relative abundances. The metadata can be found in Data S14.

Data analysis – phylogenetic placement

Phylogenetic distances between bacterial isolates were calculated with Orthofinder using STAG support values and STRIDE was used to root the phylogenetic tree of the isolates (Figure 1a, Data S2) (Emms & Kelly, 2019).

Data analysis – AtSC and HvSC natural community data

To investigate what proportion of the natural bacterial root community we cover with the culture collections, we matched the metagenomic data of natural communities with the bacterial genomic data of the culture collections. For the AtSC, we classified the shotgun metagenomics reads from Stringlis *et al.* (2018) (PRJNA435676) using Kaiju (1.8.2) (Menzel et al., 2016) on the genus level (Data S3). The AtSC originates from the same soil batch that the experiments in this study were conducted in. The bacterial genera count data was normalized and lowly abundant genera (<0.05%) were removed. Finally, the relative abundances of the bacterial genera present in the AtSC were summed to calculate the AtSC coverage (Figure 1b, Data S3).

Data from natural barley root communities grown in the same soil as the HvSC was used to assess the HvSC coverage. Barley root microbiome samples were subjected to *16S rRNA* V5-V7 sequencing using the 799F and 1192R primers:

Forward: 5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAACMGGATTAGATACCKG-3'

Reverse: 5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACGTCATCCCCACCTTCC-3'

ASV sequences generated by Qiime2-DADA2 (2022.8) (Bolyen *et al.*, 2019) were matched on the 97% sequence identity level with the HvSC *16S rRNA* V5-V7 sequences using the Qiime2-VSEARCH plugin (2022.8). These *16S rRNA* V5-V7 sequences were retrieved from a direct alignment of an arbitrary *16S*

rRNA sequence to the bacterial genomes followed by an *in silico* extraction of the V5-V7 amplicon (Forward - AACMGGATTAGATACCCCKG and Reverse – ACGTCATCCCCACCTTCC) in Qiime2 (2022.8). HvSC coverage was assessed by summing up the relative abundance of the HvSC-matched ASVs (Figure 1c, Data S3).

Data analysis – 1,000 SynCom simulation

To illustrate the uniqueness of the SynCom size implemented in this study with regard to functional diversity, we simulated 1,000 SynCom compositions. These SynCom compositions were based on a one-isolate-per-family principle, randomly picking one isolate from one of the 25 bacterial families that are present in the three combined culture collections. Subsequently, the number of unique KOs as well as the KO intravariability for each of these 1,000 SynComs was calculated to compare it to the SSC (Figures 1e, S1b, Data S6, S9).

Data analysis – KO intravariability

KO diversity, as presented in Figure S1, was assessed by establishing a kmer-based phylogenetic tree of each KO in each inoculum using the gene sequences as nodes with the ape package in R (5.7-1) (Paradis & Schliep, 2019). The sum of all edge lengths is used to obtain the total branch length or total gene diversity (Figure S1; x-axis) (Data S8). The number of nodes in the phylogenetic tree represents the total number of genes (Figure S1; y-axis).

Published soil and rhizosphere KO overlap with KOs identified in this study

To support the large KO overlap found across the Arabidopsis, Barley, and Lotus culture collection, we compared the KO overlap found in natural soil metagenomes with the KO overlap found in natural rhizosphere metagenomes. We used precompiled data from Ofek-Lalzar et al. (2014), containing a KO overview of Maon soils and cucumber and wheat rhizospheres grown in this soil, and from (Stringlis et al., 2018) and an in-house dataset (BioProject ID: PRJNA1108916), containing a KO overview of Reijerscamp soil and Arabidopsis rhizospheres grown in this soil. Using binomial test, we evaluated whether the rhizosphere KO overlap deviates from the soil KO overlap (Data S10, Table S2).

Data analysis – phenotypic data

Phenotypic data such as shoot weight and the number of nodules were statistically assessed by a Kruskal Wallis test followed by a Dunn post hoc test (Figures S3, S4, Data S13).

Data analysis – diversity analyses

Alpha diversity analyses (Figures 2b-e, S7, S9) were performed using for each sample the isolates with relative abundance >0.05% or KOs with relative abundance >1.5e-5% (0.05% divided by the average number of KOs per genome: 3,295). To calculate the number of observed isolates, the number of observed KOs, and the Shannon diversity of the isolate and KO compositions we used the phyloseq package in R (1.44.0) (McMurdie & Holmes, 2013). Statistical differences between the inocula and/or hosts were assessed by an ANOVA with post-hoc Tukey HSD test (Figures 2b-e, S7, S9). Beta diversity analyses to calculate the Bray-Curtis distances between samples (Figures 2f-g, S10) were conducted on the original microbiome and functional abundance tables using the vegan package in R (2.6-4, Oksanen

et al., 2013). Statistical significance of compositional differences by different variables was inferred by an Adonis test/PERMANOVA from the vegan package in R. The R^2 of the Adonis test was calculated to investigate what proportion of the total compositional variance is explained by different variables such as inoculum or host, calculated on multiple taxonomic levels and the functional level in Figures 2h, 3a-c. In Figure 3b-c, the nitrogen-fixing symbionts of Lotus or *Rhizobacter* sp. P2_G4, representing the dominant isolates in the dataset, were computationally excluded to quantify their effect on the microbiome composition. To investigate variation within samples of the same treatment (unique combination of host, inoculum, and nutrient condition), we determined the centroid of the treatment in the PCoA plane and calculated the distance from each sample to the centroid (Figure S10).

Data analysis – simulation of within and between community diversity

To investigate how Adonis/PERMANOVA R^2 values of the microbiome composition correlate to the R^2 values of the functional composition, we conducted a simulation with 1,000 iterations (Extended Data 2). In each iteration, we generated a randomized microbiome dataset consisting of 54 samples (equal to those present in our dataset): nine groups of six replicates. We determined that the average number of isolates that are recruited across the root samples was 119 and thus selected this number of random isolates for every host-specific inoculum thrice, thus totaling nine groups. The random abundance values resembled the abundance distribution in the actual SSC dataset (Table S3). The microbiome dataset was then extrapolated to a functional dataset using the PICRUST2 (2.5.1) algorithm (Douglas et al., 2020) and the Bray-Curtis distances between the samples on both the genus and KO composition were calculated as before. We categorized the dataset into three, six, or nine groups by k-means clustering before calculating the Adonis R^2 values that separate these three, six, or nine groups in both the isolate and functional dataset. The isolate and functional R^2 values were subsequently plotted to each other, and a Pearson correlation was used to calculate the significance and slope of the correlation and deviations of this correlation were statistically assessed through standardized residual testing (Table S4).

Data analysis – effect of plant species on host R^2

To investigate the effect of each host plant species on the microbiome and functional composition, we subsetted the data per inoculum and excluded one host at a time to recalculate the R^2 value and compare it with that obtained R^2 value for the entire data set. An overview on how the R^2 value changes at each taxonomic level and functional level in the different datasets (with and without dominators or without nodulators or P2_G4) can be found in Table S5.

The functional PERMANOVA R^2 values in Table S5 were used to calculate the contribution of each plant species to the host R^2 presented in Figure 3a-b. For each plant species, the difference between the host R^2 value with all three hosts and the host R^2 value of the dataset with the other two hosts is calculated. The resulting value divided by the sum of all these values leads to the host plant species contribution (Formula C) (outer ring in Figure 3a-b). To investigate the degree to which the plant species affect the host R^2 in each inoculum, the contributions of the three hosts in each inoculum are summed and divided by the summed contributions of the three hosts across all inocula (Formula D) (inner ring in Figure 3a-b).

$$C1) \text{ Arabidopsis contribution} = 100\% * \frac{\text{Host } R^2 - \text{Host } R^2_{\text{Barley} \propto \text{Lotus}}}{\sum \text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Barley}} + \text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Lotus}} + \text{Host } R^2 - \text{Host } R^2_{\text{Barley} \propto \text{Lotus}}}$$

$$C2) \text{ Barley contribution} = 100\% * \frac{\text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Lotus}}}{\sum \text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Barley}} + \text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Lotus}} + \text{Host } R^2 - \text{Host } R^2_{\text{Barley} \propto \text{Lotus}}}$$

$$C3) \text{ Lotus contribution} = 100\% * \frac{\text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Barley}}}{\sum \text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Barley}} + \text{Host } R^2 - \text{Host } R^2_{\text{Arabidopsis} \propto \text{Lotus}} + \text{Host } R^2 - \text{Host } R^2_{\text{Barley} \propto \text{Lotus}}}$$

$$D1) \text{ Host effect in AtSC} = 100\% * \frac{\sum \text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in AtSC}}{\sum \text{AtSC}(\text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in AtSC}) + \sum \text{HvSC} + \sum \text{LjSC} + \sum \text{SSC}}$$

$$D2) \text{ Host effect in HvSC} = 100\% * \frac{\sum \text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in HvSC}}{\sum \text{AtSC}(\text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in AtSC}) + \sum \text{HvSC} + \sum \text{LjSC} + \sum \text{SSC}}$$

$$D3) \text{ Host effect in LjSC} = 100\% * \frac{\sum \text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in LjSC}}{\sum \text{AtSC}(\text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in AtSC}) + \sum \text{HvSC} + \sum \text{LjSC} + \sum \text{SSC}}$$

$$D4) \text{ Host effect in SSC} = 100\% * \frac{\sum \text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in SSC}}{\sum \text{AtSC}(\text{Arabidopsis contribution} + \text{Barley contribution} + \text{Lotus contribution in AtSC}) + \sum \text{HvSC} + \sum \text{LjSC} + \sum \text{SSC}}$$

Data analysis – intrafamily functional diversity and root colonization

To assess whether higher intrafamilial functional diversity was associated with increased root colonization, we calculated, for each bacterial strain, the proportion of intragenomic non redundant KOs relative to the total set of KOs identified within its corresponding family (Figures 3d, S11b, S12) (Formula A). Similarly, the relative abundance of each bacterial strain was expressed as a proportion of the cumulative relative abundance of its family (Formula B). Both proportions were calculated separately for each bacterial family within the host-specific inocula from which they originated (AtSC, HvSC and LjSC). Finally, the KO proportion and the proportional relative abundance (log-linear scale) were used to evaluate the presence of a correlation between intrafamilial functional diversity and root colonization.

$$A) \text{ KO proportion} = \frac{\text{Number of unique KOs in isolate}}{\text{Number of unique KOs in family pangenome}}$$

$$B) \text{ Relative abundance proportion} = \frac{\text{Relative abundance isolate in dataset}}{\text{Relative abundance family in dataset}}$$

To evaluate whether the observed correlations deviated from random expectations, a statistical randomization test was performed. KO proportions and relative abundance values were randomly shuffled within each bacterial family and SynCom 1,000 times. For each permutation, the corresponding *p*-value was recalculated, generating a null distribution of *p*-values. The 97.5% confidence interval of this distribution was then determined and compared with the observed (non-randomized) *p*-value (Table S6).

Arabidopsis multi-SynCom root colonization experiment - SynCom design and preparation

To further investigate and validate the intrafamilial relationship between functional diversity and root colonization, an additional complex SynCom inoculation experiment was performed using SynComs designed based on high functional diversity. Functionally complex, 200-member AtSC-derived SynComs were designed through a multi-step computational pipeline. Firstly, 1,000 random assortments of 200 bacteria were generated using *NumPy random.choice* (Harris et al., 2020) with size=200 and replace=False while picking from the set of 447 AtSC members. Secondly, the community functional complexity was quantified by calculation of KO diversity (Shannon index) using *SciPy entropy* (Virtanen et al., 2020) across the sum of the genome sequences corresponding to the 200 selected bacteria per assortment. Thirdly, the assortments were ranked by Shannon index (high-to-low), and finally a random selection of seven diverse SynComs from the top-50 was made. Notably, maximization of functional diversity simultaneously led to taxonomically balanced communities with members from all major families. Overview of the 1,000 random assortments, corresponding functional diversity (Shannon index) and final selection of SynComs for the community assembly validation experiment are in Data S18.

SynComs were prepared largely as for the SSC experiments. In short, bacterial suspensions were transferred from -80 °C glycerol stocks and inoculated into 1/10th TSB in deep-well plates, five days prior to plant inoculation. On the day of plant inoculation, bacteria were harvested by repeated centrifugation at 4,700 rpm and resuspension in Hoagland plant-growth solution (trice). Finally, optical

density at 600 (OD₆₀₀) was used as a proxy for bacterial growth and used to determine the appropriate volume for equimolar pooling at a final OD₆₀₀ equal to 0.02.

Arabidopsis multi-SynCom root colonization experiment - plant preparation, inoculation and sampling

Arabidopsis thaliana ecotype Columbia-0 (Col-0) plants were grown using a gnotobiotic FlowPot system adapted from previously described methods (Kremer et al., 2021), as this system is better suited for growing *Arabidopsis* as compared to the leca grain-riversand mix experimental set-up. FlowPots were assembled from 50-mL plastic Falcon tubes cut at the 30-mL mark and filled with double autoclaved and oven-baked river sand. Assembled FlowPots were fixed onto a support at the bottom of a box container and autoclaved twice. After cooling, the substrate was saturated under sterile conditions by flushing bacterial inocula and placement of *Arabidopsis* seeds.

Prior to sowing, *Arabidopsis* seeds were surface-sterilized in a desiccator with a mix of 100 ml bleach and 3.2 ml HCl (37%) for 3-4 hours followed by stratification at 4 °C for ~48 hours. Three to four seeds were sown per FlowPot under sterile conditions. FlowPots were irrigated through the bottom port with either sterile Hoagland solution (control) or defined microbial synthetic community (SynCom) suspensions prepared from the AtCC collection. In total, we tested seven 200-member SynComs (9-12 biological replicates each) (Data S18).

Plants were grown in a controlled-environment growth chamber under short-day conditions (10 h light/14 h dark) at 21 °C, 200 $\mu\text{mol m}^{-2} \text{s}^{-1}$ and harvested 6-weeks post-inoculation. Under sterile conditions each FlowPot was disassembled, shoots removed with a sterile scalpel, and roots quickly rinsed in sterile water and gently shaken to remove residual substrate. Root tissue from each FlowPot, thus composed of multiple plants, was then used for DNA extraction using the MagMAX™ Plant DNA isolation kit (Thermo Fisher Scientific), and followed by HackFlex-based Illumina sequencing at Useq as before using the Illumina NovaSeq X platform (PE150, 2x150bp) (Gaio et al., 2022).

Arabidopsis multi-SynCom root colonization experiment – data analysis

The data was preprocessed similarly to the SSC data, including host-read removal with Bowtie2 (2.2.5) (Langmead & Salzberg, 2012), pseudoalignment of metagenomic reads to bacterial genome references (based on the 200-member SynCom composition for each SynCom separately) with Salmon (1.9.0) set with a 0.95 minimum threshold score (Patro et al., 2017) (Data S19), and normalization of isolate counts using the bacterial genome lengths (Data S20). Subsequently, for each SynCom and for the six largest bacterial families (*Beijerinckiaceae*, *Burkholderiaceae*, *Caulobacteraceae*, *Pseudomonadaceae*, *Rhizobiaceae*, and *Xanthomonadaceae*), the log-linear correlation between KO proportion (Formula A) and relative abundance proportion (Formula B) was assessed.

To assess whether these correlations are consistent across SynComs of varying composition, bacterial strains of each combination of family and SynCom were grouped into clusters based on the intragenomic number of unique KOs using mclust (version 6.1.1) (Scrucca et al., 2023) and ranked from highest to lowest. These groups were subsequently ranked in a similar fashion based on their cumulative relative abundance from highest to lowest, and a Kendall's rank test was conducted for each

bacterial family across SynComs to assess whether there is a significant rank correlation (Table S7, Data S21).

Data analysis – family effect on SynCom and host R^2

To assess the influence of bacterial families on microbiome composition—analogous to evaluating the host's effect—we performed an *in silico* bacterial family-specific exclusion analysis (Figure 4a, Data S22). For each bacterial family, we computationally removed the corresponding isolate abundances from the microbiome data. For the remaining isolates we then extrapolated the isolate abundances to KO abundances using the PICRUSt2 (2.5.1) algorithm. Subsequently, we calculated the change in inoculum and host PERMANOVA R^2 compared to the original dataset to assess the family effect.

Data analysis – sPLS-DA

To investigate which KOs contribute to compositional differences between inocula or hosts, we implemented sparse Partial Least Squares Discriminant analysis (sPLS-DA) (Data S23). sPLS-DA is a type of supervised machine learning designed for a classification task that emphasizes feature contribution to two or multiple classes. The R package mixOmics (6.24.0) (Rohart et al., 2017) first optimizes the hyperparameters through a grid search, leading to the most optimal number of principal components, data subsets and sizes, and iterations. The tuned hyperparameters were then implemented in sPLS-DA to generate mean contribution scores for the KOs to the compositional differences between inocula or hosts. These contribution scores were normalized for the variation explained by the principal component they contribute to (Formula E) (Figure 4b).

$$E) \text{ Weighted contribution} = \text{sPLS – DA contribution of KO} * \text{PCoA explained variation}$$

The cumulative abundances of bacterial family-derived KOs with contribution scores to specific hosts in specific inoculum contexts were subsequently investigated in two subsets of the data. These subsets included a dataset in which all bacterial strains were removed except for the family in question, and an opposite dataset in which all bacterial strains were kept except for the family in question (Figure 4c). Statistical differences were assessed using an ANOVA followed by a post-hoc Tukey HSD test.

Data analysis – differential abundance analysis

Differential abundance testing was performed using DESeq2 (version 1.40.0) (Love et al., 2014) (Data S24). DESeq2 has an internal normalization step that normalizes KO abundances by the geometric mean of the KO, followed by a local regression and a log2foldchange shrinkage using the apegglm shrinkage estimator (Zhu et al., 2019). Significance was tested by a Wald test and p -values were adjusted for multiple testing using the Benjamini and Hochberg method (significance $p < 0.05$). DESeq2 was run to find enriched KOs in the root microbiome as compared to the initial inoculum and compare this between hosts and inocula (Figures 5, Extended Data 4, S14, S19, Table S8).

Pathway and KO annotations were derived from the KEGG database (Data S17). KOs without pathway annotation (3,443 KOs) were given an annotation based on their KO annotation, e.g. transporter, transcriptional regulator, or secretion, to generate a new KO pathway annotation table (Data S17).

Data analysis – differential abundance analysis – fold change calculation

For KOs and pathways, the fold change of isolates with the KO or pathway was calculated between the root samples and the input samples to assess whether these are selected by the host (Figures 5b-c, e-f, Extended Data 5, 6, S15, S16, S17, S20). For each host-inoculum combination, the cumulative relative abundances of isolates with the KO or pathway were averaged across sample replicates and divided by the average cumulative relative abundances of isolates with the KO or pathway in the respective input inocula (Formula F1) (Data S25, S26, S27, S31). For pathways, only significant KOs were taken into consideration for the calculation of the fold change. In this analysis, inocula for which the KO/pathway was not significant for any host were excluded. The KO/pathway average was then calculated for the remaining inocula in which it is significant.

$$F1) FC \text{ present} = \frac{\frac{\sum_{\text{root sample replicates}} (\sum_{\text{isolates with KO | pa}} \text{relative abundance in root})}{\text{number of root sample replicates}}}{\frac{\sum_{\text{input sample replicates}} (\sum_{\text{isolates with KO | pathway}} \text{relative abundance in initial inoculum})}{\text{number of initial inoculum sample replicates}}}$$

The median fold change across inocula was taken to produce Figure 5b-c, e (Data S25, S26, S27). The respective fold change for KOs and pathways across all host-inoculum combinations were averaged to produce Figures 5f, S20.

In Figures 5f, S20, the fold change differences on the y-axes are calculated between the isolates with the KO or pathway as compared to the isolates without the KO (Formula F2 and F3).

$$F2) FC \text{ Absent} = \frac{\frac{\sum_{\text{root sample replicates}} (\sum_{\text{isolates without KO | pathway}} \text{relative abundance in root})}{\text{number of root sample replicates}}}{\frac{\sum_{\text{input sample replicates}} (\sum_{\text{isolates without KO | pa}} \text{relative abundance in initial inoculum})}{\text{number of initial inoculum sample replicates}}}$$

$$F3) \text{ Enrichment} = \log_2 \left(\frac{FC \text{ present}}{FC \text{ absent}} \right)$$

For the pathway figures (Figures 5f, S20), a Wilcoxon test was conducted to compare the fold changes of the isolates with and without the pathway and only significant differences (Bonferroni adjusted p -value < 0.05) were kept (Data S31).

Data analysis – differential abundance analysis – host-specific KOs/pathways

To select host-specific KOs, the DESeq2 output was filtered to only include KOs that were significant in the presence of at least three inocula for one host in the original dataset. Similar to the general KOs, the fold change between the relative abundance of isolates with the KO in the root samples and the input samples was calculated for each host individually and the median value from the four inocula was taken (Formula F1) (Data S26). From this dataset, KOs were considered host-specific when the fold change for one host was higher than three (>3) and lower than three (<3) for the other two hosts (Data S26). These host-specific sets were used to produce Figure 5c-d. In consideration that many of these KOs could be derived from genomes of the nitrogen-fixing symbionts, we repeated the same analysis for Lotus with the computational exclusion of LjNodule214, P1_H10, P2_A12, and P2_D6, from the dataset (Data S27).

From the host-specific lists of KOs, we selected Arabidopsis, Barley, and Lotus-specific gene clusters for in-depth exploration (SR2, Extended Data 5, 6, S15, S16, S17). The gene abundances, isolate abundances, isolate fold changes between root and input, and the presence of these genes across the bacterial phylogeny of these gene clusters is shown in these figures, with an explanation of the pathway in which they function.

The most frequent host-specific pathway—ABC transporters—was examined in detail, focusing on their distribution and diversity among the SSC isolates (SR3, Extended Data 8), as well as the classes of compounds they are predicted to transport (SR3, Figure S18, Data S28, S29). For Figure S18, the p -values were averaged across gene cassettes that together make up the ABC transporter (Data S29) using Stouffer's method (Stouffer et al., 1949).

Data analysis – effect of soil origin on KO selection

Because the three culture collections originated from different host plants and soil types, we performed a direct comparison between the LjSC collection and the Arabidopsis root-derived bacterial isolates obtained from the same Cologne agricultural soil (Bai et al., 2015) (Data S30). Using a Kolmogorov–Smirnov test, the bimodal distributions of KO profiles in these two culture collections were compared to statistically assess whether they differed significantly.

Data analysis – differential abundance analysis – general KOs/pathway subsets

To produce Figures 5b-c, e-f, S20, general KOs were selected from the DESeq2 output (Data S25). In Figure 5b, the threshold was lenient, incorporating KOs that were significantly enriched in more than six out of the 12 host-inoculum combinations (852 KOs). For Figures 5e-f, S20, we used a stricter threshold by only incorporating KOs that were significantly enriched in all 12 host-inoculum combinations after removing KOs of the dominant Lotus symbionts and P2_G4 (266 KOs). For Figures 5f, S20, only the KOs with pathway annotations from these 266 KOs were plotted (203 out of the 266 KOs were assigned to pathways) (Data S25).

Aside from the in-depth analysis that was conducted in this manuscript, we generated datasets that can produce Figures 5, S20, and the individual plots from Extended Data 5, 6, S15, S16, S17, meaning the isolate abundances, gene abundances, ternary plots, and the family pie charts, for any KO or pathway (pathways only 5f) (Data S46, S47).

Data analysis – Distribution of 266 KOs across SSC isolates

To identify KOs associated with highly abundant isolates within each bacterial family, we performed a bacteria enrichment analysis between highly abundant (Top) and lowly abundant (Non-Top) isolates for each KO. For each plant-SynCom combination, the SynCom members were ordered by abundance and the bacterial strains that cumulatively account for 90% of the relative abundance were categorized as top (Data S32). The KO feature table containing 266 KOs was used to construct 2x2 contingency tables, categorizing isolates based on two criteria: isolate relative abundance—Top vs. Non-Top—and KO presence. Contingency tables were generated using Python (v3.12.5), and statistical association was tested using Barnard's exact test, as implemented in SciPy (v1.13.1) (Barnard, 1947). Families with fewer than five isolates were excluded from this analysis. Additionally, contingency tables containing

two or more empty cells (equal to zero) were omitted to avoid artifacts and unreliable statistics. For each KO, Barnard's exact test yielded Z-scores and *p*-values, which were subsequently adjusted for multiple comparisons using Benjamini-Hochberg procedure in R (v4.1.2). Finally, results were saved as tables (Data S33), and statistical significance (adjusted *p*-value < 0.05) was visualized as dot plots using ggplot2 (v3.5.1) (Figure 5g).

Lotus bacterial family drop out SynCom root colonization experiment – SynCom preparation

To study the effect of individual bacterial families on the root microbiome's functional composition and corroborate the *in-silico* analysis, another SynCom inoculation experiment was conducted on *Lotus japonicus* Gifu in which several prominent bacterial families were each dropped out in different SynComs (Data S34). Using only bacterial isolates from the LJSC (177), bacterial isolates were cultivated in a similar fashion to the SuperSynCom experiment and individually adjusted to an OD₆₀₀ of 0.1. The bacterial isolates were then pooled into the following family groups: *Burkholderiaceae* (n=69), *Caulobacteraceae* (n=12), *Pseudomonadaceae* (n=21), *Rhizobiaceae* (n=27), and others (n=48), which included all remaining bacteria. The OD₆₀₀ of each family group was then adjusted to 0.02 using low nutrient SLA (Data S12), pooled together to create five SynComs with all families minus the dropout family, and the OD₆₀₀ was readjusted to 0.02.

The experimental set-up and seed surface-sterilization of *Lotus* seeds was similar to the SSC treatment, with a few exceptions: six replicate *Lotus* plants per treatment were used instead of four with one plant per pot, only a low nutrient condition (SLA) was used, and each pot received 10 mL of inoculum instead of 25 mL (OD₆₀₀ 0.02). Similarly, negative controls only received 10 mL of sterile nutrient solution. The harvest of *Lotus* plants, the DNA extraction and WGS Hackflex library preparation was similar to the SSC treatment, with the following exception: the amplified metagenome fragments were sequenced on an Illumina NovaSeq X Plus platform (PE150, 2x150 bp).

Lotus bacterial family drop out SynCom root colonization experiment – Data analysis

The data was preprocessed similarly to the SSC data and Arabidopsis multi-SynCom root colonization experiment data, including host-read removal with Bowtie2 (2.2.5) (Langmead & Salzberg, 2012), pseudoalignment of metagenomic reads to bacterial genome references (based on the bacterial strains minus the dropped out bacterial family for each SynCom separately) with Salmon (1.9.0) set with a 0.95 minimum threshold score (Patro et al., 2017) (Data S35), normalization of isolate counts using the bacterial genome lengths (Data S36), and extrapolation of bacterial abundances to KO abundances using PICRUSt2 (2.5.1) (Douglas et al., 2020) (Data S37). Sequencing revealed that a member of *Rhizobiaceae* was present in samples with *Rhizobiaceae* drop-out, hence data from this SynCom was not considered for further analyses. As only one input inoculum sample was sequenced for each SynCom, due to low sequencing depth, we created four synthetic input samples for each SynCom by rarefaction. For each SynCom input sample, a rarefaction curve was created using phyloseq (with step 250) (McMurdie & Holmes, 2013) and the depth was extracted at which 95% of all isolates in the sample are present. These depths were then averaged across all input inoculum samples and used as the rarefaction depth to create four synthetic input inoculum samples.

To investigate the effect of individual bacterial families on the root microbiome's functional composition, a niche replacement score was calculated (Formula G). Similarly to the SSC, differential abundance analyses were performed using DESeq2 between the root microbiome and the input inoculum (Love et al., 2014) (Data S38). Subsequently, the differential abundance analyses output was used to calculate the isolate fold changes (Formula F1). The comparison of the fold change between the SynCom that misses a bacterial family and the full LjSC condition is the niche replacement score (Formula G). These scores were calculated for the plant-specific KO set that includes the nodulator data (Data S39) and core root-associated KO set in which the nodulators were computationally removed before the extrapolation step (Data S40). These scores were aggregated to pathway scores and shown in Figure S21.

$$G) \text{ Niche replacement score} = \log_{10}\left(\frac{\text{Fold change}_{\text{Dropped-out family}}}{\text{Fold change}_{\text{Full LjSC}}}\right)$$

To investigate whether specific bacterial families are responsible for the selection of the 266 KOs from the root microbiome, the number of enriched KOs was compared between the full LjSC condition and the SynComs that miss one family. The difference between these numbers points to KOs that are no longer selected for, and thus suggests that these KOs may be contributed by the dropped-out bacterial family. Afterwards, a binomial test was conducted to statistically assess whether this difference represents an over or underrepresentation of the 266 KOs compared to the total KOs (Table S9).

Data analysis – comparison to Levy et al. (2018)

To investigate the universality of the KOs identified in this study for root competence, we investigated the distribution of the 266 KOs across the SSC isolates (Extended Data 9a). In addition, we also investigated whether the 266 KOs are plant-associated in the data of Levy et al. (2018). The KO profiles of the 3,837 bacteria investigated by Levy et al. (2018) were downloaded (from http://labs.bio.unc.edu/Dangl/Resources/gfobap_website/index.html), and the presence of the 266 KOs in the KO profiles of all groups of bacteria (non-plant-associated, plant-associated, and soil bacteria) was recorded (Figure Extended Data 9b, Data S41, S42, S43). In addition, the Venn diagram in Extended Data 9b shows the overlap for the 266 KOs with the plant-associated KOs found by Levy et al. (2018) (Data S42, S43). A binomial test was performed to determine whether the overlap between the plant-associated KOs reported by Levy et al. (2018) and the 266 KOs identified in this study was greater than expected by chance. Specifically, enrichment was evaluated by comparing the observed overlap with the expected overlap based on the total set of KOs reported by Levy et al. (2018) and the KO set used in the present analysis.

Data analysis – KOs in natural soils and rhizospheres

To assess whether the lenient (852 KOs) and stringent (266 KOs) KO sets identified in this study were also enriched in natural rhizospheres compared with bulk soils, we analyzed data from López et al., (2023) and an in-house dataset (BioProject ID: PRJNA1108916). For the López et al. (2023) dataset, KOs associated with the rhizosphere niche across bacterial phyla were extracted (Data S44). For the in-house dataset, differential abundance analysis was performed using ANCOM-BC2 (version 2.6.0) (Lin & Peddada, 2024) to compare metagenomes from the rhizosphere of *Arabidopsis thaliana* grown in

natural Reijerscamp soil with corresponding bulk soil metagenomes (Data S45). The resulting list of rhizosphere-associated and rhizosphere-enriched KOs from both datasets were subsequently compared with the full set of associated and enriched KOs identified in this study using hypergeometric and Fisher's exact enrichment tests (Table S10).

Data availability

Data deposition: The data reported in this article have been deposited in the National Center for Biotechnology Information Short Read Archive BioProject database. The relevant accession numbers can be found in the separate, specific sections below.

Arabidopsis culture collection

Genomic reads from the bacterial cultures are deposited under accession number PRJNA1131834, while the genomic sequences are deposited under accession numbers PRJNA1138681 (isolated at three days post transplantation) and PRJNA1139421 (isolated at seven days post transplantation). Genomes, PROKKA gene predictions, and EggNOG annotations are also available at <https://zenodo.org/records/10992416> (Selten *et al.*, 2024).

Barley culture collection

Genomic reads from the bacterial cultures are deposited under accession number PRJNA1131819, while the genomic sequences are deposited under accession number PRJNA1139693.

Lotus culture collection

Genomic reads as well as genome assemblies from the bacterial cultures have been deposited previously from Wippel *et al.* (2021) under accession number PRJEB37696. Genomic sequences can also be found at <https://www.at-sphere.com/>. Bacterial isolates LjRoot204, LjRoot204_2, LjRoot205, LjRoot206, and LjRoot208 were not included under this accession number and can be found under accession numbers PRJNA1142758 (genomic reads) and PRJNA1142862 (genomes).

Metagenome sequencing data

Shotgun metagenome-sequenced reads from the root microbiomes of Arabidopsis, Barley, and Lotus inoculated with the AtSC, HvSC, LjSC, and SSC generated in this study have been deposited under accession number PRJNA1131994. The shotgun metagenome-sequenced reads from the Arabidopsis multi-SynCom root colonization experiment are deposited under accession number PRJNA1449013, while the shotgun metagenome-sequenced reads from the Lotus bacterial family drop out SynCom root colonization experiment are deposited under accession number PRJNA1444504.

Shiny app

To assist the community in investigating specific KOs in our data, we have built a shiny app under the following link: https://pm-bacterial-genetics-au.shinyapps.io/SSC_community_app/. This app will allow users to investigate KO abundances across hosts and inocula, abundances of isolates with a specific KO/pathway across hosts and inocula, and the functional enrichment of KOs for each host that is similarly calculated in this manuscript. The user can specify the original dataset as well as the dataset in which the dominator isolates, Lotus symbionts (LjNodule 214, P1_H10, P2_A12, and P2_D6), and the

HvSC-dominant P2_G4, are computationally excluded. The data to generate the figures in the shiny app are included in Data S46 (original dataset) and S47 (exclusion of nodulators and P2_G4 dataset).

Scripts

Rscripts developed for the analyses conducted in this study and to generate the figures can be found under this GitHub link: https://github.com/Gijsselten13/Scripts_SSC_Community. The processed data on which the Rscripts are run can be found under the following Zenodo link: <https://doi.org/10.5281/zenodo.20028852>.

Methods – references

- Bai, Y., Müller, D. B., Srinivas, G., Garrido-Oter, R., Pothhoff, E., Rott, M., Dombrowski, N., Münch, P. C., Spaepen, S., Remus-Emsermann, M., Hüttel, B., McHardy, A. C., Vorholt, J. A., & Schulze-Lefert, P. (2015). Functional overlap of the *Arabidopsis* leaf and root microbiota. *Nature*, **528**, 364–369.
- Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika*, **34**, 123–138.
- Berendsen, R. L., Vismans, G., Yu, K., Song, Y., de Jonge, R., Burgman, W. P., Burmølle, M., Herschend, J., Bakker, P. A. H. M., & Pieterse, C. M. J. (2018). Disease-induced assemblage of a plant-beneficial bacterial consortium. *The ISME Journal*, **12**, 1496–1507.
- Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., ... Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, **37**, 852–857.
- Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., Peplies, J., Gloeckner, F. O., Amann, R., Eickhorst, T., & Schulze-Lefert, P. (2012). Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature*, **488**, 91–95.
- Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution*, **38**, 5825–5829.
- Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., ... Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.
- Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P., & Parks, D. H. (2020). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics*, **36**, 1925–1927.
- Coil, D., Jospin, G., & Darling, A. E. (2015). A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics*, **31**, 587–589.
- de Muinck, E. J., Trosvik, P., Gilfillan, G. D., Hov, J. R., & Sundaram, A. Y. M. (2017). A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform. *Microbiome*, **5**, 68.
- Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., Huttenhower, C., & Langille, M. G. I. (2020). PICRUSt2 for prediction of metagenome functions. *Nature Biotechnology*, **38**, 685–688.
- Emms, D. M., & Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology*, **20**, 1–14.
- Gaio, D., Anantanawat, K., To, J., Liu, M., Monahan, L., & Darling, A. E. (2022). Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. *Microbial Genomics*, **8**, 000744.
- Gonzalez-y-Merchand, J. A., Estrada-Garcia, I., Colston, M. J., & Cox, R. A. (1996). A novel method for the isolation of mycobacterial DNA. *FEMS Microbiology Letters*, **135**, 71–77.
- Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M.,

- Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, **585**, 357–362.
- Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., & Glöckner, F. O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41**, e1–e1.
- Kremer, J. M., Sohrabi, R., Paasch, B. C., Rhodes, D., Thireault, C., Schulze-Lefert, P., Tiedje, J. M., & He, S. Y. (2021). Peat-based gnotobiotic plant growth systems for *Arabidopsis* microbiome research. *Nature Protocols*, **16**, 2450–2470.
- Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**, 357–359.
- Levy, A., Salas Gonzalez, I., Mittelviehhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., Wang, K., Devescovi, G., Stillman, K., Monteiro, F., Rangel Alvarez, B., Lundberg, D. S., Lu, T.-Y., Lebeis, S., Jin, Z., McDonald, M., Klein, A. P., Feltscher, M. E., Rio, T. G., ... Dangl, J. L. (2018). Genomic features of bacterial adaptation to plants. *Nature Genetics*, **50**, 138–150.
- Lin, H., & Peddada, S. Das. (2024). Multigroup analysis of compositions of microbiomes with covariate adjustments and repeated measures. *Nature Methods*, **21**, 83–91.
- López, J. L., Fourie, A., Poppeliers, S. W. M., Pappas, N., Sánchez-Gil, J. J., de Jonge, R., & Dutilh, B. E. (2023). Growth rate is a dominant factor predicting the rhizosphere effect. *The ISME Journal*, **17**, 1396–1405.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**, 1–21.
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, **17**, 10–12.
- McMurdie, P. J., & Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLOS ONE*, **8**, e61217.
- Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, **7**, 11257.
- Ofek-Lalzar, M., Sela, N., Goldman-Voronov, M., Green, S. J., Hadar, Y., & Minz, D. (2014). Niche and host-associated functional signatures of the root surface microbiome. *Nature Communications*, **5**, 4950.
- Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Wagner, H., & others. (2013). Package 'vegan.' *Community Ecology Package*, **2**, 1–295.
- Olm, M. R., Brown, C. T., Brooks, B., & Banfield, J. F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME Journal*, **11**, 2864–2868.
- Overgaard, C. K., Tao, K., Zhang, S., Christensen, B. T., Blahovska, Z., Radutoiu, S., Kelly, S., & Dueholm, M. K. D. (2022). Application of ecosystem-specific reference databases for increased taxonomic resolution in soil microbial profiling. *Frontiers in Microbiology*, **13**, 942396.
- Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.
- Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, **25**, 1043–1055.

- Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, **14**, 417–419.
- Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: an R package for ‘omics feature selection and multiple data integration. *PLOS Computational Biology*, **13**, e1005752–e1005752.
- Scrucca, L., Fraley, C., Murphy, T. B., & Adrian E., R. (2023). *Model-based clustering, classification, and density estimation using mclust in R*. Chapman and Hall/CRC.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.
- Selten, G., Stassen, M. J. J., de Rooij, P., Berendsen, R. L., Stringlis, I., & de Jonge, R. (2024). Draft genome sequences of *Arabidopsis thaliana*-associated micro-organisms from Reijerscamp soil, the Netherlands [Data set]. In *Zenodo*.
- Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness (pp. 227–245).
- Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, **35**, 1026–1028.
- Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M. J. (1949). The American soldier: adjustment during army life vol. 1. *J Consult Psychology*, **13**, 310.
- Stringlis, I. A., Yu, K., Feussner, K., de Jonge, R., Van Bentum, S., Van Verk, M. C., Berendsen, R. L., Bakker, P. A. H. M., Feussner, I., & Pieterse, C. M. J. (2018). MYB72-dependent coumarin exudation shapes root microbiome assembly to promote plant health. *Proceedings of the National Academy of Sciences*, **115**, E5213–E5222.
- Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, **17**, 261–272.
- Wippel, K., Tao, K., Niu, Y., Zgadzaj, R., Kiel, N., Guan, R., Dahms, E., Zhang, P., Jensen, D. B., Logemann, E., Radutoiu, S., Schulze-Lefert, P., & Garrido-Oter, R. (2021). Host preference and invasiveness of commensal bacteria in the *Lotus* and *Arabidopsis* root microbiota. *Nature Microbiology*, **6**, 1150–1162.
- Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A., & Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, **2**, 1–18.
- Zhu, A., Ibrahim, J. G., & Love, M. I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*, **35**, 2084–2092.

Supplemental figures and results

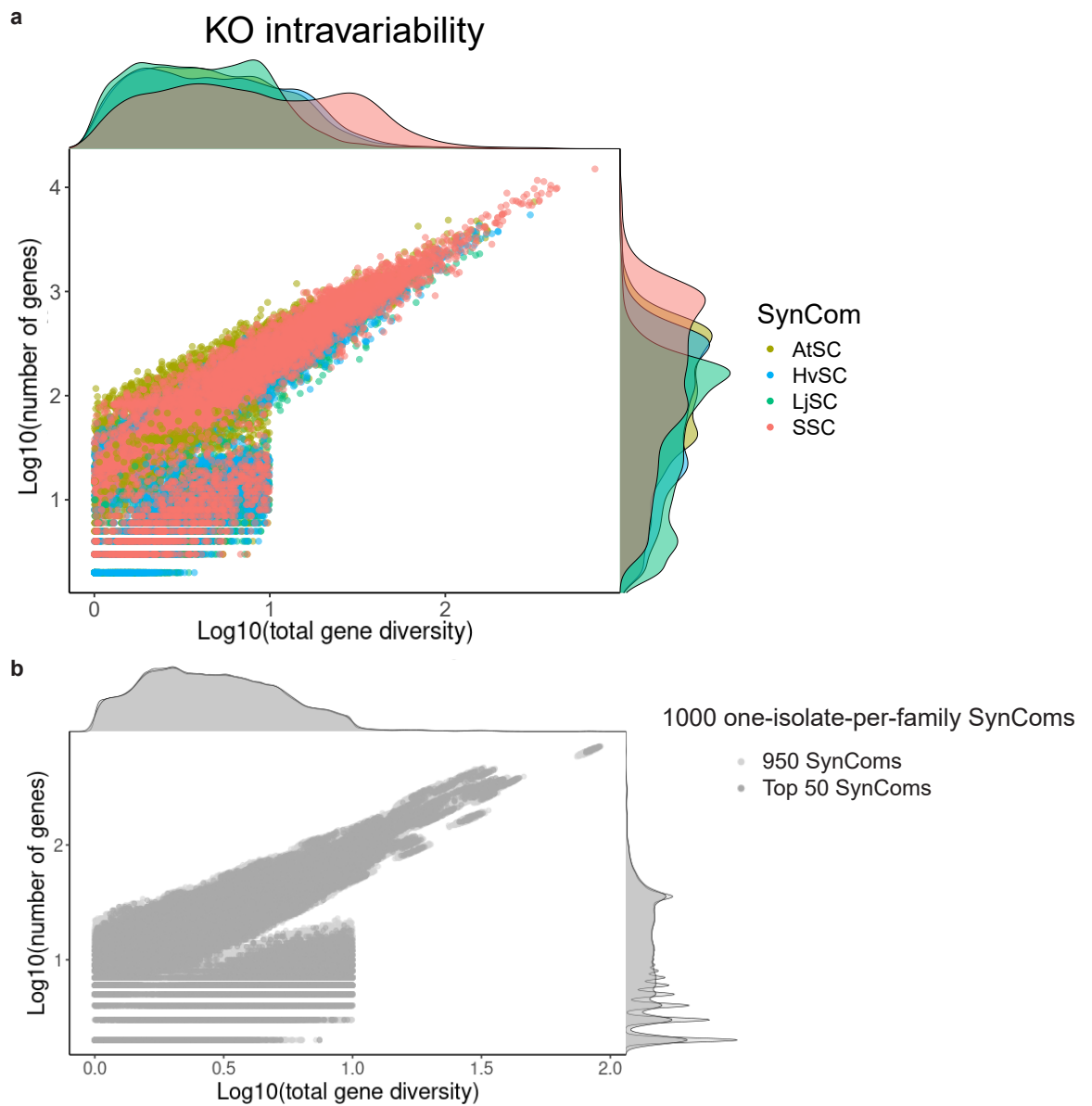


Figure S1. Gene sequence variation within SynComs. Each point represents one KO. The distribution of number of genes per KO in each of the four microbial collections (a) or the one-isolate-per-family SynComs (b) is shown on the y-axis. The x-axis displays the gene diversity, calculated by summing all the branch lengths of the gene dendrogram of each KO found in the SynCom. In b, $n = 1,000$ independently simulated one-isolate-per-family SynComs (Data S6) and the SynComs with the highest cumulative functional diversity are indicated by a darker grey color. A higher gene diversity indicates a larger variation in nucleotide sequence between genes in a specific KO. A higher complexity in the four microbial collections displays a higher gene diversity compared to SynComs of 25 isolates.

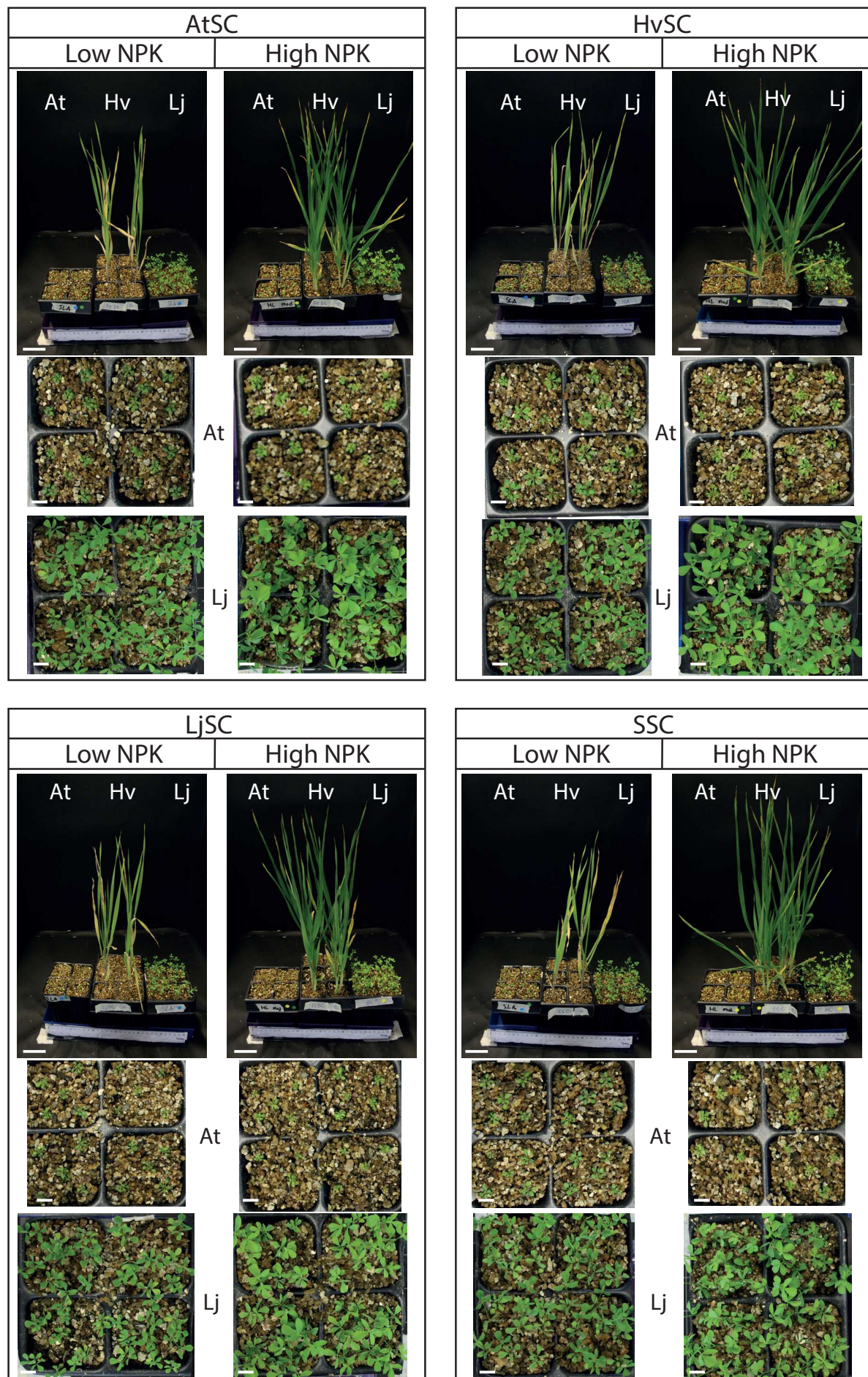


Figure S2. Experimental photos Arabidopsis, Barley and Lotus shoots. Representative phenotypes of Arabidopsis, Barley, and Lotus three weeks after inoculation with microbial communities and grown either in low or high nutrient conditions (NPK). Similar results were obtained in two independent experiments. All the plants were photographed at the same scale. A close-up top view is added for Arabidopsis and Lotus plants. Scale bars: 5 cm (side); 1 cm (top).

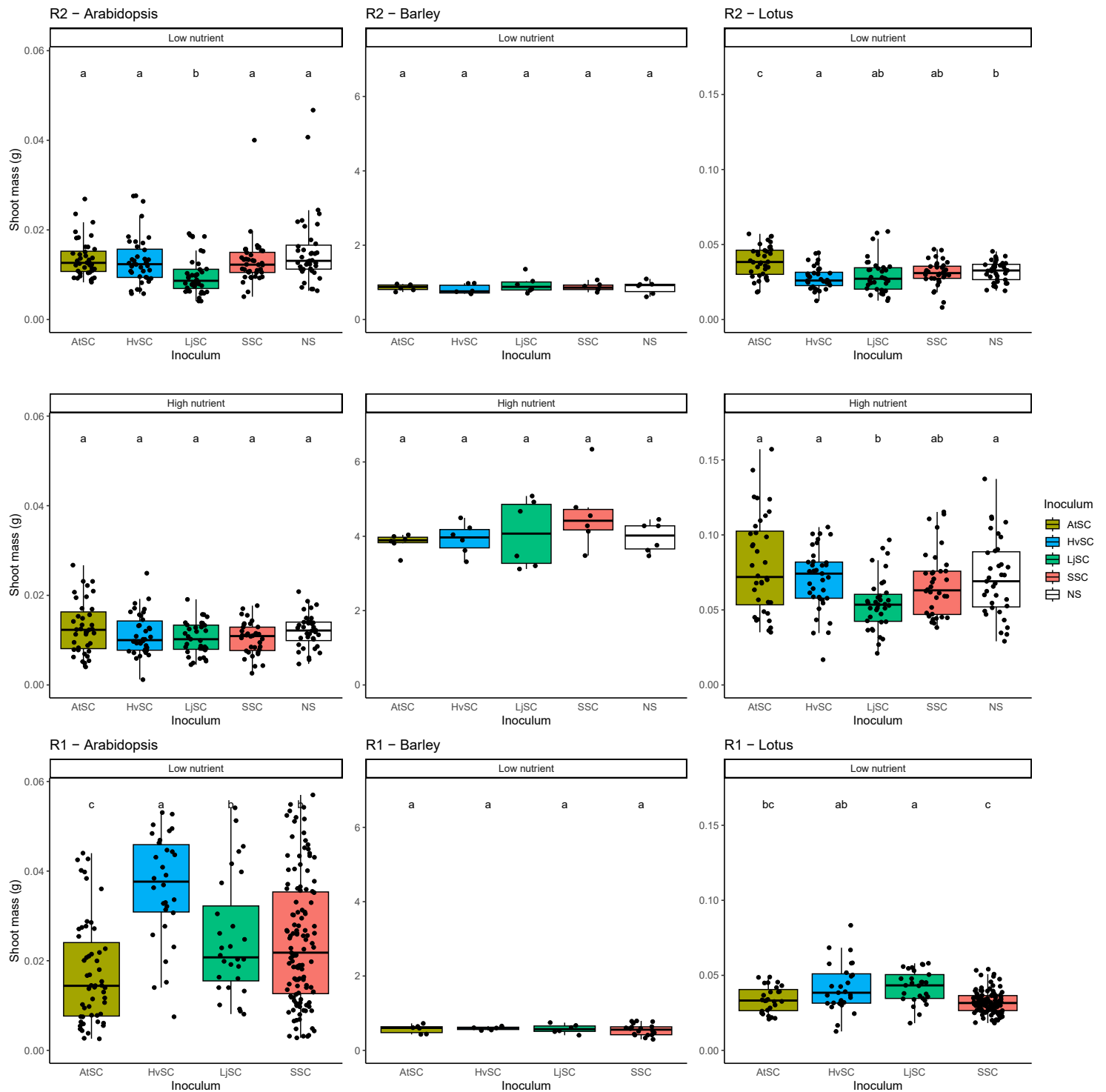
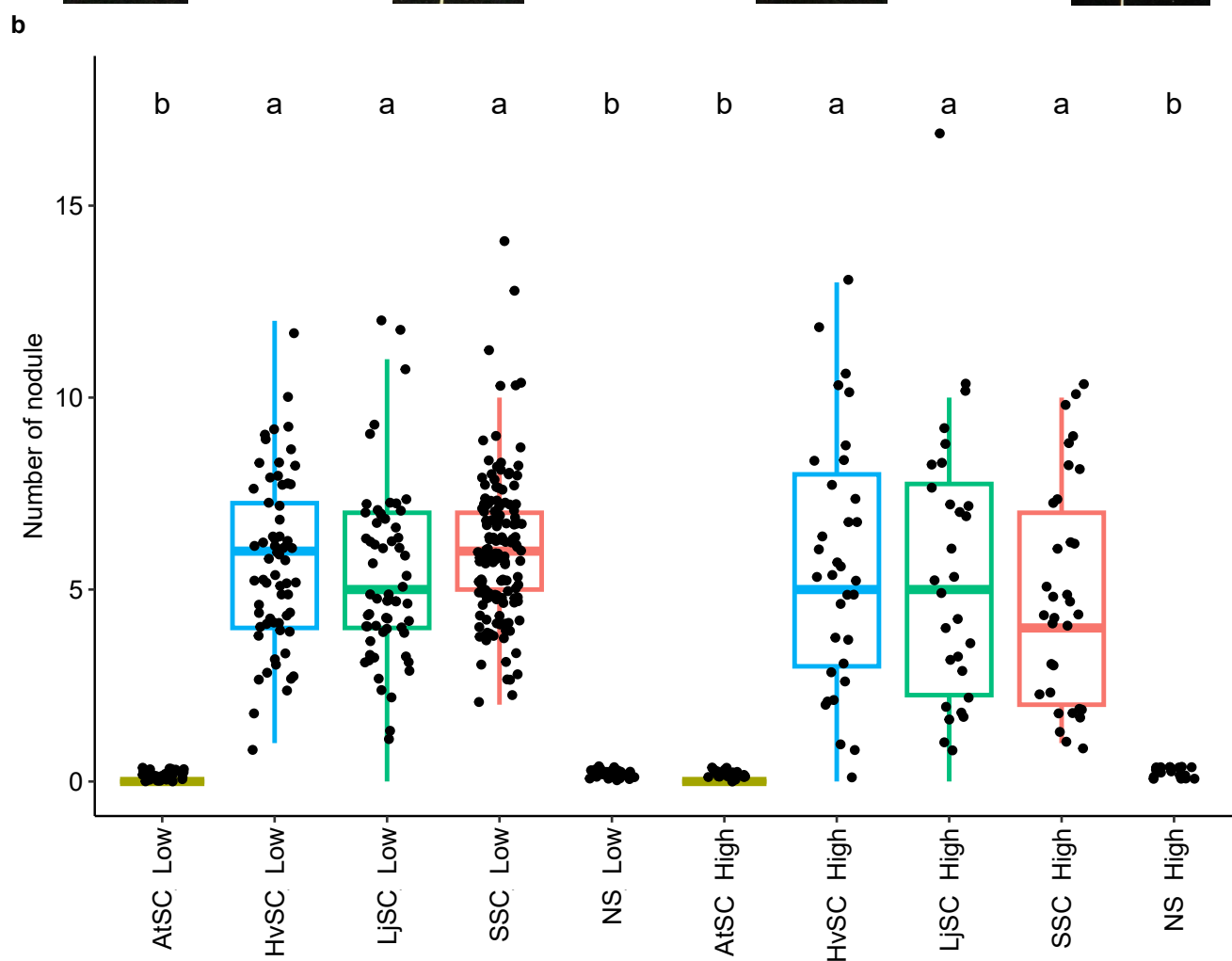
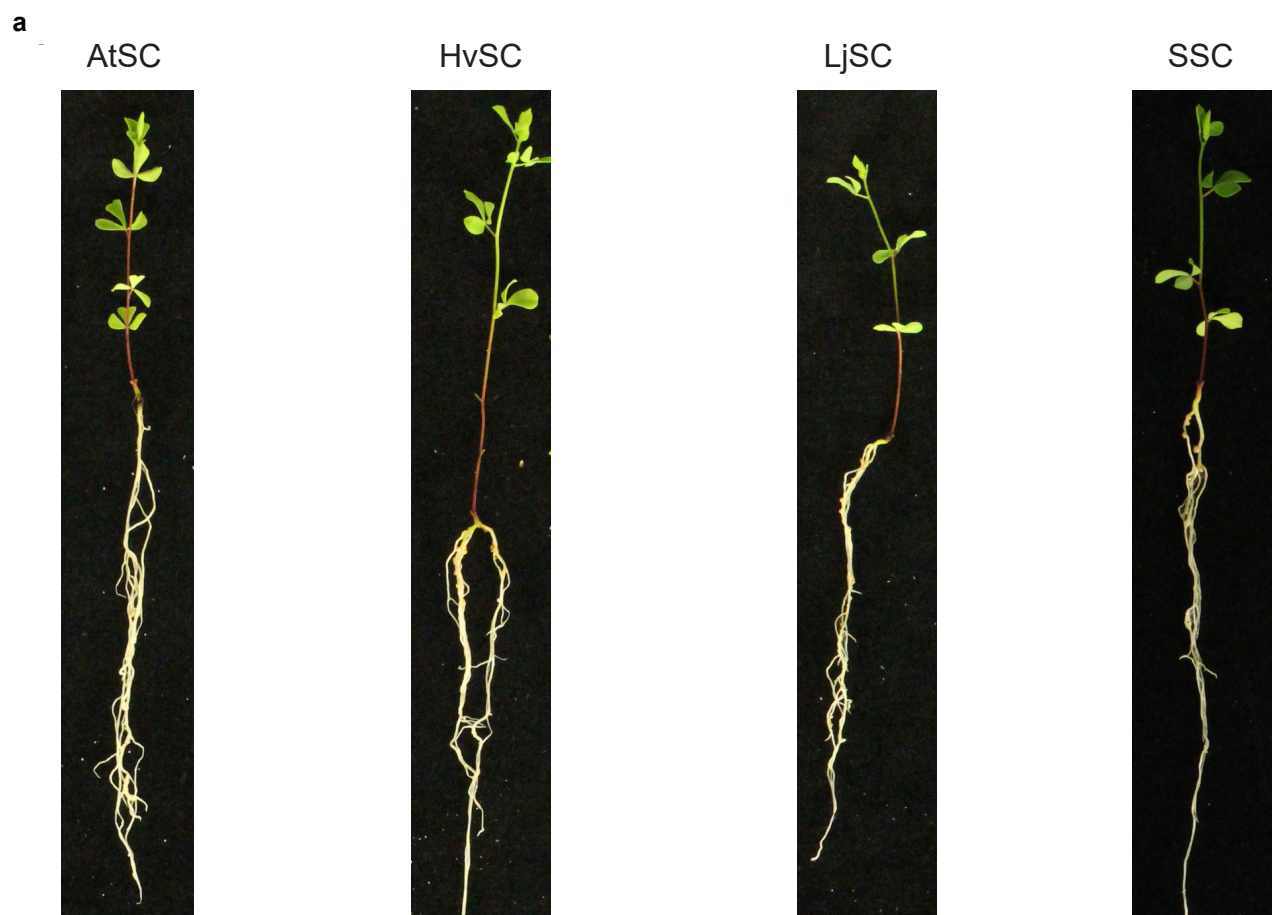


Figure S3. Fresh shoot biomass of Arabidopsis, Barley and Lotus following microbial community inoculations. Fresh shoot biomass for Arabidopsis (left), Barley (middle) and Lotus (right) plants when grown under low nutrient conditions in the second replicate experiment (top), high nutrient conditions in the second replicate experiment (middle), and low nutrient conditions in the first replicate experiment in the presence or absence (NS) of host-derived inocula (AtSC, HvSC and LjSC) or the combined SuperSynCom (SSC) inoculum. Each point represents one plant ($n = 4$ biological replicate pots per treatment containing 10 Arabidopsis, 1 Barley, or 10 Lotus plants per pot). Boxes indicate the median and interquartile range and whiskers indicate the most extreme values within 1.5 x interquartile range. Kruskal-Wallis followed by a Dunn's post-hoc test was conducted to investigate statistical significance between inocula subsetted by host and nutrient condition ($p < 0.05$). Groups not sharing a letter are significantly different. Shoot weights and exact p -values are provided in Source data S13.



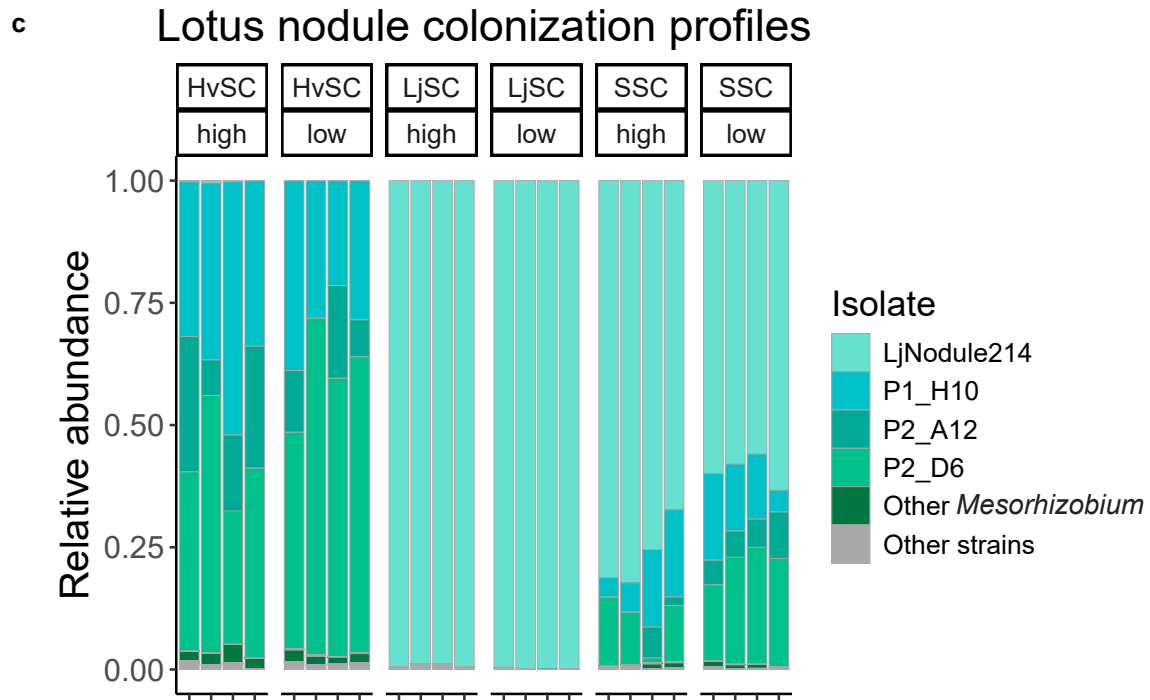


Figure S4. Lotus nodulation phenotypes and nodule microbial composition across inocula. Representative pictures of Lotus plants inoculated with the four inocula under low nutrient conditions (a), nodule number of Lotus when inoculated with the four inocula under low and high nutrient conditions (b), and isolate profiles of nodule samples (c). For a, similar results were obtained in two independent experiments. In b, each point represents one plant ($n = 4$ biological replicate pots per treatment containing 10 Lotus plants per pot). Boxes indicate the median and interquartile range and whiskers indicate the most extreme values within 1.5 x interquartile range. A Kruskal wallis test followed by a Dunn post hoc test was conducted to analyze statistical significance between groups ($p < 0.05$) (Data S13).

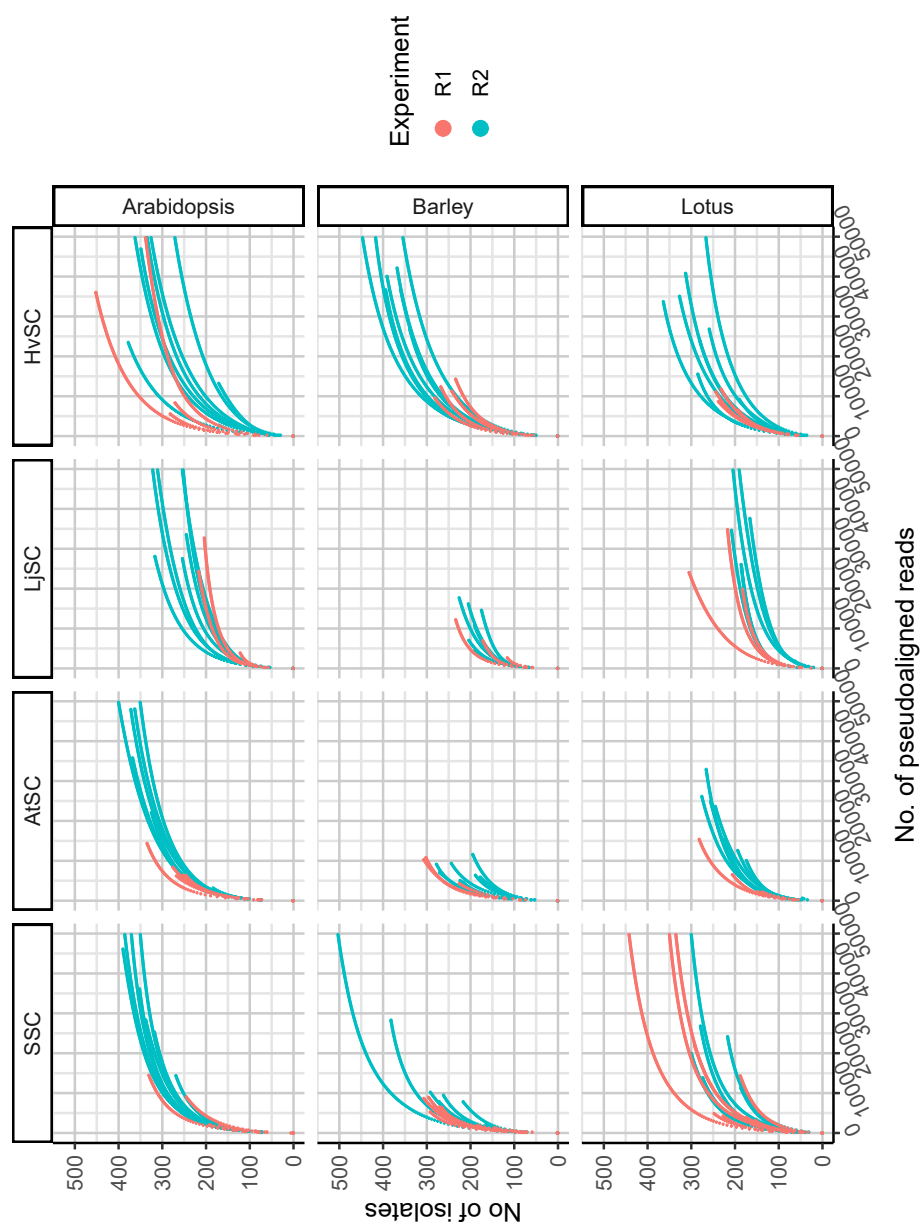


Figure S5. Rarefaction curves of SynCom pseudoaligned reads. For each group of samples (hosts and SynComs) the number of SynCom isolates (y-axis) is shown as a function of increasing microbiome reads (x-axis). Each curve represents one biological sample. Nutrient regimes are not indicated in this plot. Colors indicate whether the curve is from a sample from the first or second SuperSynCom experiment.

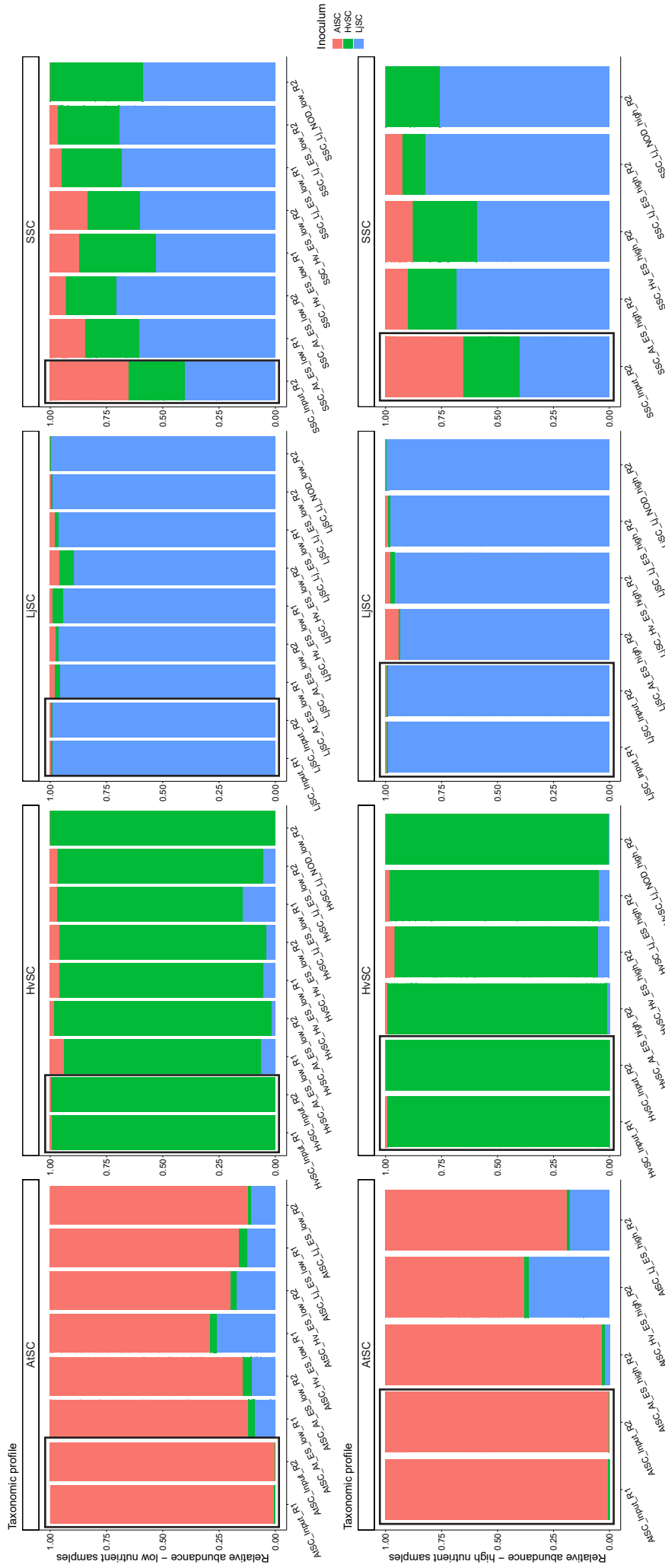


Figure S6. Specificity of read assignment across microbial community inocula. Taxonomic profile of the input inocula (bars in black boxes) and the Arabidopsis, Barley and Lotus root microbiome when inoculated under low nutrient conditions (top) and high nutrient conditions (bottom) with the AtSC, HvSC, LjSC and SSC (left to right). Each bar represents the average of $n=4$ biological replicates and the y-axis indicates relative abundance. The bars are colored by the inoculum that the isolates derive from. The x-axis provides information on the sample's nature (inoculum, host, compartment, nutrient condition, and experiment), for which the meaning can be found in Data S14.

Pseudoalignment of reads to the Salmon indices can lead to read misalignment, with, e.g., reads from AtSC samples pseudoaligning to HvSC or LjSC indices. This does not necessarily mean LjSC or HvSC isolates are contaminating the AtSC samples, as seed endophyte-derived reads could have a high genomic similarity to LjSC or HvSC isolates, which would be reflected in these contamination levels. Despite cross-index assignment of a number of reads, most evident in AtSC samples, we demonstrate that our data contains clean input inocula, and marginal contamination levels in the host-specific inocula.

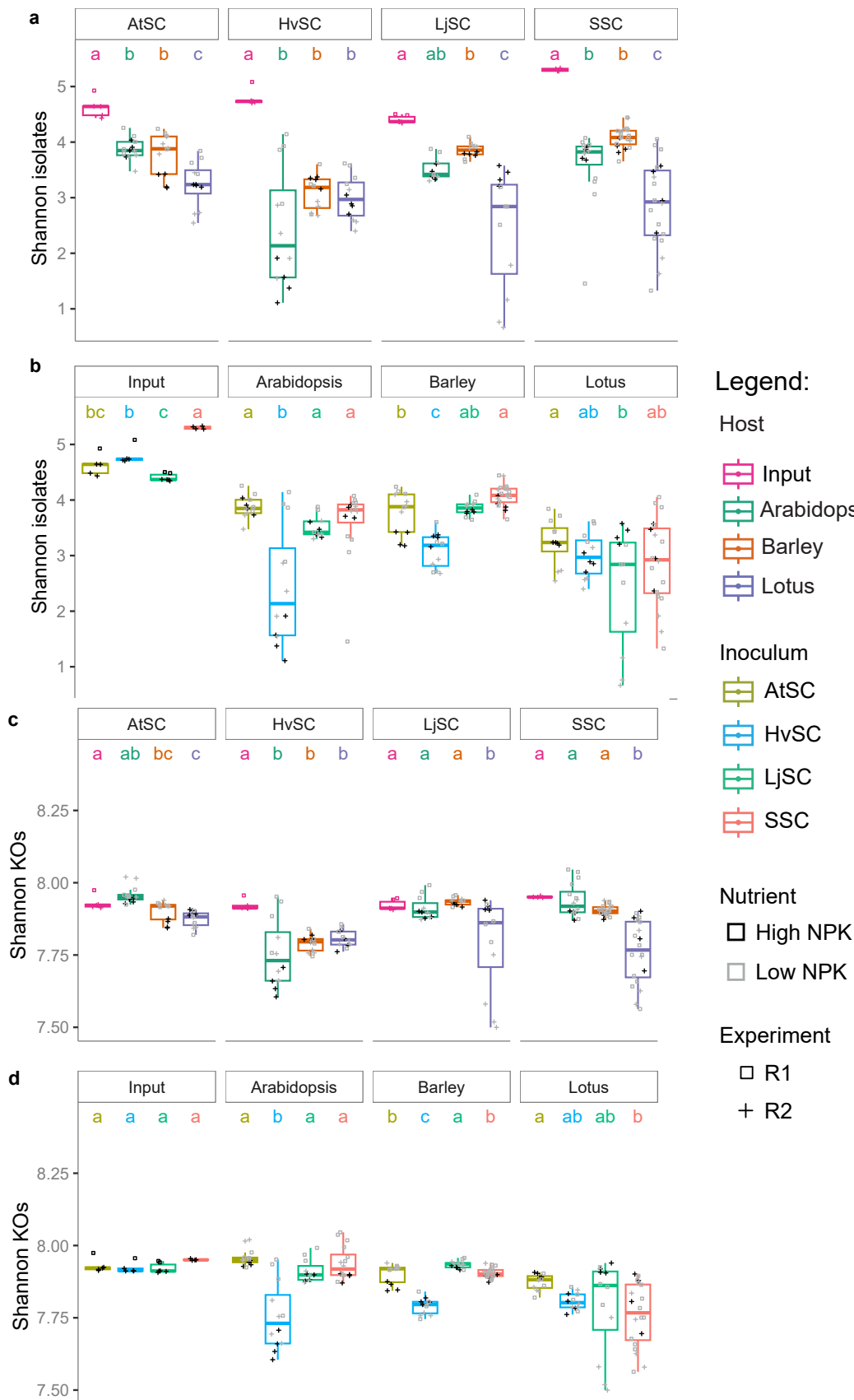
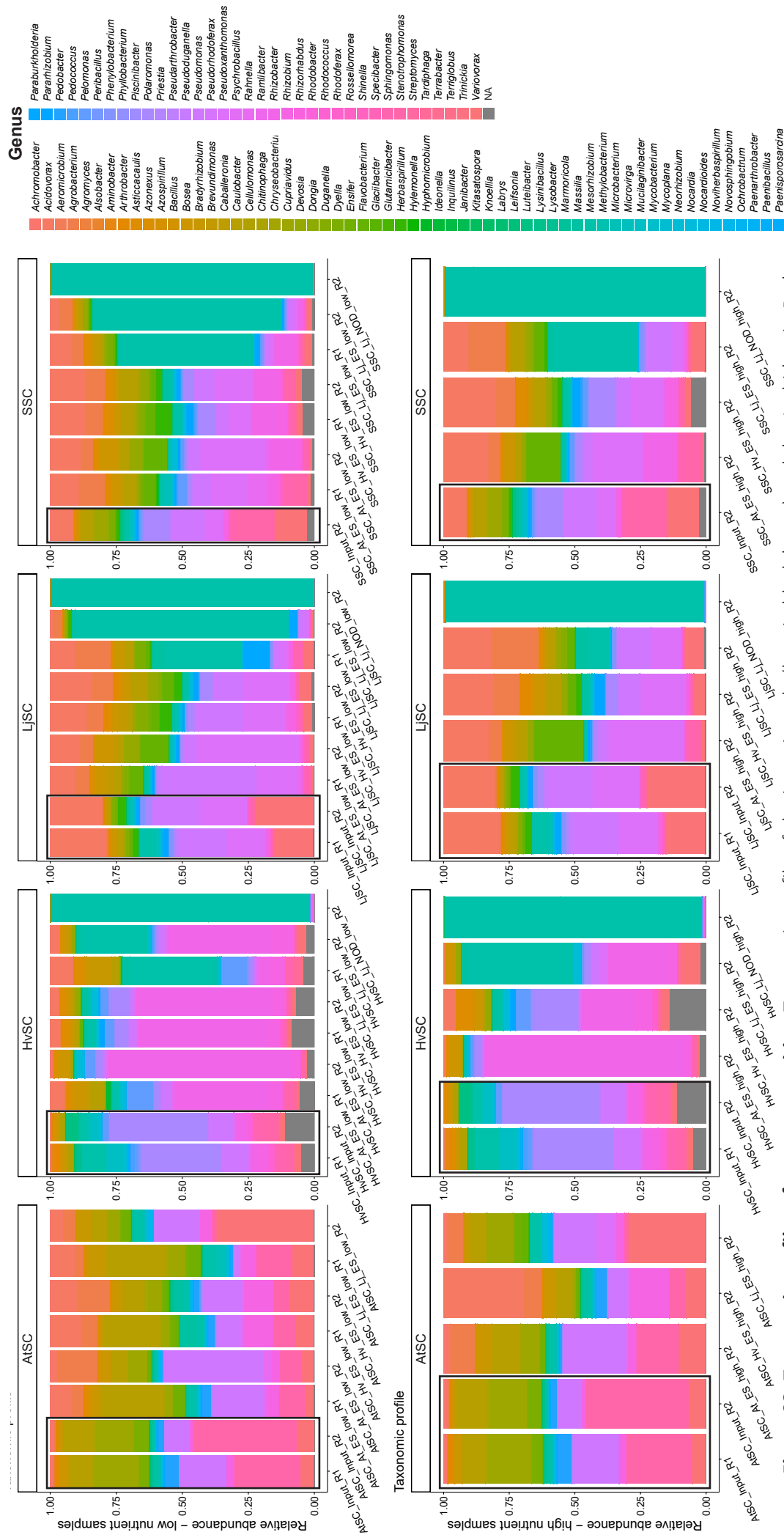


Figure S7. Shannon-based isolate and functional alpha diversity across hosts and inocula. Shannon diversity of microbial communities based on isolates (a and b) or on KOs (c and d) for each host and inoculum. Plots a and b are the same but the groups are grouped either by inoculum (a) or host (b) allowing statistical comparison between groups. This is similar for plots c and d. For each host, $n = 12$ biological replicates were analysed for AtSC, HvSC and LjSC, and $n = 20$ for SSC. Boxes show the median and interquartile range; whiskers extend to the most extreme values within $1.5 \times$ the interquartile range. Individual jittered points represent biological samples. The alpha diversity values were statistically assessed with an ANOVA with post-hoc Tukey HSD test ($p < 0.05$). Groups not sharing a letter are significantly different.



Supplemental results 1 - Nutrient condition effect on diversity analyses

The reduction in the number of isolates and KOs that the roots of Arabidopsis, Barley, and Lotus accommodate compared to the initial inoculum is substantial (Figure 2b-c). The effect of nutrient condition on root microbiome assembly in this analysis was found to be small, although it appears to contribute differentially in specific host and inoculum contexts (Figure S9). This limited effect in relation to host and inoculum effect is comparable to recently published studies (Wang et al., 2024; Zhang et al., 2022), suggesting that nutrient levels in these conditions are not the dominant factor explaining microbiome assembly. Nonetheless, specific nutrient-driven changes in plant and bacterial physiology likely exist but these are not the primary focus of this study. The low-nutrient condition seems to positively affect the number of isolates and KOs, most evident in the Arabidopsis and Lotus contexts. In the context of a host or inoculum-dependent enrichment of specific community members—illustrated by the enrichment of Lotus symbionts in the low-nutrient Lotus-context and the enrichment of *Rhizobacter* sp. P2_G4 in the high-nutrient HvSC-context—the communities are more uneven in terms of isolates and functions (Figure S9).

These findings suggest that a low-nutrient environment can enrich for increased taxonomic and functional diversity, which could be related to an increased nutritional dependence of the plant on its microbiota (Singh et al., 2022). In addition, oligotrophs may be prevented from dominating the microbiome in low-nutrient, scarce conditions, leading to more diverse communities (Wang et al., 2018; Zeng et al., 2016). However, in contrast, long-term nutrient enrichment has also been shown to increase bacterial diversity (Bledsoe et al., 2020). The only exception to the correlation between nutrient concentration and bacterial diversity is when Lotus is grown in low-nutrient conditions, inevitably leading to the recruitment of symbionts to engage in nitrogen-fixing symbiosis, greatly affecting the microbiome composition (Figure S9) (Zgadza et al., 2016).

In line with the findings described in the main text and in Figure 2f-g, we demonstrate that the variation in microbiome composition among samples with the same treatment (host, inoculum, and nutrient condition) is lower for the functional composition than for the genus composition (Figure S10). This indicates that functional convergence not only occurs between hosts and inocula, but also within samples from the same treatments. The variation in microbiome composition between nutrient conditions seems to be largely similar between high- and low-nutrient conditions, with slightly higher variation in the low-nutrient condition (Figure S10). This supports the argument that the host's interactions with its microbiota are more stochastic when nutrients are limited, thus leading to more variation and diversity, though this difference between high- and low-nutrient conditions is relatively small. Interestingly, across inocula and nutrient conditions, more variation exists in the Arabidopsis and Lotus root microbiome as compared to the Barley root microbiome. Interestingly, groups of samples that have a relatively high compositional variation appear to be characterized by the presence of dominant strains such as nodulating *Mesorhizobium* strains or a dominant *Rhizobacter* strain (Extended Data 1, S10). This observation suggests that dominant strains can lead to larger variation in bacterial and functional composition, most likely driven by their own variable relative abundances. The Barley root microbiome harbors the highest taxonomic and functional diversity (Figure 2b, d). However, this remains relatively stable across nutrient conditions (Figures S9, S10) even if Barley plants show clear response to the amount of supplemented nutrients (Figure S2).

One potential driver of the observed differences in microbiome assembly characteristics across these plant species might be the domestication processes to which cereals like Barley have been subjected. Breeding practices can negatively impact root microbiota diversity in crops, as they may disrupt a million-year-long evolutionary process (Edwards et al., 2015; Peiffer et al., 2013). Conversely, studies on domesticated grass species demonstrate that microbial diversity may remain unaffected by domestication (Naylor et al., 2017; Ling et al., 2022).

References

- Bledsoe, R. B., Goodwillie, C., & Peralta, A. L. (2020). Long-term nutrient enrichment of an oligotroph-dominated wetland increases Bacterial Diversity in bulk soils and plant rhizospheres. *MSphere*, **5**.
- Edwards, J., Johnson, C., Santos-Medellín, C., Lurie, E., Podishetty, N. K., Bhatnagar, S., Eisen, J. A., & Sundaresan, V. (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. *Proceedings of the National Academy of Sciences*, **112**.
- Ling, N., Wang, T., & Kuzyakov, Y. (2022). Rhizosphere bacteriome structure and functions. *Nature Communications*, **13**, 836.
- Naylor, D., DeGraaf, S., Purdom, E., & Coleman-Derr, D. (2017). Drought and host selection influence bacterial community dynamics in the grass root microbiome. *The ISME Journal*, **11**, 2691–2704.
- Peiffer, J. A., Spor, A., Koren, O., Jin, Z., Tringe, S. G., Dangl, J. L., Buckler, E. S., & Ley, R. E. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proceedings of the National Academy of Sciences*, **110**, 6548–6553.
- Singh, S. K., Wu, X., Shao, C., & Zhang, H. (2022). Microbial enhancement of plant nutrient acquisition. *Stress Biology*, **2**, 3.
- Wang, C., Liu, D., & Bai, E. (2018). Decreasing soil microbial diversity is associated with decreasing microbial biomass under nitrogen addition. *Soil Biology and Biochemistry*, **120**, 126–133.
- Wang, M., Ge, A.-H., Ma, X., Wang, X., Xie, Q., Wang, L., Song, X., Jiang, M., Yang, W., Murray, J. D., Wang, Y., Liu, H., Cao, X., & Wang, E. (2024). Dynamic root microbiome sustains soybean productivity under unbalanced fertilization. *Nature Communications*, **15**, 1668.
- Zeng, J., Liu, X., Song, L., Lin, X., Zhang, H., Shen, C., & Chu, H. (2016). Nitrogen fertilization directly affects soil bacterial diversity and indirectly affects bacterial community composition. *Soil Biology and Biochemistry*, **92**, 41–49.
- Zgadaj, R., Garrido-Oter, R., Jensen, D. B., Koprivova, A., Schulze-Lefert, P., & Radutoiu, S. (2016). Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. *Proceedings of the National Academy of Sciences*, **113**.
- Zhang, L., Zhang, M., Huang, S., Li, L., Gao, Q., Wang, Y., Zhang, S., Huang, S., Yuan, L., Wen, Y., Liu, K., Yu, X., Li, D., Zhang, L., Xu, X., Wei, H., He, P., Zhou, W., Philippot, L., & Ai, C. (2022). A highly conserved core bacterial microbiota with nitrogen-fixation capacity inhabits the xylem sap in maize plants. *Nature Communications*, **13**, 3361.

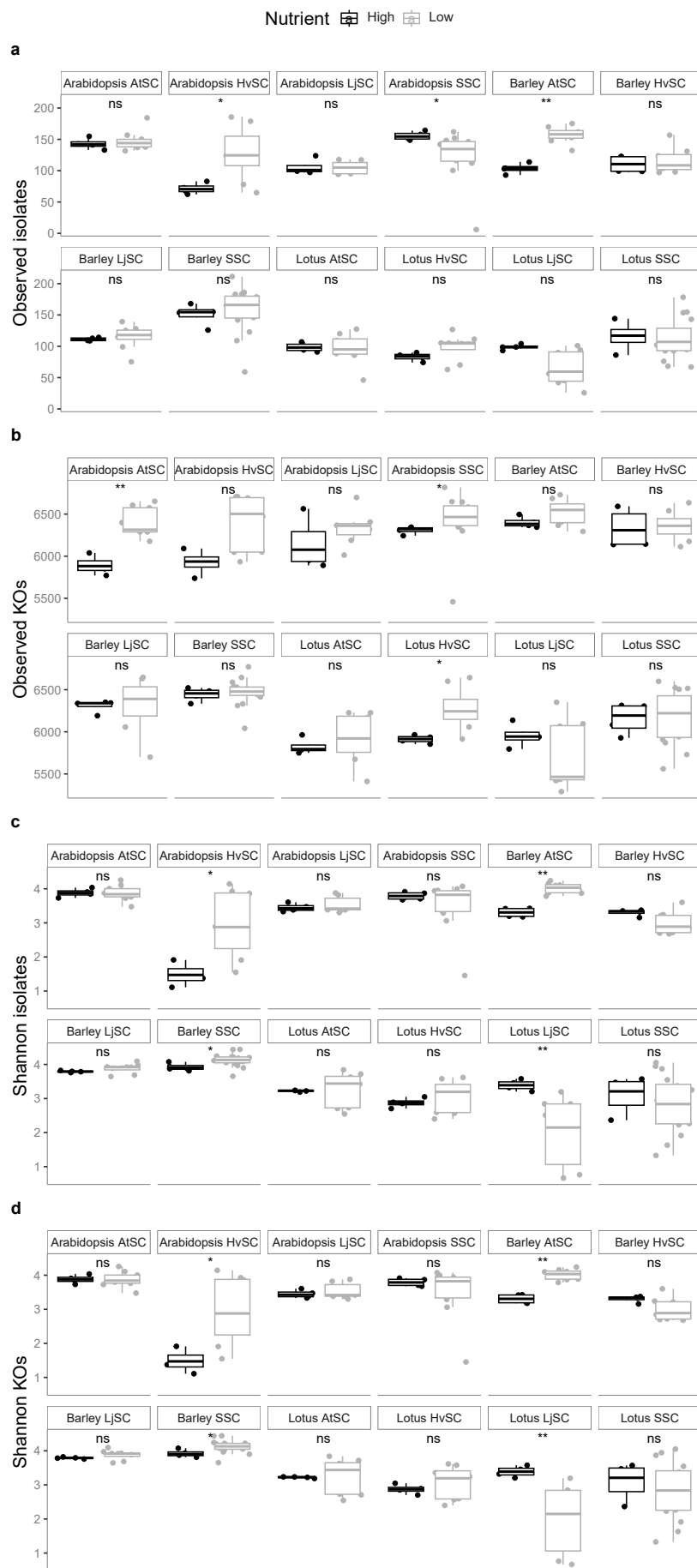


Figure S9. Nutrient condition effect on alpha diversity. The data is subsetting to show the observed isolates (a), observed KOs (b), Shannon diversity for isolates (c) and Shannon diversity for KOs (d) between low and high nutrient conditions. For each host, $n = 6$ biological replicates were analysed for AtSC, HvSC and LjSC, and $n = 10$ for SSC. Boxes show the median and interquartile range; whiskers extend to the most extreme values within $1.5 \times$ the interquartile range, while individual jittered points represent biological samples. Statistical differences are indicated by * assessed by a Wilcoxon test (p -value < 0.05).

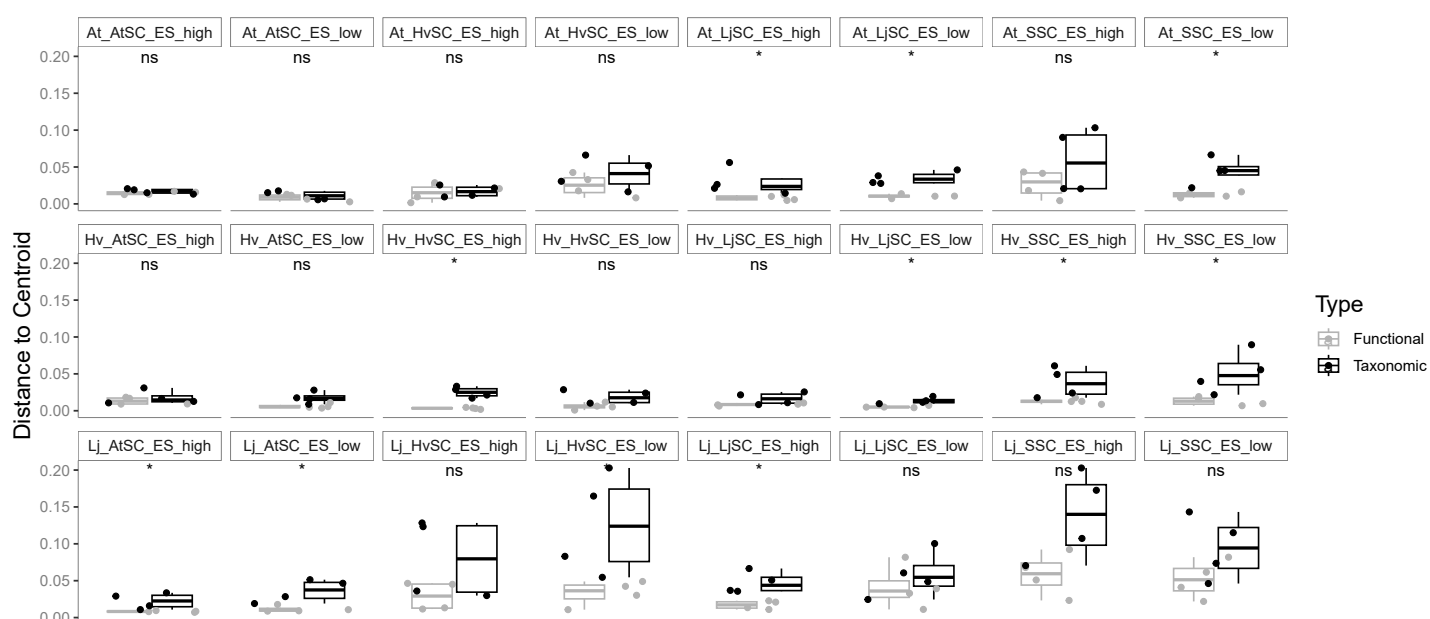
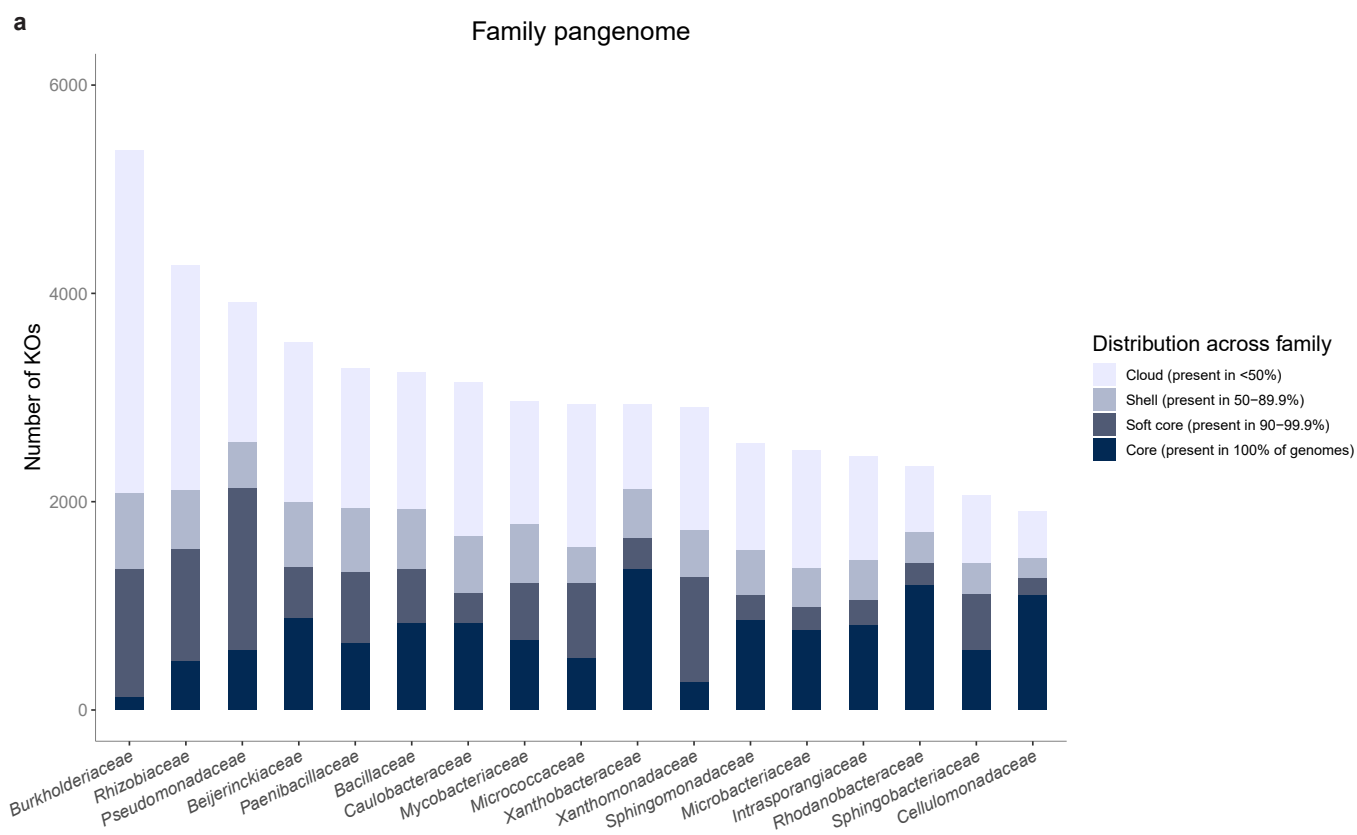


Figure S10. Distance to centroid for genera and functions. Each boxplot panel indicates the distance between the individual samples and the centroid of the whole group of samples for both the genus (taxonomic) and KO (functional) composition shown in Figure 2f-g (only for the SSC experiment R2). For each treatment, $n = 4$ biological replicates were used. Boxes show the median and interquartile range; whiskers extend to the most extreme values within $1.5 \times$ the interquartile range. Individual jittered points represent biological samples, and values beyond the whiskers are shown individually. A high distance indicates a greater compositional variation between replicates of the same treatment. Statistical differences are indicated by * assessed by a Wilcoxon test ($p < 0.05$). The panel titles provide information on the sample's nature (host, inoculum, compartment, and nutrient condition), for which the meaning can be found in Data S14.



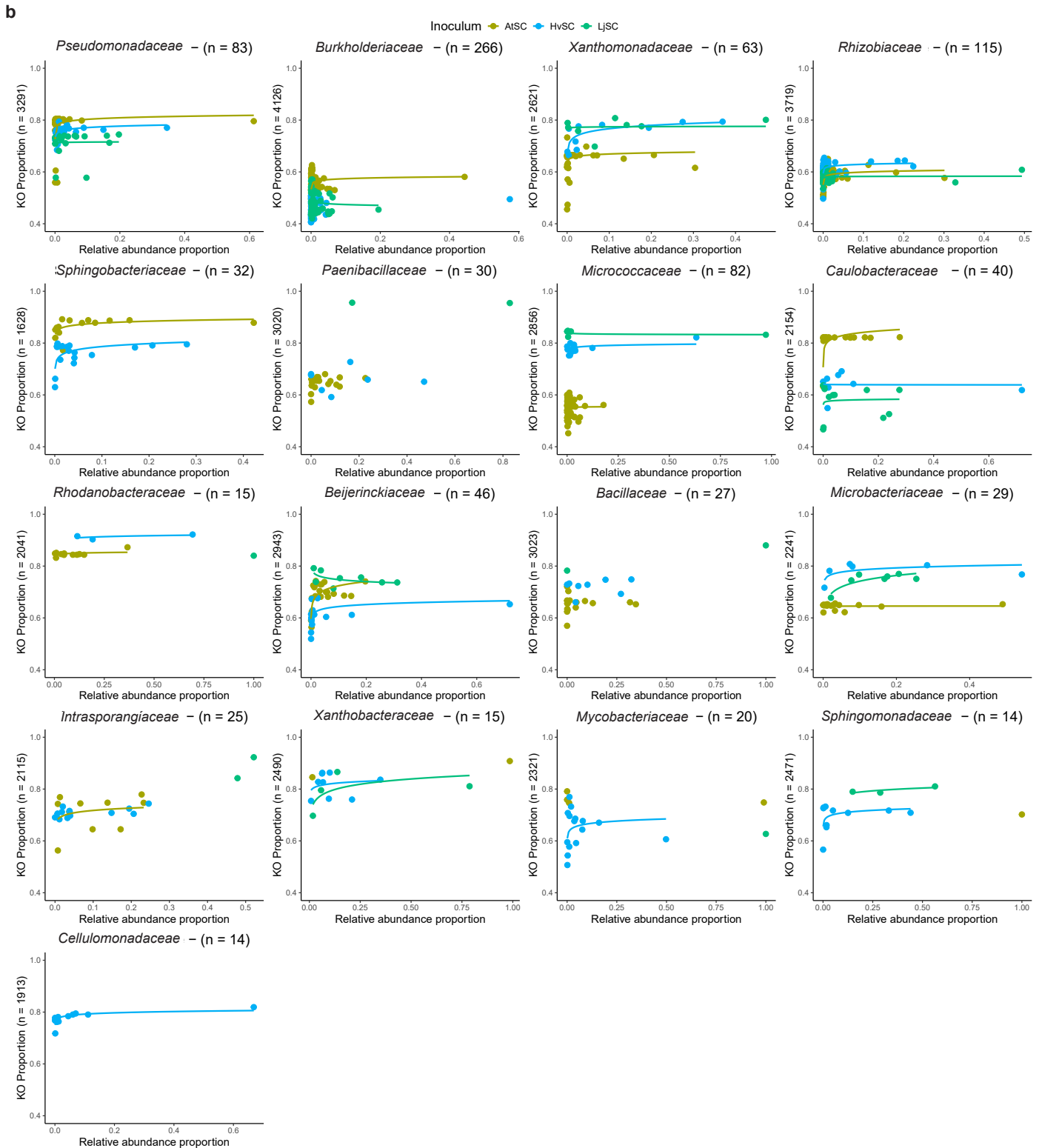


Figure S11. Functional variation and diversity in bacterial families. (a) Functional variation indicated by core, soft core, shell, and cloud KOs within each bacterial family. (b) The relation between genomic diversity of isolates (y-axis) and the relative abundance (x-axis) in the dataset subsetted by family. The KO proportion indicates the number of KOs an isolate has as compared to the number of KOs in the family pangenome. The relative abundance proportion indicates the isolate's relative abundance across all host microbiomes but for each inoculum separately as compared to the family's relative abundance. Each point represents one bacterial isolate. Statistical significance and exact p -values are provided in Table S6. The numbers in the brackets indicate the number (n) of isolates in the family (plot titles) and number (n) of KOs in the family pangenome (y-axes). Colors indicate from which inoculum the isolates originate. Loglinear correlations calculated in each these plots are displayed in the heatmap of Figure 3d.

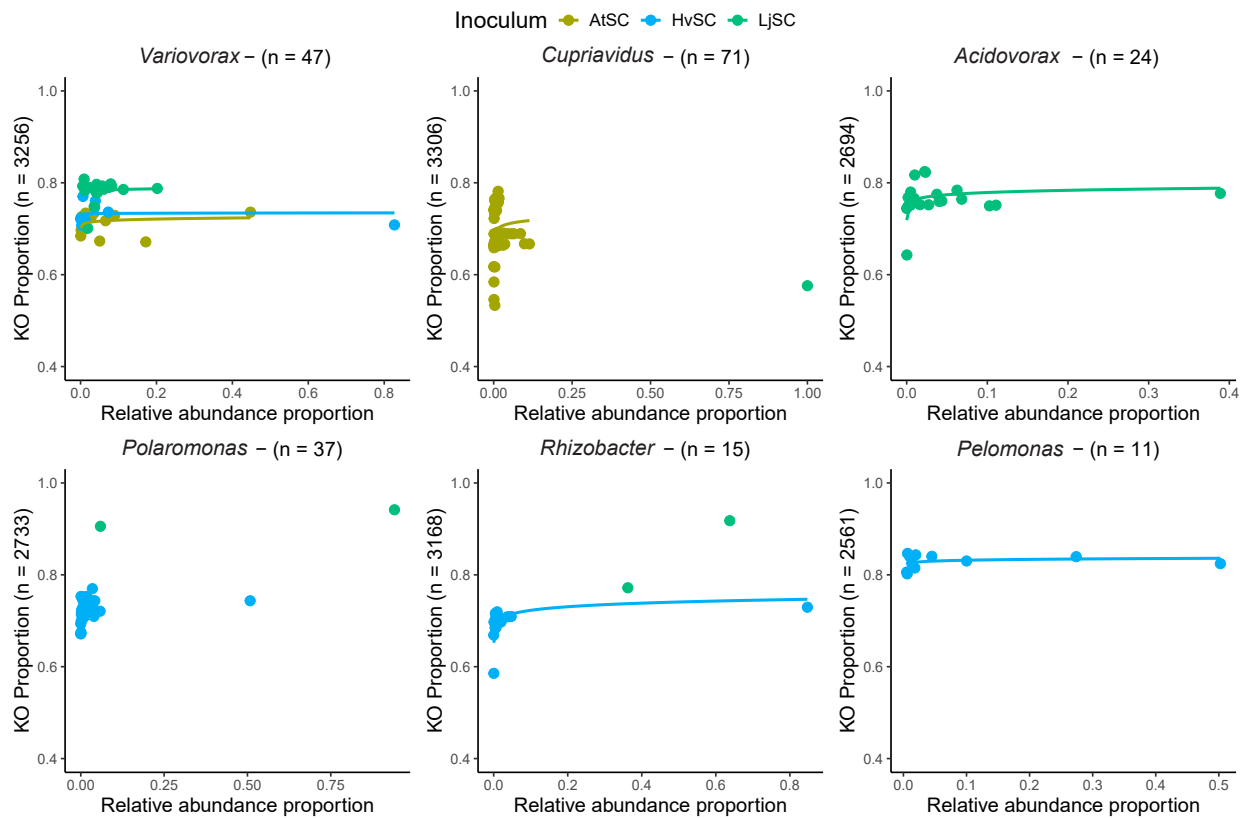
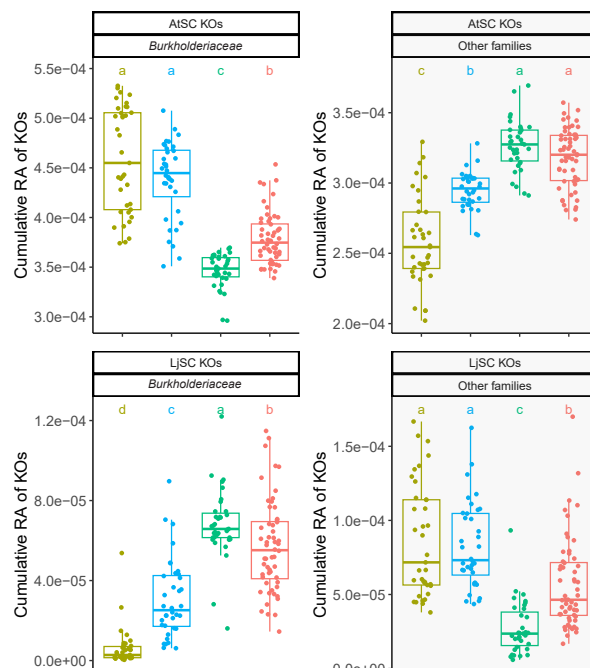
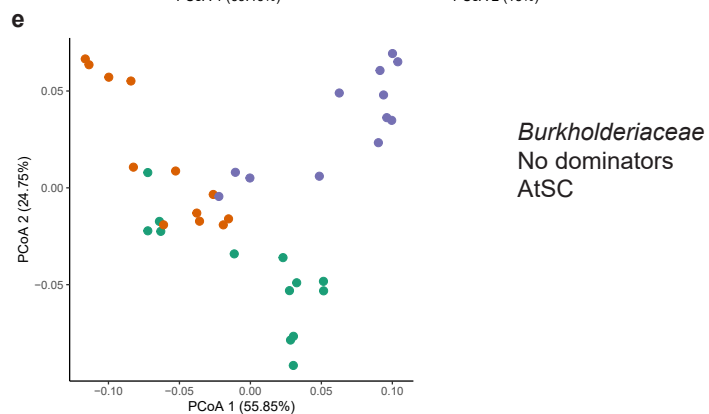
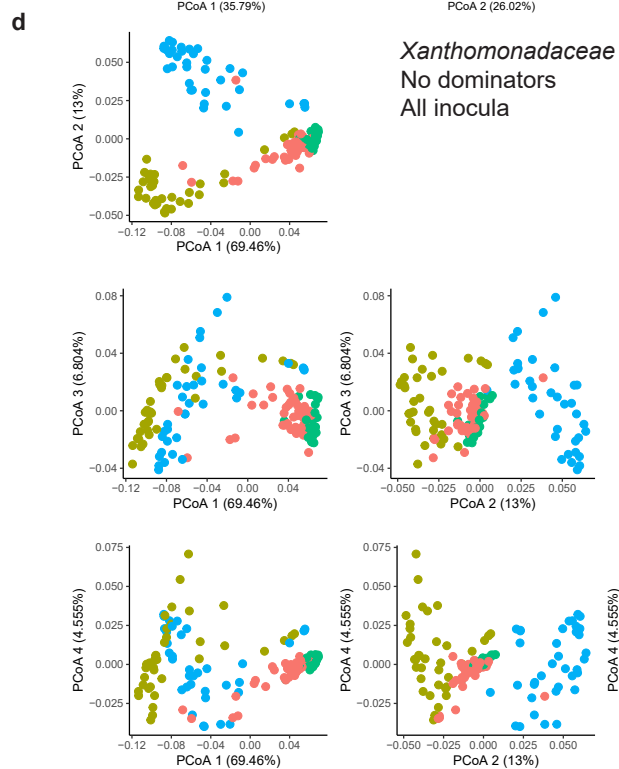
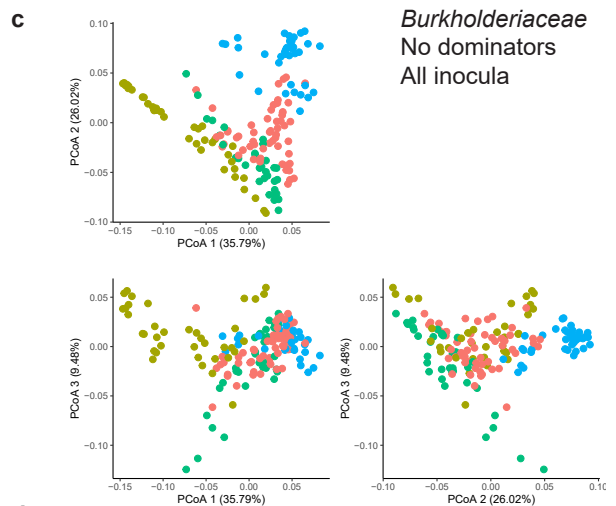
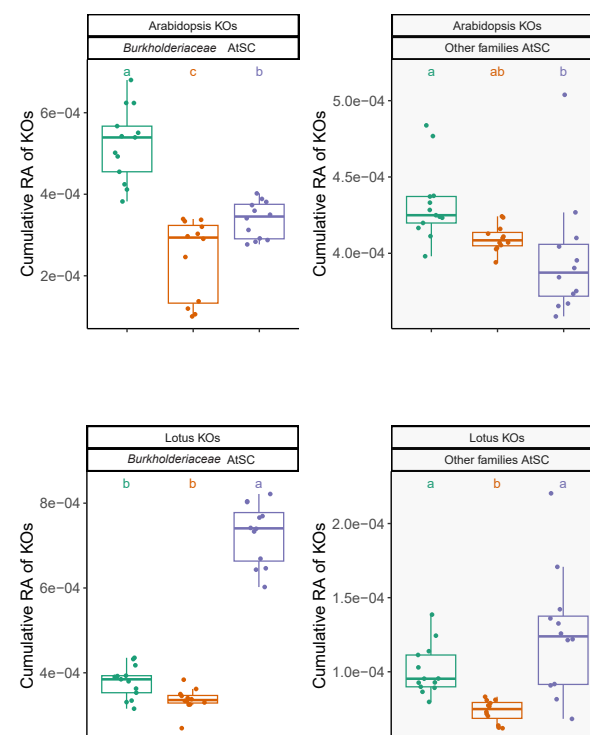
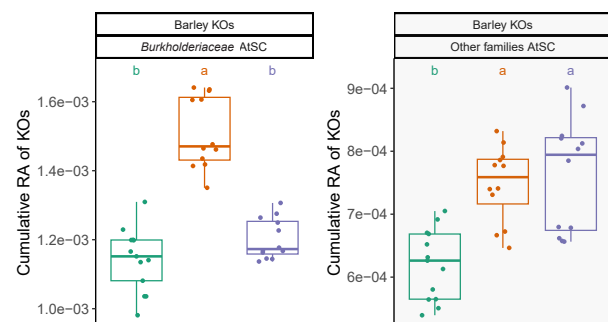


Figure S12. Correlation KO diversity and relative abundance per genus in the largest family. The relation between genomic diversity of isolates (y-axis) and the relative abundance (x-axis) in the dataset subsetted per genus in the *Burkholderiaceae*. The KO proportion indicates the number of KOs in an isolate compared to the number of KOs in the genus pangenome. The relative abundance proportion indicates the isolate's relative abundance across all host microbiomes but for each inoculum separately as compared to the genus' relative abundance. Each point represents one bacterial isolate. Statistical significance and exact *p*-values are provided in Table S6. The numbers in the brackets indicate the number (*n*) of isolates in the family (plot titles) and number (*n*) of KOs in the family pangenome (y-axes). Loglinear correlations calculated in each these plots are displayed in the heatmap of Figure 3d. The colors indicate from which inoculum the isolate derives from.



● Arabidopsis
● Barley
● Lotus



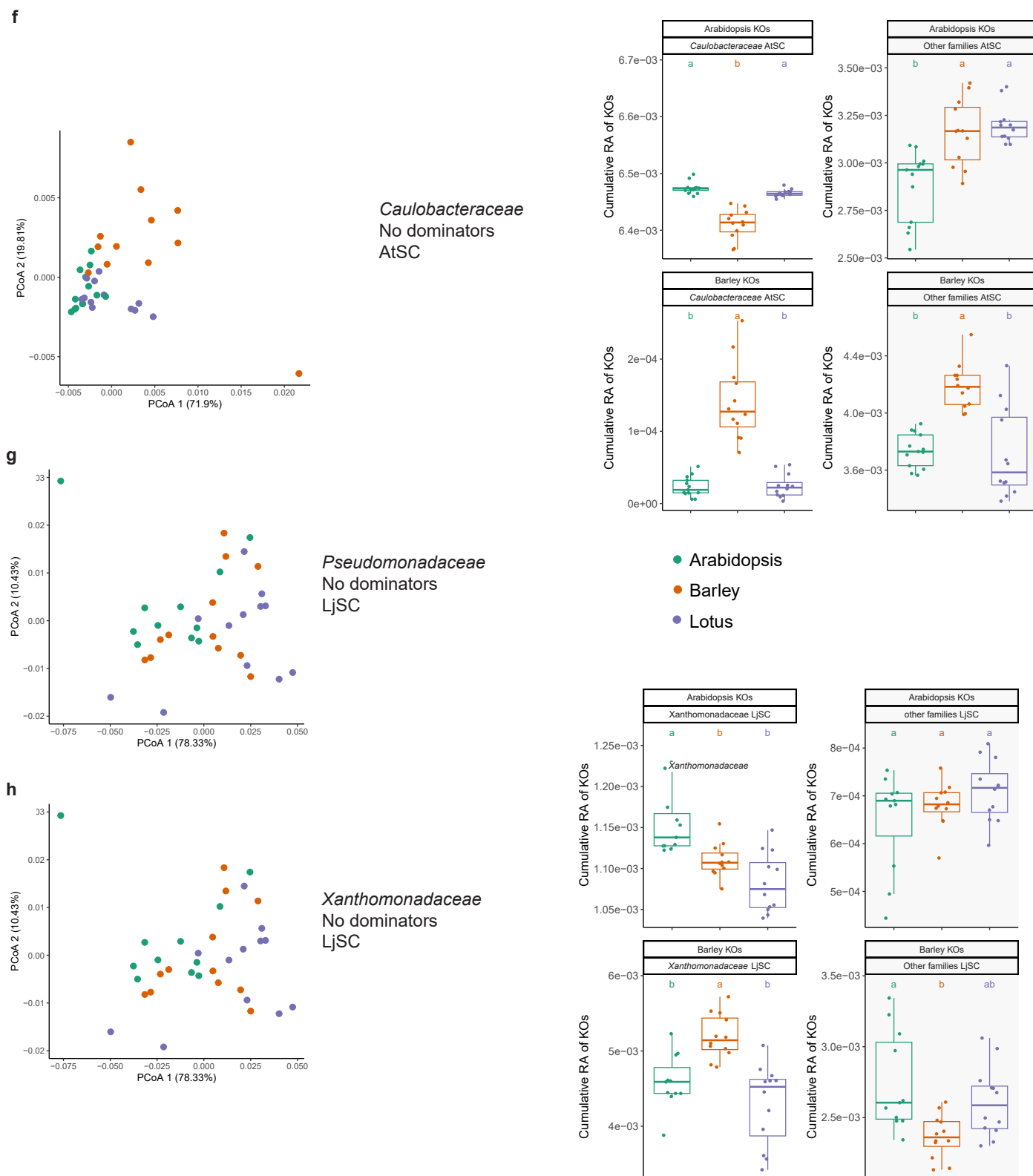


Figure S13. Contribution of KOs to family-subsetted SynCom data. (a) Contribution of the different bacterial families to the observed community-level split by plant host (right) and inoculum (left) and in the absence of dominator isolates (tainted bars) - full figure version of 4a. Family Bray-Curtis PCoA on the KO composition when the data is subsetting for a single family; *Rhizobiaceae* (b), *Burkholderiaceae* (c, e), *Xanthomonadaceae* (d, h), *Caulobacteraceae* (f), and *Pseudomonadaceae* (g). The data was subsetting according to the dataset or SynComs in which the families have a large contribution to the compositional differences (Figure 4b). For c, e-f, h, box plots for the relative abundance of contributing KOs are shown, similar to Figure 4c. Each point in the PCoA and boxplots indicates a biological replicate with $n = 12$ for plant comparisons and $n = 36$ for AtSC, HvSC, and LjSC for SynCom comparisons, and $n = 60$ for SSC for SynCom comparisons. Boxes indicate the median and interquartile range; whiskers indicate the most extreme values within $1.5 \times$ IQR. The relative abundances are taken from the KOs that contribute to the compositional differences as indicated in Figure 4b. The relative abundances per group were statistically assessed with an ANOVA followed by a post-hoc Tukey HSD test ($p < 0.05$). Contribution values in a and exact p -values in b can be found in Source Data S22 and S23 respectively.

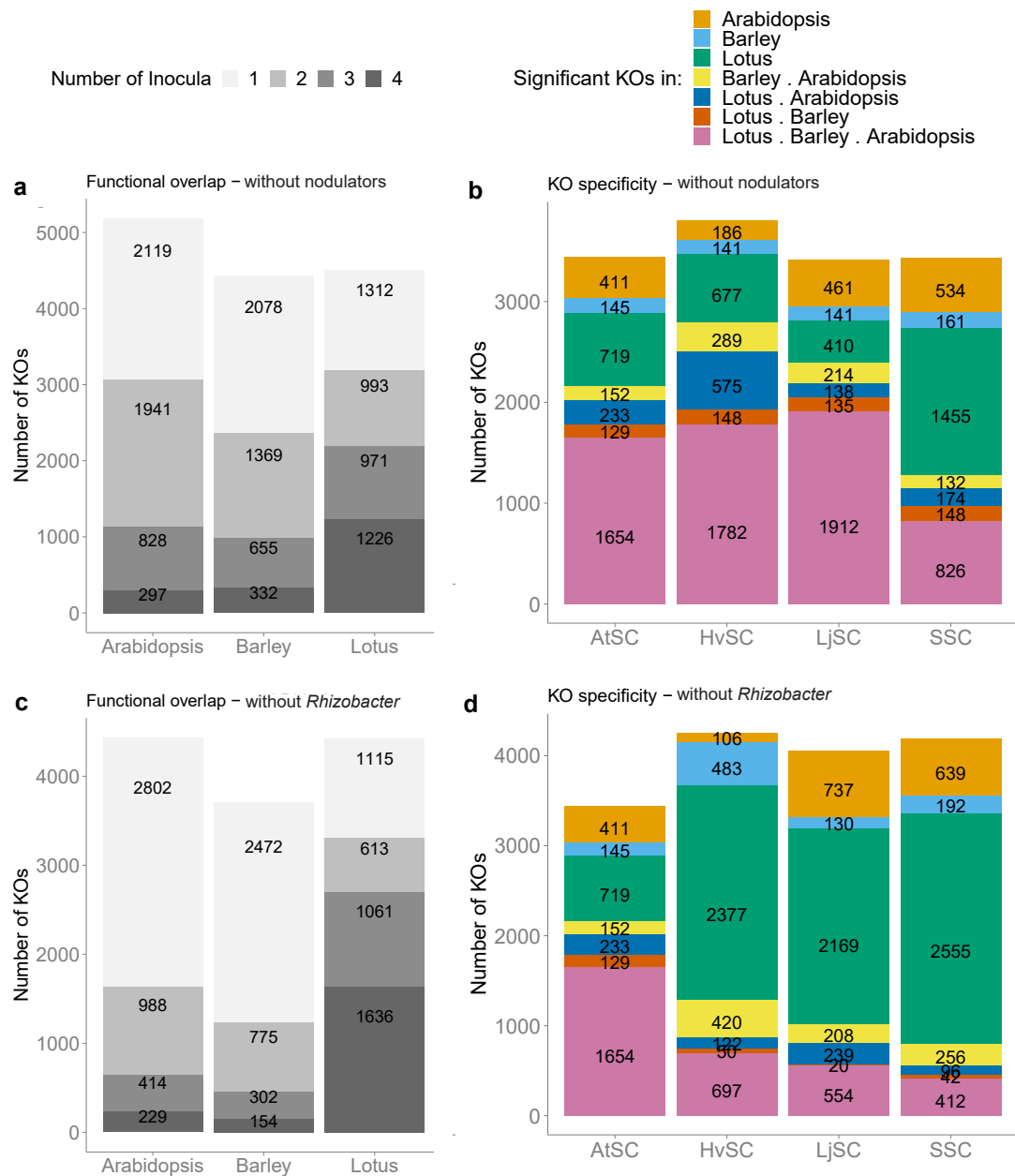


Figure S14. Overlap and number of root microbiome-enriched KOs across hosts and inocula. The overlap in differentially abundant bacterial KOs recruited by Arabidopsis, Barley or Lotus compared to the initial inoculum. Overlap between the four inocula and the three hosts shown in number of differentially abundant KOs in the dataset without the Lotus symbionts (a-b respectively), or the dataset without the *Rhizobacter* sp. P2_G4 (c and d respectively).

Supplemental results 2 – Host-specific functions

When examining host-dependent bacterial functions, we found that the number of Lotus-specific bacterial genes (355) greatly exceeds that of Arabidopsis-specific (48) and Barley-specific (40) bacterial genes (Figure 5c-d, Data S26). This suggests that Lotus exerts stronger selective pressure compared to the other two hosts. To identify the most relevant host-specific bacterial functions, we focused on functions from gene cassettes in which most or all genes consistently exhibited host-specificity in our analysis (Data S26).

Arabidopsis-specific bacterial genetic features - bacteriochlorophyllides

We observed that genes within the photosynthetic bacteriochlorophyllide biosynthesis gene cluster (*bch*) were enriched specifically in Arabidopsis-associated communities (Extended Data 5, Data S26). Notably, 11 out of 18 *bch* genes showed Arabidopsis-specific enrichment (Extended Data 5a), especially when inoculated with the SSC (Extended Data 5b). This finding is surprising because bacteriochlorophyllide is involved in harnessing light energy for photosynthesis, yet light penetrates only about 1 cm into the soil, reaching only a small fraction of the Arabidopsis root (Wu *et al.*, 2014). However, a positive correlation between soil moisture and light penetration depth has been observed (Williams *et al.*, 2016). This correlation is consistent with the fact that the Arabidopsis plants were kept covered during the experiment, likely affecting soil moisture. Bacteria encoding this photosynthetic gene cluster mostly derive from the *Flavobacteriaceae*, *Pseudomonadaceae*, and *Burkholderiaceae* families (Extended Data 5c), representing common colonizers of Arabidopsis roots (Pieterse *et al.*, 2021; Thiergart *et al.*, 2019; Wang *et al.*, 2021). We observed that the abundance of two *bch* genes, *bchE* and *bchU*, was markedly larger for Lotus. BchE has been reported to be mainly active in strictly anaerobic photosynthetic bacteria (Ouchane *et al.*, 2004), while bchU converts chlorophyllide *a* to chlorophyll, a more efficient chlorophyllide due to its ability to harvest energy from a greater bandwidth of light (Malina *et al.*, 2021). This suggests that photosynthetic traits are Arabidopsis-specific, though photosynthetic isolates that are strictly anaerobic and have more efficient light-harvesting abilities could be Lotus-specific. Alternatively, imprecise, or wrongful annotation of *bchE* homologues could be responsible for this observation, and its high abundance likely underlies the unexpected Lotus, as opposed to Arabidopsis, specificity when isolate abundances are taken into consideration (Figure Extended Data 5d).

Barley-specific bacterial genetic features – exopolysaccharide succinoglycans

Among the Barley-specific genes, multiple genes can be found that constitute thiamine and inositol transporters, as well as two succinoglycan biosynthesis genes—*exoV* and *exoZ*—when inoculated with the AtSC (Extended Data 6a-c, Data S26). These two *exo* genes derive mainly from *Rhizobiaceae* isolates (Extended Data 6d). Succinoglycans are extracellular polysaccharides associated with the bacterial cell wall, with low-molecular-weight forms in particular playing a role in nodule infection (Maillet *et al.*, 2020; Mendis *et al.*, 2016; Skorupska *et al.*, 2006). ExoV and exoZ add pyruvate and acetyl groups to the end of the polysaccharide chain (Extended Data 6a), therewith stabilizing it and leading to a higher proportion of high-molecular-weight forms (Jeong *et al.*, 2022; York & Walker, 1998). The high relative abundance of both genes in Barley-associated communities indicates that these high molecular weight succinoglycans could be important for Barley root colonization (Extended Data 6a).

Lotus-specific bacterial genetic features – symbiosis-related

Lotus shows a substantially higher amount of enriched host-specific genes as compared to Arabidopsis or Barley (Data S26). Nonetheless, in the absence of symbionts, only a fraction of those remains (Data S27) (61 out of the 355). In the dataset including symbionts, Lotus shows, as predicted, an enrichment of *nif* and *nod* genes in its root microbiome, which are involved in nitrogen fixation and nodulation, respectively (Göttfert, 1993; Spaink *et al.*, 2012) (Figures S15, S16, Data S26). The *nif* and *nod* gene clusters are entirely Lotus-specific (Koirala & Brözel, 2021), except for *nifV* (Arabidopsis-specific), *nifH* in the AtSC (Barley-specific), and *nodU* (not host-specific) (Figures S15, S16). The *nifV* gene encodes a homocitrate synthase required for nitrogen fixation. However, many nitrogen-fixing symbionts such as *Mesorhizobium* spp. lack this gene, and homocitrate is provided by the host legume via FEN1 (Hakoyama *et al.*, 2009). Enrichment of *nifV* in Arabidopsis roots was identified in this dataset indicating the presence of diazotrophic rhizobacteria in the microbiota of non-symbiotic plants, a phenomenon that was previously described (Santos *et al.*, 2017). Barley was found here to provide an environment promoting the selection of bacteria containing nitrogen-fixation genes, even if it does not have the natural capacity for this symbiosis. This was observed for Barley also when inoculated with AtSC which does not contain compatible symbionts for Lotus (Figure S15c). Together, these reiterate the well-established knowledge that bacteria with nitrogen-fixing symbiotic capacity have the capacity to occupy successfully the environments provided by different plant roots, including non-symbiotic hosts (Guo *et al.*, 2023; Haskett *et al.*, 2022; Pankiewicz *et al.*, 2021; Rosenblueth *et al.*, 2018).

One of the most enriched Lotus-specific gene clusters aside from the *nif* and *nod* genes are the *ery* genes that are involved in the transport and catabolism of the sugar erythritol (Figure S17, Data S26) (Geddes *et al.*, 2010). While erythritol is exuded by many plants (Carvalhais *et al.*, 2015; Vranova *et al.*, 2013), its exudation by Lotus nodules (Colebatch *et al.*, 2004; Ranner *et al.*, 2023) suggests that bacterial erythritol catabolism might be uniquely associated with Lotus nodules. This process could serve as another signal for symbionts to be enriched in the Lotus root microbiome.

Host immunity in bacterial function recruitment

Aside from the release of root exudates to attract rhizobacteria, plant hosts also shape the root microbiome through its immune system (Yu *et al.*, 2019). One way of doing this, is by releasing antibiotics or toxins into the rhizosphere that selectively target bacterial groups (Negre Rodríguez *et al.*, 2025). This cocktail of metabolites and antimicrobial compounds may differ between plant hosts and play a large role in host-specific signatures in root microbiome composition irrespective of the soil microbial community composition (Russ *et al.*, 2025). Host specificity in this study, however, mainly appears to arise from released metabolites as these are more prevalent among the host-specific functions observed within the microbiota (Data S26). A few host immunity modulating bacterial functions may suggest the existence of microbial host specificity related to immune processes. Arabidopsis-specific *ipdC*, for example, is involved in the conversion of the indole derivative tryptophan to auxin (Spaepen *et al.*, 2007), a possible plant-growth promoting trait, but may also play a role in the breakdown of other indole derivatives including the *Brassicaceae*-specific glucosinolates. Other examples include the Lotus-specific multidrug efflux pumps *mdlB*, *opmE*, *vexE*, and *vexF*, which may actively remove antimicrobial compounds from the bacterial cell to ensure resistance (Gaurav *et al.*, 2023). Interestingly, such efflux pumps have displayed host specific recruitment in other studies as well (Russ *et al.*, 2025).

References

- Carvalhais, L. C., Dennis, P. G., Badri, D. V., Kidd, B. N., Vivanco, J. M., & Schenk, P. M. (2015). Linking jasmonic acid signaling, root exudates, and rhizosphere microbiomes. *Molecular Plant-Microbe Interactions*, **28**, 1049–1058.
- Colebatch, G., Desbrosses, G., Ott, T., Krusell, L., Montanari, O., Kloska, S., Kopka, J., & Udvardi, M. K. (2004). Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *The Plant Journal*, **39**, 487–512.
- Gaurav, A., Bakht, P., Saini, M., Pandey, S., & Pathania, R. (2023). Role of bacterial efflux pumps in antibiotic resistance, virulence, and strategies to discover novel efflux pump inhibitors. *Microbiology*, **169**.
- Geddes, B. A., Pickering, B. S., Poysti, N. J., Collins, H., Yudistira, H., & Oresnik, I. J. (2010). A locus necessary for the transport and catabolism of erythritol in *Sinorhizobium meliloti*. *Microbiology*, **156**, 2970–2981.
- Göttfert, M. (1993). Regulation and function of rhizobial nodulation genes. *FEMS Microbiology Letters*, **104**, 39–63.
- Guo, K., Yang, J., Yu, N., Luo, L., & Wang, E. (2023). Biological nitrogen fixation in cereal crops: progress, strategies, and perspectives. *Plant Communications*, **4**, 100499.
- Haskett, T. L., Paramasivan, P., Mendes, M. D., Green, P., Geddes, B. A., Knights, H. E., Jorin, B., Ryu, M.-H., Brett, P., Voigt, C. A., Oldroyd, G. E. D., & Poole, P. S. (2022). Engineered plant control of associative nitrogen fixation. *Proceedings of the National Academy of Sciences*, **119**.
- Jeong, J., Kim, Y., Hu, Y., & Jung, S. (2022). Bacterial succinoglycans: structure, physical properties, and applications. *Polymers*, **14**, 276.
- Koirala, A., & Brözel, V. S. (2021). Phylogeny of nitrogenase structural and assembly components reveals new insights into the origin and distribution of nitrogen fixation across bacteria and archaea. *Microorganisms*, **9**, 1662.
- Malina, T., Koehorst, R., Bína, D., Pšenčík, J., & van Amerongen, H. (2021). Superradiance of bacteriochlorophyll c aggregates in chlorosomes of green photosynthetic bacteria. *Scientific Reports*, **11**, 8354.
- Mendis, H. C., Madzima, T. F., Queiroux, C., & Jones, K. M. (2016). Function of succinoglycan polysaccharide in *Sinorhizobium meliloti* host plant invasion depends on succinylation, not molecular weight. *MBio*, **7**.
- Negre Rodríguez, M., Pioppi, A., & Kovács, Á. T. (2025). The role of plant host genetics in shaping the composition and functionality of rhizosphere microbiomes. *MSystems*, **10**.
- Ouchane, S., Steunou, A.-S., Picaud, M., & Astier, C. (2004). Aerobic and anaerobic Mg-protoporphyrin monomethyl ester cyclases in purple bacteria. *Journal of Biological Chemistry*, **279**, 6385–6394.
- Pankiewicz, V. C. S., do Amaral, F. P., Ané, J.-M., & Stacey, G. (2021). Diazotrophic bacteria and their mechanisms to interact and benefit cereals. *Molecular Plant-Microbe Interactions*, **34**, 491–498.
- Pieterse, C. M. J., Berendsen, R. L., de Jonge, R., Stringlis, I. A., Van Dijken, A. J. H., Van Pelt, J. A., Van Wees, S. C. M., Yu, K., Zamioudis, C., & Bakker, P. A. H. M. (2021). *Pseudomonas simiae* WCS417: star track of a model beneficial rhizobacterium. *Plant and Soil*, **461**, 245–263.
- Ranner, J. L., Schalk, S., Martyniak, C., Parniske, M., Gutjahr, C., Stark, T. D., & Dawid, C. (2023). Primary and secondary metabolites in *Lotus japonicus*. *Journal of Agricultural and Food Chemistry*, **71**, 11277–11303.
- Rosenblueth, M., Ormeño-Orrillo, E., López-López, A., Rogel, M. A., Reyes-Hernández, B. J., Martínez-Romero, J. C., Reddy, P. M., & Martínez-Romero, E. (2018). Nitrogen fixation in cereals. *Frontiers in Microbiology*, **9**.
- Russ, D., Fitzpatrick, C. R., Saha, C., Law, T. F., Jones, C. D., Kliebenstein, D. J., & Dangl, J. L. (2025). An effluent pump family distributed across plant commensal bacteria conditions host- and organ-specific glucosinolate detoxification. *Nature Communications*, **16**, 5699.
- Santos, A. R. S., Etto, R. M., Furmam, R. W., Freitas, D. L. de, Santos, K. F. d'Eça N., Souza, E. M. de, Pedrosa, F. de O., Ayub, R. A., Steffens, M. B. R., & Galvão, C. W. (2017). Labeled *Azospirillum brasilense* wild type and excretion-ammonium strains in association with barley roots. *Plant Physiology and Biochemistry*, **118**, 422–426.
- Skorupska, A., Janczarek, M., Marczak, M., Mazur, A., & Król, J. (2006). Rhizobial exopolysaccharides: genetic control and symbiotic functions. *Microbial Cell Factories*, **5**, 7.
- Spaepen, S., Vanderleyden, J., & Remans, R. (2007). Indole-3-acetic acid in microbial and microorganism-plant signaling. *FEMS Microbiology Reviews*, **31**, 425–448.
- Spaink, H. P., Kondorosi, A., & Hooykaas, P. J. J. (2012). *The Rhizobiaceae: molecular biology of model plant-associated bacteria*. Springer Science & Business Media.
- Thiergart, T., Durán, P., Ellis, T., Vannier, N., Garrido-Oter, R., Kemen, E., Roux, F., Alonso-Blanco, C., Ågren, J., Schulze-Lefert, P., & Hacquard, S. (2019). Root microbiota assembly and adaptive differentiation among European *Arabidopsis* populations. *Nature Ecology & Evolution*, **4**, 122–131.
- Vranova, V., Rejsek, K., Skene, K. R., Janous, D., & Formanek, P. (2013). Methods of collection of plant root exudates in relation to plant metabolism and purpose: a review. *Journal of Plant Nutrition and Soil Science*, **176**, 175–199.

- Williams, E. A., Cobbina, S. M., & Bentil, J.** (2016). The effects of light intensity on soil depth of different moisture contents using laser sensor. *International Journal of Scientific and Research Publications*, **6**, 488–493.
- Wu, X., Ge, T., Yuan, H., Li, B., Zhu, H., Zhou, P., Sui, F., O'Donnell, A. G., & Wu, J.** (2014). Changes in bacterial CO₂ fixation with depth in agricultural soils. *Applied Microbiology and Biotechnology*, **98**, 2309–2319.
- York, G. M., & Walker, G. C.** (1998). The succinyl and acetyl modifications of succinoglycan influence susceptibility of succinoglycan to cleavage by the *Rhizobium meliloti* glycanases ExoK and ExsH. *Journal of Bacteriology*, **180**, 4184–4191.
- Yu, K., Pieterse, C. M. J., Bakker, P. A. H. M., & Berendsen, R. L.** (2019). Beneficial microbes going underground of root immunity. *Plant, Cell & Environment*, **42**, 2860–2870.

Lotus-specific *nif* genes

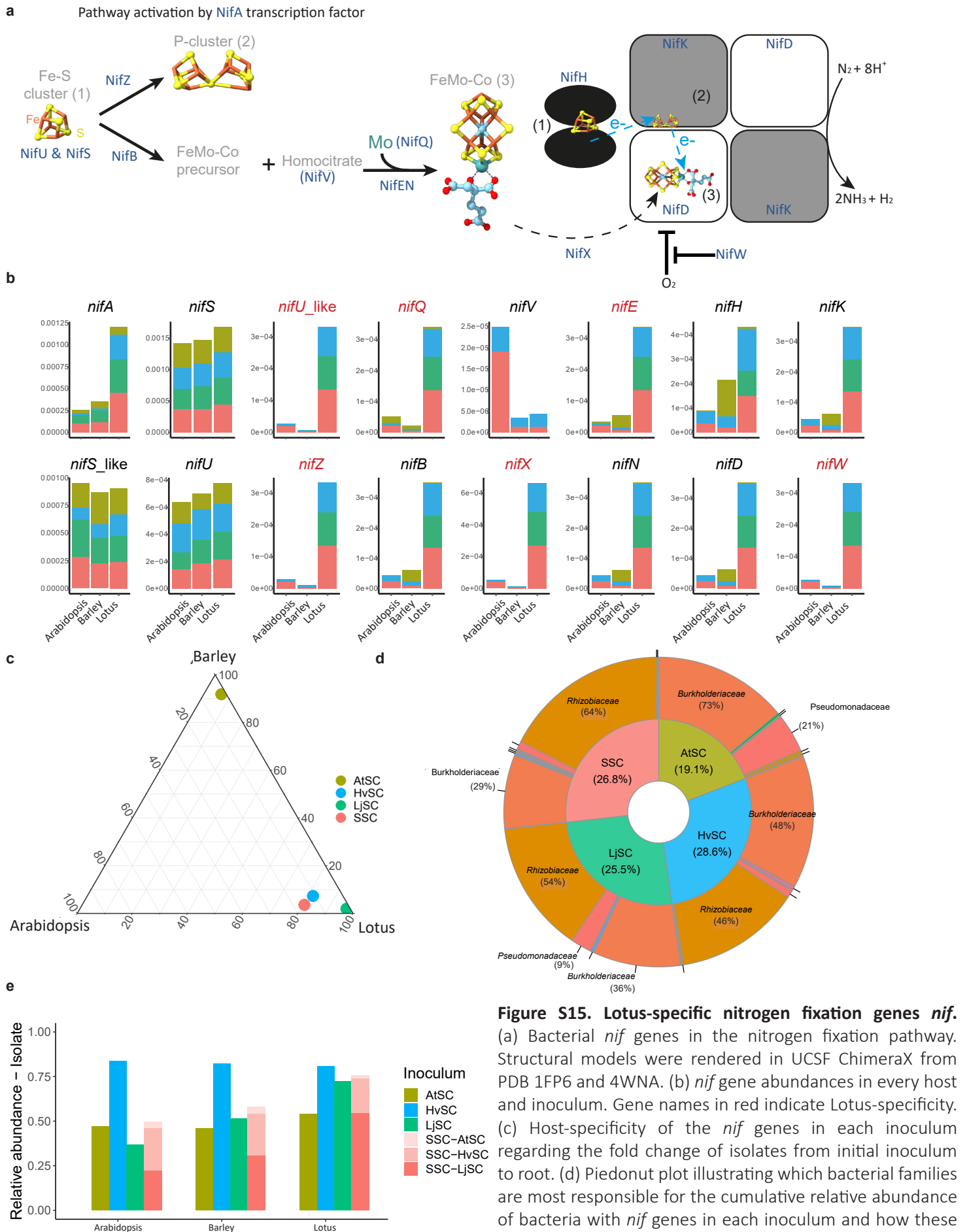


Figure S15. Lotus-specific nitrogen fixation genes *nif*.

(a) Bacterial *nif* genes in the nitrogen fixation pathway. Structural models were rendered in UCSF ChimeraX from PDB 1FP6 and 4WNA. (b) *nif* gene abundances in every host and inoculum. Gene names in red indicate Lotus-specificity. (c) Host-specificity of the *nif* genes in each inoculum regarding the fold change of isolates from initial inoculum to root. (d) Pie chart illustrating which bacterial families are most responsible for the cumulative relative abundance of bacteria with *nif* genes in each inoculum and how these cumulative relative abundances compare to each other across inocula. (e) Cumulative relative abundance of isolates in each host and inoculum that have *nif* genes. In the SSC it shows the origin of the isolates that contribute to this cumulative relative abundance.

Lotus-specific *nod* genes

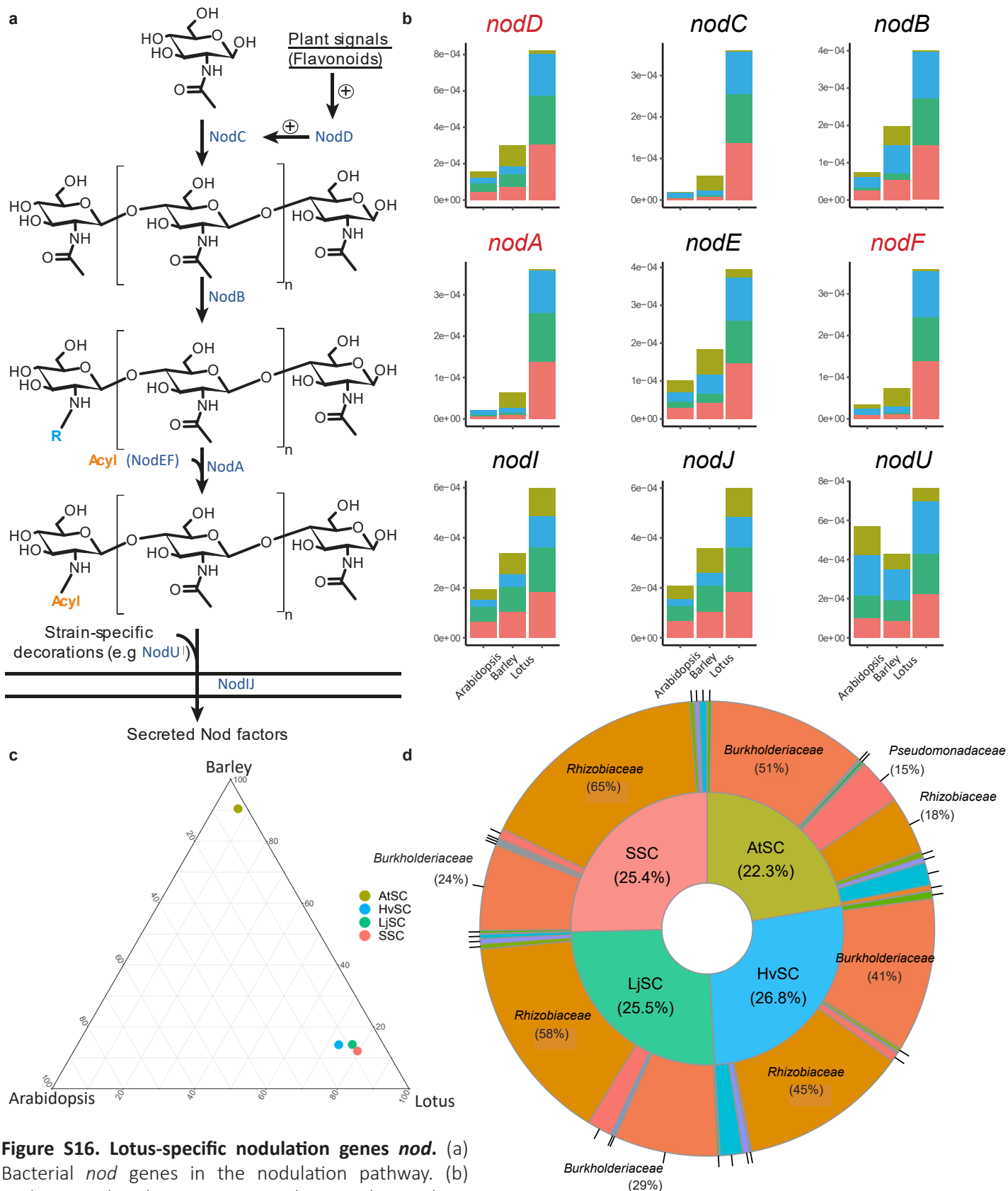


Figure S16. Lotus-specific nodulation genes *nod*. (a) Bacterial *nod* genes in the nodulation pathway. (b) *nod* gene abundances in every host and inoculum. Gene names in red indicate Lotus-specificity. (c) Host-specificity of the *nod* genes in each inoculum regarding the fold change of isolates from initial inoculum to root. (d) Piedonut plot illustrating which bacterial families are most responsible for the cumulative relative abundance of bacteria with *nod* genes in each inoculum and how these cumulative relative abundances compare to each other across inocula. (e) Cumulative relative abundance of isolates in each host and inoculum that have *nod* genes. In the SSC it shows the origin of the isolates that contribute to this cumulative relative abundance.

Lotus-specific *ery* genes

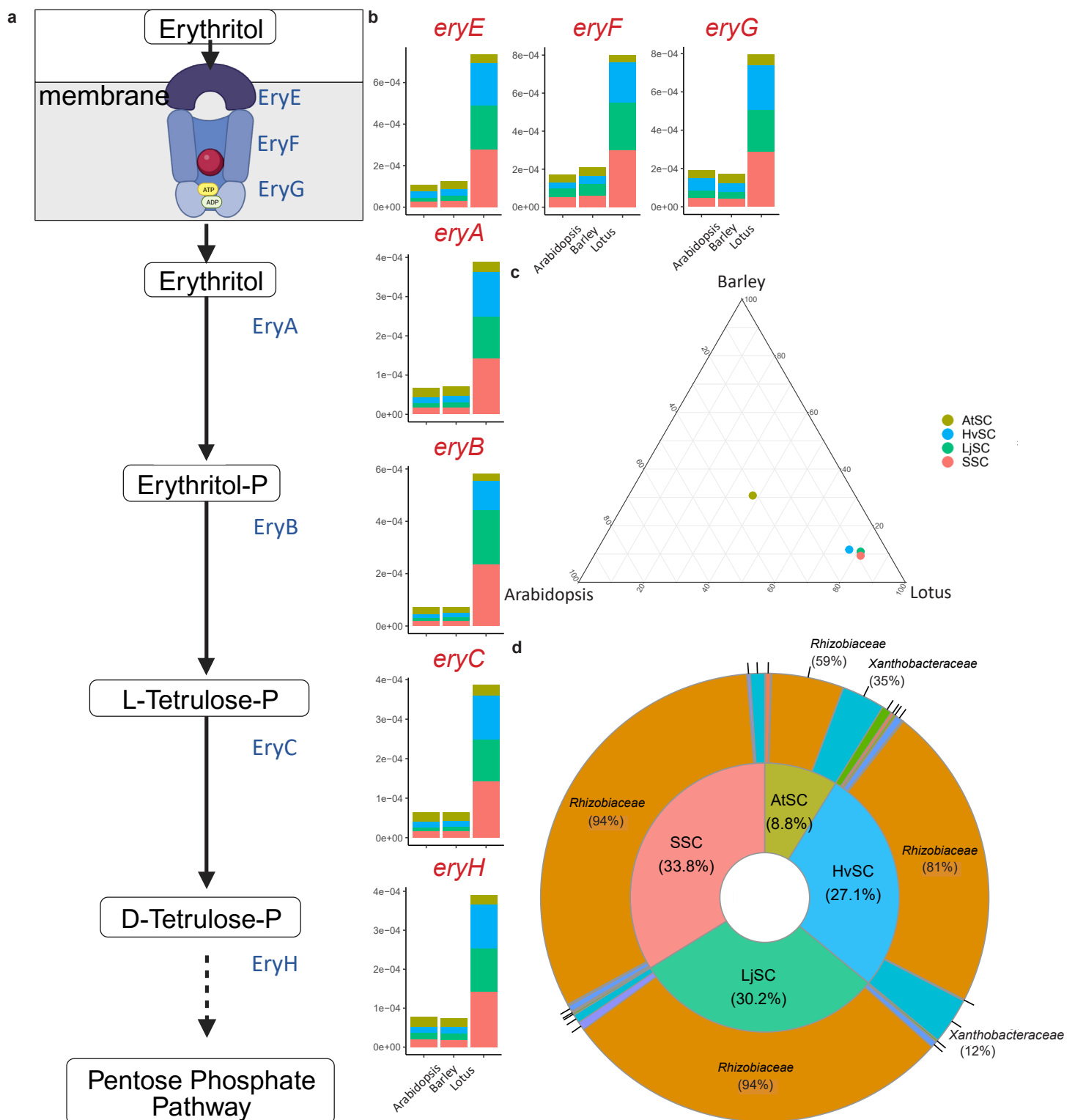


Figure S17. Lotus-specific genes *ery*. (a) Bacterial *ery* genes in the erythritol metabolism pathway. (b) *ery* gene abundances in every host and inoculum. Gene names in red indicate Lotus-specificity. (c) Host-specificity of the *ery* genes in each inoculum regarding the fold change of isolates from initial inoculum to root. (d) Piedonut plot illustrating which bacterial families are most responsible for the cumulative relative abundance of bacteria with *ery* genes in each inoculum and how these cumulative relative abundances compare to each other across inocula. (e) Cumulative relative abundance of isolates in each host and inoculum that have *ery* genes. In the SSC it shows the origin of the isolates that contribute to this cumulative relative abundance.

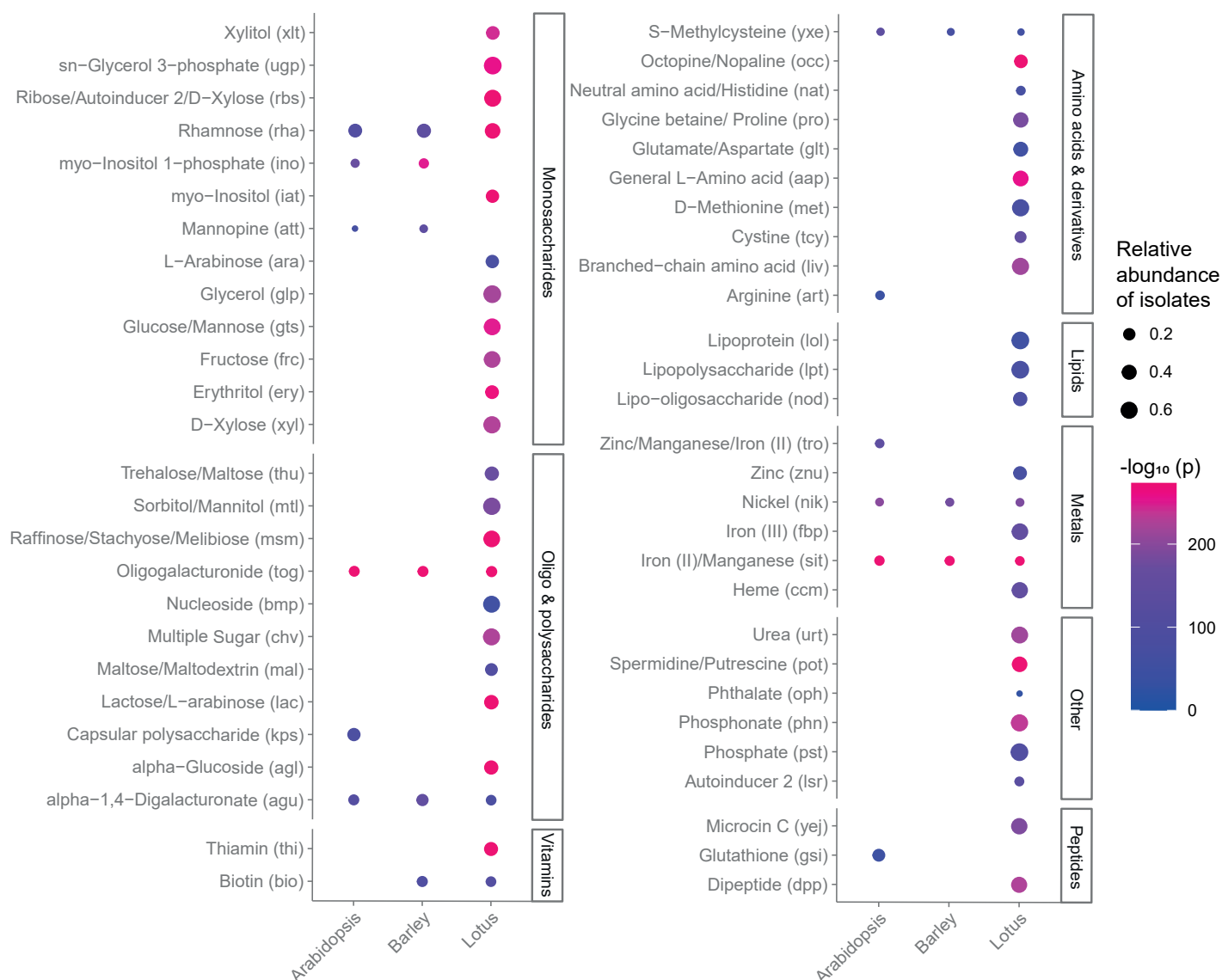
Supplemental results 3 – Host-specific functions - ABC transporters

The analysis of differential abundance of KOs between inocula and root communities revealed that ABC transporters emerged as the category with the largest number of host-specific bacterial functions (Figure 5d, Data S28). To better understand the role and contribution of ABC transporters in root microbiota assembly, we first investigated the number and diversity of ABC transporters in the three host-specific inocula (Extended Data 8). Analysis at the bacterial isolate level revealed that LjSC encompassed isolates with the largest number of ABC transporters, followed by HvSC and AtSC (Extended Data 8, solid lines). All three hosts enriched their root microbiota for isolates that encode a broad spectrum of ABC transporter illustrated by the many top colonizers encoding a relatively large number of ABC transporters. This included the so-called dominator strains, i.e., the Lotus symbionts and the *Rhizobacter* sp. P2_G4, indicating that ABC transporter diversity might contribute to their observed dominance (Extended Data 8). Root communities assembled by the three hosts in the presence of LjSC, however, were dissimilar, revealing little to no correlation—or even a negative correlation—between the number of ABC transporters and the relative abundance of isolates (Extended Data 8). We hypothesize that this could be a consequence of the higher overall number of ABC transporters in LjSC compared to the other SynComs, therewith raising the baseline and reducing the impact of additional transporters (Extended Data 8, Figure S18).

Illustratively, when investigating the number and variety of ABC transporters selected for by the three hosts, we find that Lotus shows the most frequent enrichment of a wide range of ABC transporters (Figure S18). Combined with the finding that LjSC isolates possess a higher diversity of ABC transporters, we hypothesize that the Lotus root microbiome is characterized by fluctuating metabolite availability, where the ability to metabolize a broad variety of compounds enhances microbial competitiveness. In Lotus, these bacteria could play a crucial role in the microbial community by fostering cooperation—such as the production and exchange of amino acids, vitamins, or other compounds—with other community members (Mataigne *et al.*, 2021).

References

Mataigne, V., Vannier, N., Vandenkoornhuyse, P., & Hacquard, S. (2021). Microbial systems ecology to understand cross-feeding in microbiomes. *Frontiers in Microbiology*, **12**.



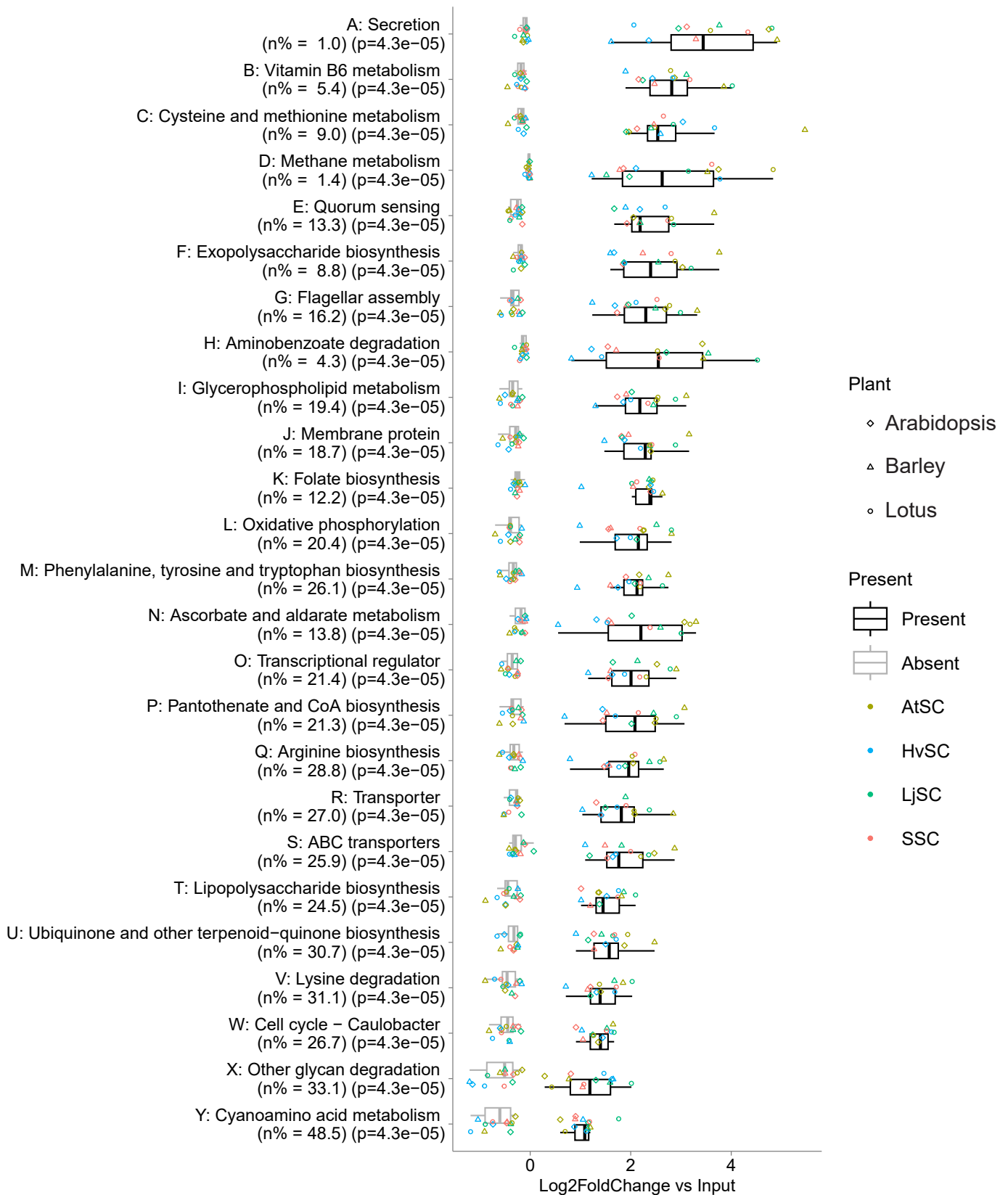
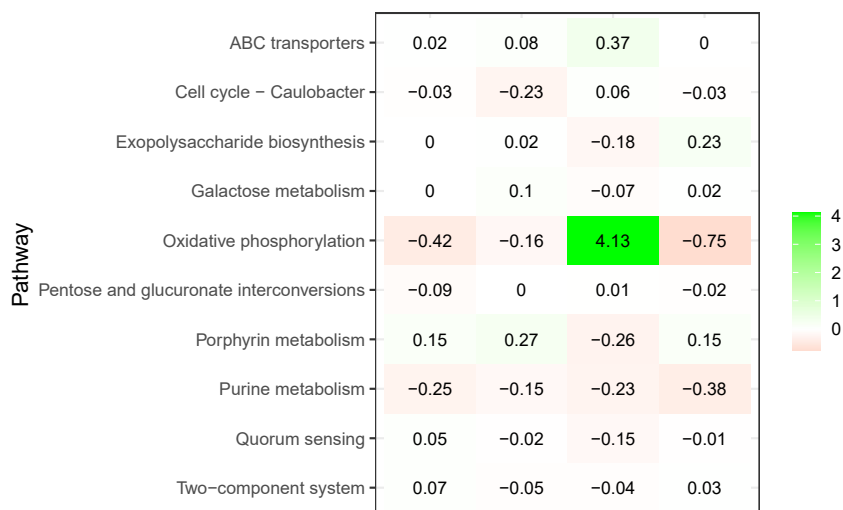


Figure S20. Pathway enrichment in the root microbiome. Fold change of bacterial strains with (Present) and without (Absent) significant pathway KOs from input inoculum to root microbiome. Only KOs that were significant from the DESeq2 analyses were collated to pathway level. Pathways are ordered based on the highest positive distance to an exponential decay function fitted to the relative abundance of strains with the function and the fold change (Figure S19b). Each point is a treatment ($n = 12$) consisting of 12 biological replicates. Boxes show the median and interquartile range; whiskers extend to the most extreme values within $1.5 \times$ the interquartile range. Individual jittered points represent biological samples, and values beyond the whiskers are shown individually. Colors and shapes indicate the inoculum and host respectively. $n\%$ indicates the percentage of strains with the pathway, and the p -value indicates statistical differences between the present and absent group based on a Wilcoxon test, with a p -value Holm adjustment.

a

Niche replacement score - plant-specific pathways



Niche replacement score - core pathways

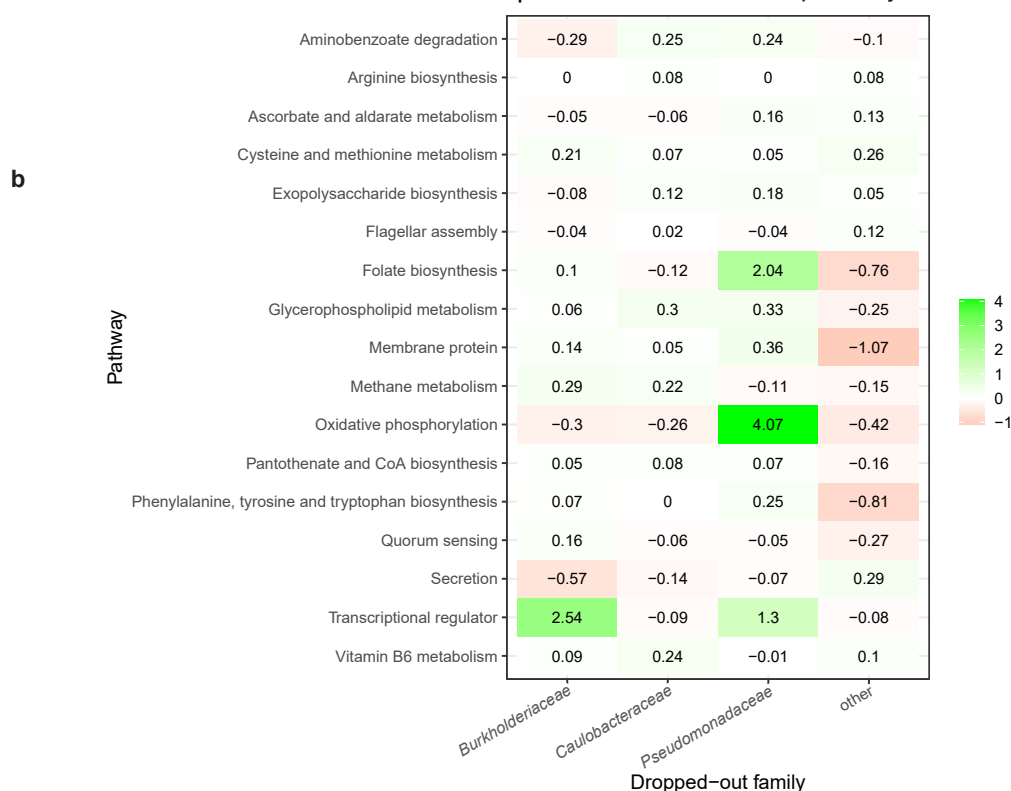


Figure S21. Niche replacement score for the plant-specific and core pathways. Niche replacement score, calculated as the log10-scaled ratio of pathway abundance fold change from input inoculum to root community between the dropped family SynCom and the full LjSC SynCom, indicates to what extent the loss of a bacterial family, and thus their KOs and corresponding pathways, in the root community are “compensated” for by other families. The colors correspond to the values in the cells. A ‘0’ indicates full compensation, while negative values indicate limited compensation of KOs upon family removal. Positive values on the other hand indicate that loss of a particular family results in increased abundance of the KO and corresponding pathway. The niche replacement score is shown for plant-specific KOs and corresponding pathways they belong to from Figure 5c-d (a) and the 266 core KOs and the pathways they belong to from Figure 5e-f (b). For each pathway, only the KOs that were significant in the differential abundance analyses were used for this calculation.

As can be seen from this plot, oxidative phosphorylation displays the highest positive *Pseudomonadaceae* niche-replacement score for both plant-specific and core pathways. Together this suggests that in the absence of *Pseudomonadaceae* bacteria, bacteria from other families fill this niche and do this to a great extent. In other family drop outs, this does not happen, suggesting that under balanced familial representation *Pseudomonadaceae* occupy this niche. Other noteworthy observations include the pathway purine metabolism in (a) as this pathway displays a negative score for each family drop out, suggesting equal contribution of each family to this pathway.

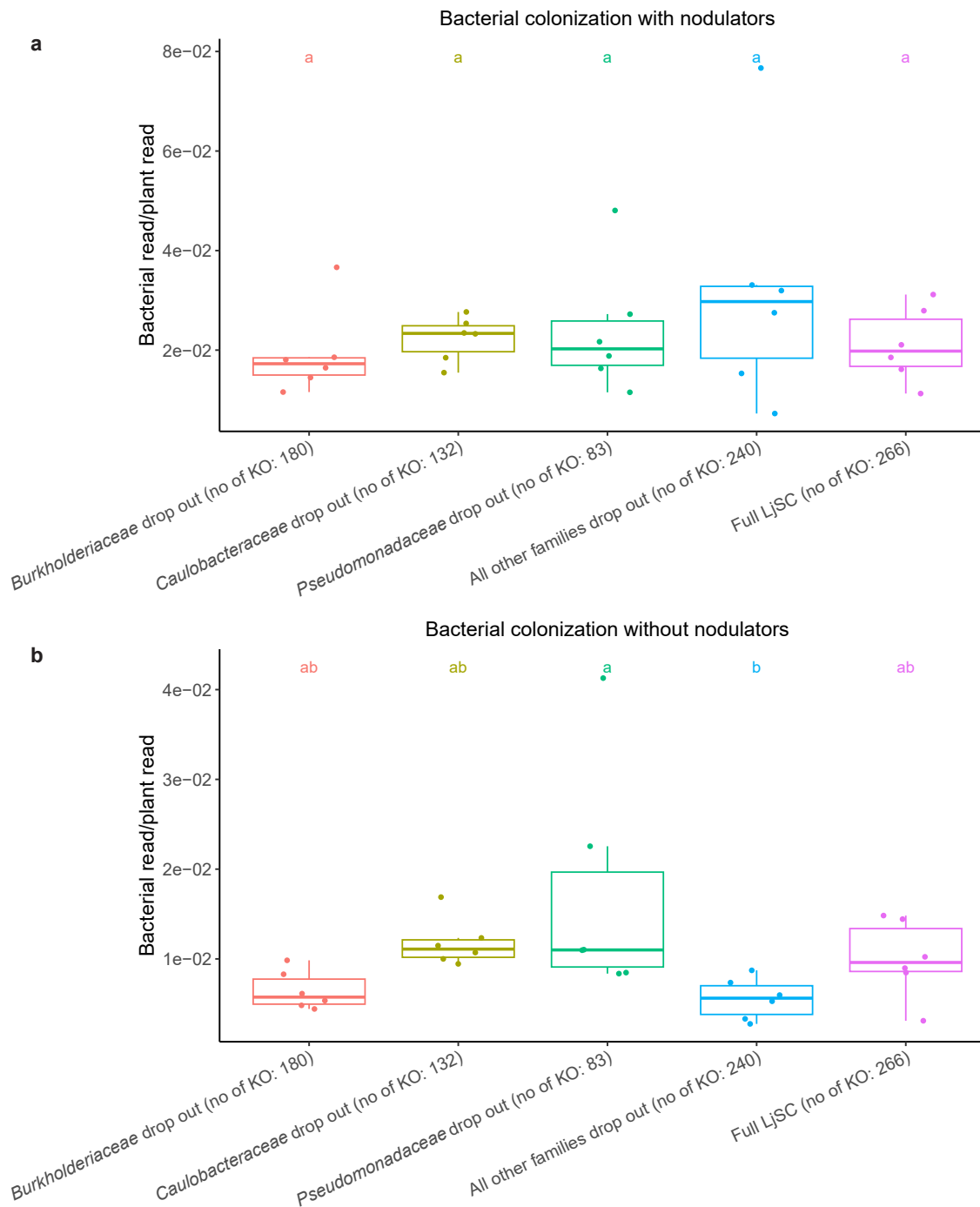


Figure S22. Bacterial load in LjSC family drop out experiment. For each SynCom in the LjSC family drop out experiment, the ratio between bacterial and plant reads was calculated providing information on the bacterial density or level of colonization on the root. Each point represents one sample and each group consists of $n = 6$ replicates. Boxes show the median and interquartile range; whiskers extend to the most extreme values within $1.5 \times$ the interquartile range. For each bacterial family the number of 266 core KOs it possesses is shown. Note that these KOs are not unique and shared across families (Extended Data 10) and merely a representation of their intrinsic capacities. The bacterial load numbers in the original dataset is shown (a) as well as in the dataset in which the nodulators were computationally removed (b). Statistical differences were assessed by an ANOVA followed by a Tukey post-hoc test with p -value adjustment ($p < 0.05$). Groups not sharing a letter are significantly different. Ratios and exact p -values are provided in Source Data S35.