Supplementary Information

Schlaeppi et *al.* , PNAS

Experimental design

We have investigated the root-inhabiting bacterial microbiota of *Arabidopsis thaliana* (L.) Heynh and the relative species *Arabidopsis lyrata* (L.) O'Kane & Al-Shehbaz, *Arabidopsis halleri* (L.) O'Kane & Al-Shehbaz and *Cardamine hirsuta* (L.). We performed two samplings at natural sites and conducted two replicate greenhouse experiments. The Table 1 provides an overview of replicate samples per sample type, plant species and experiments and the Dataset S1 provides a detailed experimental design with individual sample IDs and the sequencing effort.

Natural site experiments:

At the two sites 'Cologne' (50.982222034039 N/ 6.82718753814697 E, Widdersdorf, Germany) and 'Eifel' (50.45012819440579 N/ 6.936978399753571 E, Dümpelfeld, Germany) we have collected side by side naturally growing (i.e. not planted by the authors) *A. thaliana* and *C. hirsuta* in spring 2012. We have excavated whole plants including the surrounding soil in cores of ~5 cm in diameter and 5 - 10 cm in depth. The plants in their soil cores were brought to the laboratory and the root systems were sampled within 12 h after removing the plants from their natural habitat. From each species a minimum of 25 individuals were collected, of which 20 were pooled into 4 samples, each consisting of 5 plants. From these 4 samples per species the rhizosphere and the root compartments were fractionated (see below) and used for community profiling. The remaining 5 plants (per species and site) were re-potted and grown in the greenhouse to produce seeds in order to collect the genetic material from the natural sites. The soil collected from the sites was used to obtain 4 samples for community profiling and analyzed for physical and chemical properties at the 'Labor für Boden- und Umweltanalytik' (Eric Schweizer AG, Thun, Switzerland, Table S1).

Greenhouse experiments:

Seeds of *A. thaliana* ecotypes (Shakdara (Sha), Landsberg (Ler) and Columbia (Col, CS22625)) were received from Prof. Maarten Koornneef, Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany. Col was chosen as it is the most widely used ecotype of the model plant *A. thaliana* and compared with Ler and Sha for cross validation with our previous study (3). Dr. Pierre Saumitou-Laprade (Laboratoire de Génétique et Evolution des Populations Végétales, FRE CNRS 3268, Université de Lille, Villeneuve d'Ascq, France) kindly provided seeds of *A. halleri* (Auby), which he collected in summer 2009 at a heavy metal contaminated site nearby the town of Auby, France (1, 2). The line Mn47 of *A. lyrata* was obtained from the Nottingham Arabidopsis Stock Center (Stock ID N960898) and *C. hirsuta* (Oxford) was a kind gift of Prof. Miltos Tsiantis (Department of Comparative Development and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany). In replicate greenhouse experiments we grew the three *A. thaliana* ecotypes together with the relative species *A. halleri*, *A. lyrata* and *C. hirsuta* in pots containing natural, microbe-rich soil. The natural experimental 'Cologne soil' (CS) was collected at the Max Planck Institute for Plant Breeding Research (50.958 N/ 6.856 E, Cologne, Germany) in March and in September 2010, stored and prepared for use as previously described (3). The spring soil batch 'CS-4' and fall batch 'CS-5' were used in the first and in the second greenhouse experiments, respectively. The geochemical characterization, as obtained from the 'Labor für Boden- und Umweltanalytik' (Eric Schweizer AG, Thun, Switzerland) is provided in Table S1.

Before planting, seeds were surface-sterilized recycling spin columns from the Qiaquick gel extraction kit (Qiagen, Hilden, Germany), of which the silica membrane was removed. Seeds were incubated in 700 µl ethanol (70%) on a laboratory shaker in the closed spin columns and after 20 min the alcohol was removed and the seeds washed with a „flow-through" of 700 µl ethanol (100%) through the column. Following a short spin centrifugation step the seeds were dried in open spin columns for 30 minutes in a laminar flow hood. Spin columns were closed and seeds kept sterile until sowing. Surface sterilized seeds were sown onto 7x7x9 cm (LWH) plastic pots filled with experimental soil, which were placed at 4°C in the dark for stratification during 5 days prior to translocation to the greenhouse. We grew the plants for six weeks under short day conditions 8 hours light (day) and 16 hours dark (night), 22°C during the day and 18°C during the night at a relative humidity of 70 %. After germination, surplus seedlings were removed to grow the plants at a density of 4 plants per pot. Unplanted pots were subjected to the same conditions as the planted pots to prepare the control soil samples at harvest. We harvested the root systems after six weeks, when all plant species and genotypes were at the vegetative growth stage of the rosette. For all plant species and genotypes we prepared triplicate root samples (see below) for pyrosequencing each consisting of 12 plants originating from 3 pots. For comparison we prepared also 3 soil samples from 3 unplanted pots.

Sample preparation

Roots were separated from the adhering soil particles and the defined root segment of 3 cm length starting 0.5 cm below the root base was harvested. Roots were collected in 15 ml falcons containing 3 ml PBS-S buffer (130 mM NaCl, 7 mM Na_2HPO_4, 3 mM NaH_2PO_4, pH 7.0, 0.02 % Silwet L-77) and washed for 20 minutes at 180 rpm on a shaking platform. The roots were transferred to a new falcon tube and the soil suspension was centrifuged for 20 minutes at 4,000 x *g* and the pellet, referred to as the rhizosphere, collected in liquid nitrogen and stored at -80°C. After washing a second time (20 minutes at 180 rpm in 3 ml PBS-S buffer) the roots were transferred to a new falcon tube and sonicated for 10 minutes at 160 W in 10 intervals of 30 seconds pulse and 30 seconds pause (Bioruptor Next Gen UCD-300, diagenode, Liège, Belgium) to enrich for bacteria with root endophytic lifestyle. Roots were removed from PBS-S, dipped in a fresh volume of 3 ml PBS-S buffer and shortly dried on 50 mm diameter whatman filter paper (GE Healthcare USA), transferred to 2 ml tubes and frozen in liquid nitrogen for storage at -80°C. Soil samples were collected from unplanted pots in a soil depth of -0.5 to -3.5 from the surface corresponding to 3 cm root length, frozen in liquid nitrogen and stored at -80°C until further processing.

DNA was extracted with the FastDNA® SPIN Kit for Soil (MP Biomedicals, Solon, USA) following the manufacturer's instructions with minor modifications. Samples were homogenized in the Lysis Matrix E tubes using the Precellys®24 tissue lyzer (Bertin Technologies, Montigny-le-Bretonneux, France) at 6,200 rotations per second for 30 seconds. Frozen samples were homogenized 2 times without buffer and in between cooled in liquid nitrogen. Samples were homogenized a third time after the addition of the sodium phosphate and MT buffers provided by the kit. DNA samples were eluted in 100 µl DES water and DNA concentrations were determined using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen, life technologies) on an iQ™5 real-time PCR instrument (Bio-Rad Laboratories, Hercules, USA). Briefly, fluorescence was measured (at 25°C, 30'' conditioning followed by 3 cycles of 30'' for quantification) in 96 well plates filled with 40 µl of a 1:200 dilution of PicoGreen that was added to 4 µl of DNA sample or samples of a 5x dilution series (50 to 0.5 ng/µl) of lambda DNA (in the same plate). Based on the mean fluorescence of the 3 cycles, the DNA concentrations were calculated from the generated standard curve and adjusted to a final concentration of 3.5 ng/µl.

Amplicon libraries were generated using the PCR primers 799F (AACMGGATTAGATACCCKG, reference 14) and 1193R (5'-ACGTCATCCCCACCTTCC-3', reference 3) spanning ~400 bp of the hypervariable region

V5-V7 of the bacterial 16S rRNA gene. For multiplexed pyrosequencing we utilized the 799F primer fused at the 5‘ end with a sample specific (see Dataset S1), error-tolerant 6-mer barcode (N´s) followed by a *Sfi*I restriction site containing sequence required for the ligation of the 454 adapter A (see below; 5'-GATGGCCATTACGGCC-NNNNNN-799F-3'). The 1193R primer was extended at the 5' end to contain the target sequence of 454's sequencing primers (5'-CCTATCCCCTGTGTGCCTTGGCAGTCGACT-1193R-3').

PCRs were performed on an PTC-225 Tetrade DNA Engine (MJ Research, USA) with the DFS (DNA Free Sensitive) Taq DNA Polymerase system (Bioron, Ludwigshafen, Germany) using 3 µl of 3.5 ng/µl adjusted template DNA in a total volume of 25 µl. PCR components in final concentrations included 1 U DFS-Taq DNA Polymerase, 1x incomplete reaction buffer, 0.3% BSA (Sigma-Aldrich, St. Louis, USA), 2 mM of $MgCl_2$, 200 µM of dNTPs and 300 nM of each fusion primer. The PCR reactions were assembled in a laminar flow and amplified using the touch-down protocol in Table S2. To minimize stochastic PCR effects samples were amplified with 4 independently pipetted mastermixes in triplicate reactions per mastermix. Triplicate reactions of each sample were pooled per mastermix and a 5 µl aliquot inspected on a 1% agarose gel for the lack of PCR amplicons in non-template control reactions. Subsequently, pools of the replicate master mixes were sample-wise combined and cleaned from PCR ingredients using the QIAquick PCR Clean Up kit (Qiagen, Hilden, Germany), eluted in 30 µl of 10 mM Tris-HCl (pH 7.5) and loaded on a 1.5% agarose gel. The PCR primers 799F and 1193R produce a mitochondrial product at ~800 bp and a bacterial amplicon at ~450 bp, which we cut from the gel with the x-tracta Gel Extraction Tool (Sigma-Aldrich, St. Louis, USA) and extracted from the agarose using the QIAquick Gel Extraction kit (Qiagen, Hilden, Germany). Following purification and elution in 10 mM Tris-HCl (pH 7.5) we determined the concentration of the amplicon DNA in each sample using the PicoGreen assay described above. Finally we utilized 200 ng per barcoded DNA sample to build an amplicon library that was purified twice with the Agencourt AMPure XP PCR Purification system (1:1 ratio library/AMPure beads) to remove traces of PCR primers and primer dimers and thereby concentrating the final volume to 200 µl and further to 70 µl (in 10 mM Tris-HCl, pH 7.5). We prepared separate amplicon libraries for each natural site and each greenhouse experiments and we also prepared a combined library with a subset of samples from both greenhouse replicates (see Dataset S1).

Amplicon libraries containing a *Sfi*I restriction enzyme site at the 5' end were prepared for ligation of the 454 adaptor A by digestion for 1 h at 50°C with *Sfi*I (NEB, Frankfurt, Germany) and afterwards purified with the MinElute PCR Purification Kit (Qiagen, Hilden, Germany). The 454 adaptor A with its compatible 3'-*Sfi*I-overhang was ligated to the amplicons overnight at 16°C with 1 U T4 DNA ligase (Roche, Mannheim, Germany) followed by heat inactivation (10 min at 65°C). 454 compatible amplicon libraries were purified from unligated adapters after size fractionation on 2% agarose gels with the Qiaquick gel extraction kit (Qiagen, Hilden, Germany). Amplicon libraries were bound to beads and clonally amplified using the GS FLX Titanium LV emPCR Kit (Lib-L). The amplicon libraries were then sequenced using the GS FLX Titanium Sequencing Kit XLR70 and GS FLX Titanium PicoTiterPlate Kit. All kits used were purchased from Roche and used according to the manufacturers' protocol. Sequencing was performed at the Max Planck Genome Center in Cologne (http://mpgc.mpipz.mpg.de/home/).

Sequence analysis using QIIME

Pyrosequencing reads were processed and analyzed using QIIME (4, version 1.7.0). Using the script *split_libraries.py* we splitted the reads of each of the libraries according to their assigned barcodes to the individual samples (Dataset S1). Reads with erroneous barcode or forward primer sequences or with ambiguous base calls were discarded. We defined quality sequences to require a minimal Phred score of 27 and to be at least 315 bp long. With the same script we truncated the reads to remove the reverse primer and any subsequent sequence at the 3' end. We then concatenated all quality sequences that were indexed by samples and libraries into a single fasta file, which we used as input for *pick_otus.py*. Using uclust (5) we clustered the quality sequences at 97 % sequence similarity defining the operational taxonomic units (OTUs). Chimeric OTUs/sequences were identified by ChimeraSlayer (6) using default settings and removed from the analysis. The most abundant sequence in an OTU cluster was selected as OTU representative sequence, taxonomically assigned with Greengenes (7; release gg_otus_13_05, confidence cutoff 0.5) and bound in an OTU table with *make_otu_table.py*. We subsequently identified plant-sequence-derived OTUs with a custom *R* script and removed these from the OTU table with the script *filter_otus_from_otu_table.py*. The resulting OTU table (Dataset S8) contained the quality sequences of all samples of the Cologne (L388) and Eifel (L39) site, Greenhouse replicate 1 (L28) and replicate 2 (L35) experiments as well as the common library (L40). The L40 samples, utilized to examine biological vs. technical variation (see below), were removed for downstream analyses with the script *filter_samples_from_otu_table.py*. For the phylogeny related analyses we build a tree based on PyNAST-aligned (8) and filtered set of representative sequences. The script *multiple_rarefactions.py* with the OTU table (Dataset S8) as input was used to prepare the rarefied OTU tables (100x tables from 1,000 - 6,000 sequences per sample, steps of 1000 sequences), which were used for alpha diversity analyses (*alpha_diversity.py* with metrics *observed_species* and *PD_whole_tree*). The alpha diversity data was imported into R (R Development Core Team, http://www.R-project.org) to plot Fig. S7. We randomly chose the data file #31 (Dataset S4) of the rarefied OTU tables (sampling depth of 6,000 sequences per sample) for the downstream OTU-based and beta diversity analyses. We refer to this data matrix as the threshold-independent community (TIC) for which the taxonomic overview is presented in Fig. S4*B*. We estimated beta diversity by calculating weighted UniFrac distances with the script *beta_diversity.py* and imported the distance matrix into R to generate the Fig. S5.

Technical reproducibility of community profiles

We determined the technical reproducibility of 16S rRNA gene amplicon libraries by pyrosequencing Library L28 (greenhouse replicate #1) from parallel emulsion PCRs (emPCR). The reaction products of the replicate emPCRs were sequenced on separate regions of the same 454 pico titer plate resulting in 368,675 and 416,352 raw reads from regions 1 and 2, respectively. For this analysis we performed a second QIIME run with data from the L28 samples only but split by region, whereas the sequences of the L28 samples from both regions had been combined for the main QIIME analysis (see above). We utilized the same QIIME pipeline as described above to generate an OTU table based on co-clustered quality sequences (Dataset S2). In R using the package Vegan (9), we corrected for differences in sequencing depths by rarefaction to 6,000 sequences per sample. Fig. S1*A* displays of the 3 *A. thaliana* (ecotype Col), 3 *C. hirsuta* and the 3 soil samples the pairwise variation in OTU abundance from the parallel sequencing results. To define the minimal number of sequences per OTU required for the reproducible quantification of OTU abundance, we progressively removed low abundant OTUs from the data matrix. We tested Spearman rank correlation of bacterial profiles between corresponding samples in datasets where individual OTUs were represented by a minimum of 1, 2, 3 or up to 40 sequences in one of the two samples (Fig. S1*B*).

Rarefaction analysis (Fig. S2)

Combining of sequences from both pico titer plate regions of the first greenhouse replicate experiment (Library L28) resulted in sequencing depths of 17,441 – 58,150 quality sequences per sample (Dataset S1) and permitted rarefaction analysis at augmented sequencing depth (see below, Fig. S2). Based on the non-rarefied dataset (Dataset S8) derived from the main QIIME analysis (see above), we estimated rarefaction curves for each sample individually using the function *calculateRarefaction* of the R package ShotgunFunctionalizeR (10). Similarly, we performed rarefaction analysis on the ACM dataset (see below).

Defining the Abundant Community Members (ACMs)

Based on the TIC datafile (Dataset S4) we prepared the abundant community members (ACMs) data matrix (Dataset S3) by removing OTUs, which did not reach the minimum of 20 quality sequences in at least one of the 77 samples of the natural site and greenhouse experiments. We normalized the counts of individual ACM OTUs in a sample by dividing the total counts of all ACM OTUs within that sample followed by a multiplication by 1,000 resulting in relative abundance (RA) expressed as per mill. Statistical comparisons were conducted on log2-transformed (log2(RA+1)) per mill values. Fig. S3 reports the number of quality sequences per sample in the ACM data matrix and the Fig. S4*A* displays the taxonomic composition of the ACM.

Alpha and beta diversity analyses for the ACM were performed in QIIME using the same functions and parameters as for the TIC analyses described above. To this end, the OTU-IDs of the ACM that were determined in R were used in QIIME to subset the TIC datafile (Dataset S4) to the ACM data matrix (Dataset S3). The ACM data matrix was used for multiple rarefactions (100x tables from 500 - 6,000 sequences per sample, steps of 500 sequences) to prepare the rarefied ACM containing OTU tables, which were subsequently employed for the ACM alpha diversity analyses and this data was also imported into R to generate the Fig. S7. The ACM beta diversity estimates are based on 1,400 sequences per sample (see Fig. S3) and distance matrix was imported into R to generate the Fig. 2.

Technical reproducibility of library sequencing

The greenhouse experiments were conducted with the two seasonal soil batches 'spring CS-4' and 'fall CS-5' in the first and in the second experiments, respectively. Since we noted that the environmental conditions (soil types at the natural sites and the soil batches under the controlled conditions) were the major sources of variation in ACM community composition (Fig. 2 and Fig. S5), we examined if this could arise from independent library preparation and sequencing. We therefore prepared the additional common sequencing library (L40) containing bacterial amplicons of the 3 *A. thaliana* root (ecotype Col) and 3 soil samples of each replicate greenhouse experiment. The PCR reactions with these 12 samples were conducted in parallel and we generated 531 – 1,225 quality sequences per sample for this control library (Dataset S1). To analyze the community profiles of the common sequencing library L40 with the original sequences of the samples from the libraries L28 and L35 we subsetted the non-rarefied OTU table (Dataset S8) for these 3 root and 3 soil samples of each library with the script *filter_samples_from_otu_table.py*. We corrected for differences in sequencing depths by rarefaction to 530 sequences per sample (Dataset S5) and calculated the weighted UniFrac distances using QIIME. Subsequently, the data was imported into R to prepare the Fig. S6.

Statistical analysis using ANOVA

We implemented ANOVA-based statistics to identify taxonomic groups of OTUs ('community modules') and individual OTUs ('community members') that differ quantitatively between samples (root vs. soil communities; among root communities of different plant species). For community module analyses we prepared abundance matrices both at phylum and family rank containing sample-wise the sum of OTU abundances of all OTUs in the ACM per given taxon. For example, the abundance of the phylum Bacteroidetes resulted from the summed abundances of all OTUs in the ACM assigned to this phylum. The data matrices at phylum and family rank comprised 9 and 51 taxa, respectively. For individual community member statistics, we investigated the 237 OTUs of the ACM. All statistical comparisons were performed with custom scripts in *R* on log2-transformed values (see above). For the analysis of both, natural site and greenhouse experiments, we used two models: one for comparisons of root, soil and rhizosphere samples with one factor depicting the sample groups, and the other for comparisons among root samples with all factors and their interactions. The models are described below and given in the supplementary Dataset S6.

Natural site experiments

We first searched for differentially abundant 'community modules' among the root samples of both sites as a function of the variables *site* and *host species*. The abundance of each community module was modeled for the variables *site* (levels: Eifel and Cologne) and *host species* (levels: *A. thaliana*, *C. hirsuta*): ANOVA(Abundance_of_taxon ~ *site* * *host species*, data=root_samples_both_sites). We corrected the *P* values of these F-tests for both variables and their interaction (*site:species*) for the number of tests performed using the Benjamini and Hochberg (BH) method (11). The ANOVA statistic results are presented in the Dataset S6 worksheet A. We subsequently conducted pair-wise comparisons between the sample groups of the *site:species* interaction using Tukey's HSD (honestly significant difference) post-hoc test. Sample groups included the *A. thaliana* and *C. hirsuta* root samples of the Cologne and Eifel sites: A.t.Cologne, C.h.Cologne, A.t.Eifel and C.h.Eifel. Of the Tukey results we extracted for each taxon the adjusted *P* values (Tukey corrects for multiple hypothesis testing resulting from the pair-wise comparisons between the sample groups) and we further corrected the Tukey's *P* values for the number of taxa tested using the BH method. The Tukey statistic table of this analysis is presented in Dataset S6 worksheet B. From the comparison terms 'A.t.Cologne-C.h.Cologne' and 'A.t.Eifel-C.h.Eifel' we deduced the species-specific community modules of each site as reported in Fig. S8.

Secondly, we examined the bacterial communities at the level of individual members (OTUs) in the soil, root and rhizosphere samples. We determined for both sites separately the OTUs that are enriched in the roots of each species compared to the corresponding soil community (designated 'RootOTUs'). Analogous, we identified the OTUs that are more abundant in the rhizosphere of a species compared to the corresponding soil (termed 'RhizoOTUs'). To this end the abundance of each OTU of the ACM was modeled as a function of the 'sample groups' present in the experimental design of each natural site experiment: TukeyHSD(ANOVA(Abundance_of_OTU ~ *sample_group*, data=by_site)). Sample groups comprised soil, rhizosphere samples of *A. thaliana* (A.t.Rhizosphere), root samples of *A. thaliana* (A.t.Root), rhizosphere samples of *C. hirsuta* (C.h.Rhizosphere) and root samples of *C. hirsuta* (C.h.Root). We conducted pair-wise comparisons between the *sample groups* using the Tukey method described above. For each OTU we extracted the Tukey adjusted *P* values and further corrected these for the number of OTUs tested using the BH method (Dataset S6 worksheets C and D). From the comparison terms 'A.t.Rhizosphere-Soil' and 'C.h.Rhizosphere-Soil' we deduced the RhizoOTUs for *A. thaliana* (*A.t.*RhizoOTUs) and *C. hirsuta* (*C.h.*RhizoOTUs), respectively. The Fig. S12 displays the RhizoOTUs as identified for each species at both sites. From the comparison terms 'A.t.Root-Soil' and 'C.h.Root-Soil' of the same Tukey test we determined the root-enriched OTUs for *A. thaliana* (*A.t.*RootOTUs) and *C. hirsuta* (*C.h.*RootOTUs), respectively. The Fig. S9 displays the species-specific RootOTUs at both sites. We then defined the RootOTUs for each site (reported in Figures 3*A,* 3*B* and S10A) with the union of *A.t.*RootOTUs and *C.h.*RootOTUs of a site. Finally, we compared the RootOTUs of the Cologne site with the RootOTUs of the Eifel site, of which we derived from the union and the intersection the 70 RootOTUs (Fig. S10*A*) of both natural sites and the shared 19 RootOTUs at the natural sites (Fig. S10*A*).

Thirdly, we examined the variation of the RootOTU community among the root samples of both natural sites as a function of the variables *site* and *host species*. The abundance of each of the 70 RootOTUs was modeled for the variables *site* (levels: Eifel and Cologne) and *host species* (levels: *A. thaliana*, *C. hirsuta*): ANOVA(Abundance_of_RootOTU ~ site * host species, data=root_samples_both_sites). The *P* values were calculated in the same way as for the community module analysis described above (Dataset S6 worksheet E). We subsequently conducted pair-wise comparisons between the sample groups of the *site:species* interaction using the Tukey method described above. The Tukey statistic table is presented in Dataset S6 worksheet F. From the comparison terms 'A.t.Cologne-C.h.Cologne' and 'A.t.Eifel-C.h.Eifel' we deduced the species-specific community members of each site as reported in Fig. S11.

Greenhouse experiments

The ANOVA analysis of the replicate greenhouse experiments followed the same overall strategy as for the natural site experiments. First, community modules were searched in the abundance matrices for the phylum (8 taxa, the division AD3 was not detected in the greenhouse experiments) and family (50 taxa, the family Pelobacteraceae was not detected in the greenhouse experiments) among the root samples of both replicate experiments for the variable *host species.* The abundance of each taxon was modeled for the variables *replicate* (levels: replicate 1, replicate 2) and *host species* (levels: *A. thaliana*, *A. halleri, A. lyrata* and *C. hirsuta*): TukeyHSD(ANOVA(Abundance_of_taxon ~ *replicate* * *host species*, data=root_samples_both_replicates)). We directly conducted pair-wise comparisons for the variable *host species* using the same Tukey and *P*-value correction method described for the community module analysis of the natural site experiments (Dataset S6 worksheet G). From the comparison terms with all other plant species (e.g. for *A. thaliana*: *A. thaliana-A. halleri, A. thaliana-A. lyrata and A. thaliana-C. hirsuta*) we determined the species-specific community modules (Fig. S13).

Analog to the natural site analysis we then compared the bacterial communities at the level of individual members (OTUs) between soil and root samples. We calculated in each replicate experiment separately for each species the root-enriched OTUs (RootOTUs) from the comparison with the respective soil samples. The abundance of each OTU of the ACM was modeled for the 'sample groups' present in the greenhouse experiments: ANOVA(Abundance_of_OTU ~ *sample_group*, data=by_replicate). Sample groups comprised soil samples and root samples of *A. halleri*, *A. lyrata*, *A. thaliana* and *C. hirsuta*. We performed directly pair-wise comparisons between the *sample groups* using the same Tukey method as described above for the identification of the RootOTUs at the natural sites. From the comparison terms '*A. halleri*-Soil', '*A. lyrata*-Soil', '*A. thaliana*-Soil' and '*C. hirsuta*-Soil' we identified the RootOTUs for *A. halleri* (*A.h.*RootOTUs), *A. lyrata* (*A.l.*RootOTUs), *A. thaliana* (*A.t.*RootOTUs) and *C. hirsuta* (*C.h.*RootOTUs, Dataset S6 worksheets H and I). From the union of the *A.h.*RootOTUs, *A.l.*RootOTUs, *A.t.*RootOTUs and *C.h.*RootOTUs of a replicate experiment we then defined the RootOTUs for each replicate experiment (reported in Fig. S19A). Of these we derived from the union the 76 RootOTUs (Figs. 4*A,* S14) and from their intersection the shared 39 RootOTUs of the greenhouse experiments (Fig. S19A).

We then also investigated the variation in the RootOTU community among the root samples of both replicate experiments for the variable *host species.* The abundance of each of the 76 RootOTUs was modeled for the variables *replicate* (levels: replicate 1, replicate 2), *host species* (levels: *A. thaliana*, *A. halleri, A. lyrata* and *C. hirsuta*) and *genotype* (levels: Col, Ler and Sha for the species *A. thaliana*): Tukey(ANOVA(Abundance_of_RootOTU ~ *replicate* * *host species* * *genotype*, data=root_samples_both_replicates)). We performed directly pair-wise comparisons between the *host species* using the same Tukey and *P*-value correction method described for the community module analysis of the greenhouse experiments (Dataset S6 worksheet J). In the same way we also determined the species-specific community members from the comparison terms of a species with all other plant species (e.g. for *A. thaliana*: *A. thaliana-A. halleri, A. thaliana-A. lyrata and A. thaliana-C. hirsuta;* Fig. S16).

Finally, we assessed variation in root microbiota composition between and within host species by comparing the 3 *A. thaliana* ecotypes with the 3 Arabidopsis sister species. We excluded the *C. hirsuta* samples for this direct comparison to have a balanced design of 3 levels for each of the variables *host species* and *host genotype*. We modeled the abundance of each of the ACM OTUs for the variables *replicate* (levels: replicate 1, replicate 2), *host species* (levels: *A. thaliana*, *A. halleri* and *A. lyrata*) and *genotype* (levels: Col, Ler and Sha for the species *A. thaliana*): ANOVA(Abundance_of_OTU ~ *replicate* * *host species* * *genotype*, data=Arabidopsis_root_samples_both_replicates). We determined the effect sizes of *host species* and *host genotype* from *P* values for each of variables, which were corrected for the number of tests performed using the BH method (Fig. S18 and Dataset S6 worksheet K).

Validation of ANOVA assumptions:

We have examined the suitability of the ANOVA framework for dissecting taxa/OTU tables by testing the normality of data dispersion using the Shapiro-Wilk test. For each of the data subsets (see above) we tested the distribution of data points for each taxon/OTU. The Table S3 reports the test statistics of all data subsets and the number of taxa/OTUs in the respective sub-analysis that are normally distributed. We noted that between 26 and 67 % of the taxa/OTUs have normally distributed data points and we are aware that not all taxa/OTUs meet the formal requirements of an ANOVA analysis. Therefore, we have validated the ANOVA findings with non-parametric Mann-Whitney tests and a Bayesian statistic method.

Statistical analysis using non-parametric Mann-Whitney tests

We examined the identification of RootOTUs as performed with the ANOVA approach based on the same statistical comparisons but using non-parametric Mann-Whitney tests. We tested the 237 individual OTUs of the ACM (‘community members’) for quantitative differences in abundance between soil and root communities. Analyses were performed with custom *R* scripts using the function *wilcox_test* of the library *coin*.

Natural site experiments

We conducted pair-wise comparisons between each group of root samples (*A. thaliana* root samples of the Eifel site = *A.t.*Root_Eifel; *C.h.*Root_Eifel, *A.t.*Root_Cologne and *C.h.*Root_Cologne) and their corresponding group of soil samples (Soil_Eifel, Soil_Cologne). We determined the root-enriched OTUs for *A. thaliana* from the Eifel site (*A.t.*RootOTUs_Eifel) from the comparison term ‘*A.t.*Root_Eifel-Soil_Eifel’ based on *P* values < 0.1 that were corrected for the number of OTUs tested using the BH method.

Analogous we defined the *C.h.*RootOTUs_Eifel, *A.t.*RootOTUs_Cologne and the *C.h.*RootOTUs_Cologne (Dataset S6 worksheet L). We then defined the RootOTUs for each site with the union of the corresponding *A.t.*RootOTUs and *C.h.*RootOTUs and finally, comparing the RootOTUs of the Cologne site with the RootOTUs of the Eifel site, we derived from the intersection the shared 34 RootOTUs at both natural sites (Fig. S10*B*).

Greenhouse experiments

Analogous to the natural site analysis we also compared the bacterial communities of soil samples and root samples from the replicate greenhouse experiments. We calculated in each replicate experiment separately for each species the root-enriched OTUs (RootOTUs) from the comparison with the respective soil samples using pair-wise comparisons. From the comparison terms '*A. halleri*-Soil', '*A. lyrata*-Soil', '*A. thaliana*-Soil' and '*C. hirsuta*-Soil' we identified the RootOTUs for *A. halleri* (*A.h.*RootOTUs), *A. lyrata* (*A.l.*RootOTUs), *A. thaliana* (*A.t.*RootOTUs) and *C. hirsuta* (*C.h.*RootOTUs, Dataset S6 worksheets M and N). From the union of the *A.h.*RootOTUs, *A.l.*RootOTUs, *A.t.*RootOTUs and *C.h.*RootOTUs of a replicate experiment we defined the RootOTUs for each replicate experiment (reported in Fig. S19A). Of these we derived from their intersection the shared 62 RootOTUs of the greenhouse experiments (Fig. S19*B*).

Statistical analysis using a Bayesian approach

A more direct approach to find OTUs enriched in the roots in each of the tested species compared to soil is to test a single hypothesis instead of the intersection of multiple hypotheses via a Venn-diagram. One such approach, BayesianIUT, has been implemented by van Deun et al 2009 with the aim to find genes higher (or lower) expressed in one tissue compared to many other tissues (12). In our setting this approach calculates the support for two hypotheses: (1) that an OTU has lower abundance in the soil samples compared to root samples of each species and (2) that the OTU has at least in one of the species an equal or lower abundance compared to soil. The ratio of the support for each of the two hypotheses (Bayes factor) is calculated using a Bayesian approach in an ANOVA framework (12, http://ppw.kuleuven.be/okp/software/bayesianiut/). An OTU is root-enriched across all species if there is 30 times more support for the alternative hypothesis (12). We calculated the support for each hypothesis for both natural site experiments (Fig. S10*B*) and both greenhouse experiments (Fig. S19*B*). For natural site experiments we compared the root sample groups

*A.t.*Eifel, *C.h.*Eifel, *A.t.*Cologne and *C.h.*Cologne to the group of soil samples (both sites combined). We opposed the group of soil samples (both replicates combined) to all root samples as groups by species and replicate (*A.h.*rep1, *A.h.*rep2, *A.l.* rep1, *A.l.*rep2, *A.t.*rep1, *A.t.*rep2, *C.h.*rep1 and *C.h.*rep2).

Defining shared and core RootOTUs

We defined the 'shared RootOTUs' both at natural sites (Fig. S10*B*) and in the greenhouse experiments (Fig. S19*B*) when they were supported by parametric Tukey (ANOVA), non-parametric Mann-Whitney and Bayesian statistics (see below). The intersection of the three methods revealed 14 and 26 shared RootOTUs for the natural site and the greenhouse experiments, respectively. We finally compared the shared RootOTUs of the natural sites with the shared RootOTUs of the greenhouse experiments and defined from their intersection the 9 core RootOTUs (Fig. 5*A*).

Canonical analysis of principal coordinates (CAP)

To assess the influence of the different environmental and experimental factors on the beta diversity we calculated Bray-Curtis distances and then performed a Canonical Analysis of Principal coordinates (CAP) (13) constrained by the factor of interest and conditioning by the remaining variables. We employed R package vegan v2.0-8 (9) for the constrained ordination ('capscale' function for CAP analysis) as well as for the calculation of the significance values and confidence intervals ('permutest' permutation-based testing function).

Bootstrap analysis

We tested the robustness of our findings with respect to experiment-specific compositional variations and performed a bootstrap analysis across all samples. We generated 100 bootstrap sets of the same size as our original data set (77 samples) by drawing random samples with replacement. Then, we proceeded to split each set into four subsets, resembling our original natural sites and controlled environment experiments. We ensured that each subset contained 4 soil samples and removed all duplicate samples resulting in an average of 69.54 samples (±2.51 s.e.m) per bootstrap set. We repeated for each bootstrap set the original ANOVA-based analysis and determined the core RootOTUs shared among the natural site and greenhouse experiments. Fig. S20*B* depicts the members of core RootOTUs members for each bootstrap set and provides their taxonomic assignment at order rank.

Quantitative PCR of Thermomonosporaceae OTUs

The DNA samples of the greenhouse experiment #1 and the Eifel site, which were pyrosequenced, were also used as template for quantitative PCR (qPCR) validation. From our previous study we utilized the PCR primers that were designed on the basis of an Actinocorallia OTU (3). Pyrosequencing of these Actinocorallia PCR primer amplicons revealed that they match up to 100% to sequences belonging to the order Actinomycetales. We employed the PCR primer combination of 799F (14) and 904R (15) to generate a 16S rDNA amplicon to quantify the whole bacterial community in the DNA samples. The qPCR was performed using the same DFS Taq DNA Polymerase system (Bioron, Ludwigshafen, Germany) as for the library preparation described above with the exception that 0.5 µl of EvaGreen^{TM} dye (Biotium, Hayward, USA) in a total volume of 25 µl was used. Cycling conditions were 3' at 94°C, 40 cycles with 30'' at 94°C, 30'' at 55°C and 20'' at 72°C acquiring fluorescence followed by 10' at 72°C. We performed a melting curve analysis starting from 60°C to 95°C increasing by half degrees/per 10'' to determine the uniformity of the amplicons. We normalized the abundance of Actinocorallia PCR primer amplicons with the abundance of 799F-904R community amplicons. These values were then transformed to express the proportional abundances across all samples to compare with the quantification by pyrosequencing. The pyrosequencing determined relative abundance values of all ACM OTUs, which were assigned to the family Thermomonosporaceae, were summed to obtain the cumulative Thermomonosporaceae abundance per sample. These values were then also transformed to express the proportional abundances across all samples. Overall correlation between 454 and qPCR quantifications of Thermomonosporaceae abundance is 0.82 and 0.96 for the greenhouse experiment #1 samples and the Eifel site samples, respectively. The abundances of Thermomonosporaceae quantified by qPCR and pyrosequencing are depicted in Fig. S17.

Core root microbiota comparison across studies

We have downloaded the raw sequence data of the Bulgarelli et al. (2012), Lundberg et al. (2012) and Bodenhausen et al. (2013) studies from their respective data repositories for a comparative analysis with the data of this study. Sequences were concatenated to libraries according to the experimental design described in the respective manuscripts and we processed the data using the QIIME pipeline described above. With the script *split_libraries.py* we extracted the individual soil and root samples according to their barcodes and filtered for quality sequences. We utilized for the Bulgarelli and the Bodenhausen sequences the same quality filtering criteria (read length, Phred score, no ambiguous base calls in barcode and primer) as for the data of this study (see above). We filtered the Lundberg sequence data for read length (min. 220 bp) and quality (Phred 25) utilizing the quality criteria of their study (17). The sequences of the Lundberg and the Bodenhausen datasets were reversed to the complement sequence as they were barcode-indexed with and sequenced from the reverse primer. A single fasta file containing the quality sequences of the four datasets was used as input for the script *pick_otus.py*. We utilized reference-based OTU picking based on uclust (5), the Greengenes OTUs as reference database (7; release gg_otus_13_05) allowing the formation of clusters independently of database reference seeds. The latter results in de novo OTUs, in addition to OTUs identified in the reference database. We co-clustered the quality sequences of the four datasets into OTUs at 97 % sequence similarity and a common OTU table was prepared after trimming chimeric sequences/OTUs, removing of plant-derived sequences and taxonomic assignment of the OTU representative sequences as described above. This common OTU table and OTU representative fasta sequences are provided as Dataset S9 and Dataset S10, respectively. For the subsequent analysis in R, the common OTU table was splitted into OTU tables for each study. The common origin of these OTU tables permits the direct comparison of OTU IDs between studies. OTUs assigned to the phylum Chloroflexi were removed from the OTU table of the Bulgarelli dataset (see reference 3). The data of this study allowed a minimal sampling depth of 6,000 and the Bodenhausen of 4,500 sequences per sample. We chose the samples from the Bulgarelli and Lundberg studies to contain at least 1,000 sequences per sample. Samples with fewer sequences were removed from the OTU tables and the remaining samples were rarefied according to the sampling depth of each study (Dataset S7). The Lundberg dataset was finally represented by 80 soil (number of samples: Clayton replicate 1 n_{CL1} 22, n_{CL2} 18, Mason Farm replicate 1 n_{M1} 20, n_{M2} 20) and 265 root samples (n_{CL1} 48, n_{CL2} 35, n_{M1} 78, n_{M2} 104). The Bulgarelli data was represented in the analysis with 21 soil (number of samples: Cologne replicate 1 n_{C1} 2, n_{C2} 8, Golm replicate 1 n_{G1} 6, n_{G2} 5) and 32 root samples (n_{C1} 10, n_{C2} 8, n_{G1} 7, n_{G2} 7).

Subsequently, we defined for each study the ACM, i.e. OTUs with a minimum of 20 quality sequences in at least one of the samples within a study. The ACM contained 90, 152, 77 and 260 OTUs in the Bulgarelli, Lundberg, Bodenhausen and in this study, respectively. We then normalized the counts of individual ACM OTUs by dividing the total counts of all ACM OTUs within a sample followed by a multiplication by 1,000, representing per mill RA. The normalized ACM data of each study was examined separately for the core microbiota following the same logic and same statistic analysis pipeline described above. Due to the differences between reference-based and de novo OTU clustering, the data of this study was re-analyzed for comparison with the remaining datasets. RootOTUs - OTUs that are enriched in root compared to soil samples - constitute the basis of the analysis and were calculated between soil and root samples within each replicate experiment. Then, following the analysis logic, we proceeded to identify the RootOTUs that were shared between the natural sites, and the ones shared between the two greenhouse replicate experiments. Only OTUs identified with each of the 3 statistic approaches (see above) were defined as shared RootOTUs. The overlap between the shared RootOTUs of the natural sites and the shared RootOTUs of the greenhouse experiments was referred to as the core RootOTUs. The Lundberg dataset was examined following the same procedure where the Lundberg core RootOTUs present the overlap between the shared RootOTUs of the Clayton soil experiments and the shared RootOTUs of the Mason Farm soil replicates. Also here, shared RootOTUs were defined by the 3 statistic tests. The Bulgarelli core RootOTUs present the overlap between the RootOTUs found in the Cologne soil and the Golm soil experiments (RootOTUs were defined by the 3 statistic tests). We combined the samples of the replicate Cologne and Golm soil experiments, respectively, because the few soil samples in the first Cologne soil replicate prevented a statistically sensible identification of RootOTUs. Although a limited number of root endophyte samples were harvested, we chose to include the Bodenhausen et al. (2013) data in the analysis because the same PCR primer combination was used. This study does not include soil microbiota profiles, precluding the determination of RootOTUs. Therefore, we compared the root endophyte profiles across the four natural sites tested based on OTUs that are abundant at all sites, i.e. that have a minimal abundance of 5 per mille RA in all samples. Finally, we compared the core root microbiota of each study between the four studies.

All QIIME and R scripts used for computational analyses are available via http://www.mpipz.mpg.de/R_scripts.

References:

1. Van Rossum F*, et al.* (2004) Spatial genetic structure within a metallicolous population of Arabidopsis halleri, a clonal, self-incompatible and heavy-metal-tolerant species. *Molecular ecology* 13(10):2959-2967.

2. Willems G*, et al.* (2007) The genetic basis of zinc tolerance in the metallophyte Arabidopsis halleri ssp. halleri (Brassicaceae): an analysis of quantitative trait loci. *Genetics* 176(1):659-674.

3. Bulgarelli D*, et al.* (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* 488(7409):91-95.

4. Caporaso JG*, et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335-336.

5. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.

6. Haas BJ*, et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res* 21(3):494-504.

7. McDonald D*, et al.* (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *Isme Journal* 6(3):610-618.

8. Caporaso JG*, et al.* (2010) PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26(2):266-267.

9. Oksanen J*, et al.* (2012) vegan: Community Ecology Package. Version 2.0-3.

10. Kristiansson E, Hugenholtz P, & Dalevi D (2009) ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. *Bioinformatics* 25(20):2737-2738.

11. Benjamini Y, & Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57, 289–300.

12 Van Deun K*, et al.* (2009) Testing the hypothesis of tissue selectivity: the intersection–union test and a Bayesian approach. *Bioinformatics* 25(19):2588-2594.

13. Anderson MJ & Willis TJ (2003) Canonical Analysis of Principal Coordinates: a useful method of constrained ordination for ecology. Ecology 84:511–525.

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

15. Hodkinson B & Lutzoni F (2009) A microbiotic survey of lichen-associated bacteria reveals a new lineage from the Rhizobiales. *Symbiosis* 49(3):163-180.

16. Benson AK*, et al.* (2010) Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proc Natl Acad Sci U S A* 107(44):18933-18938.

17. Lundberg DS*, et al.* (2012) Defining the core Arabidopsis thaliana root microbiome. *Nature* 488(7409):86-90.

18. Bodenhausen N, Horton MW, & Bergelson J (2013) Bacterial Communities Associated with the Leaves and the Roots of *Arabidopsis thaliana*. *PLoS ONE* 8(2).

Supplementary Figures:

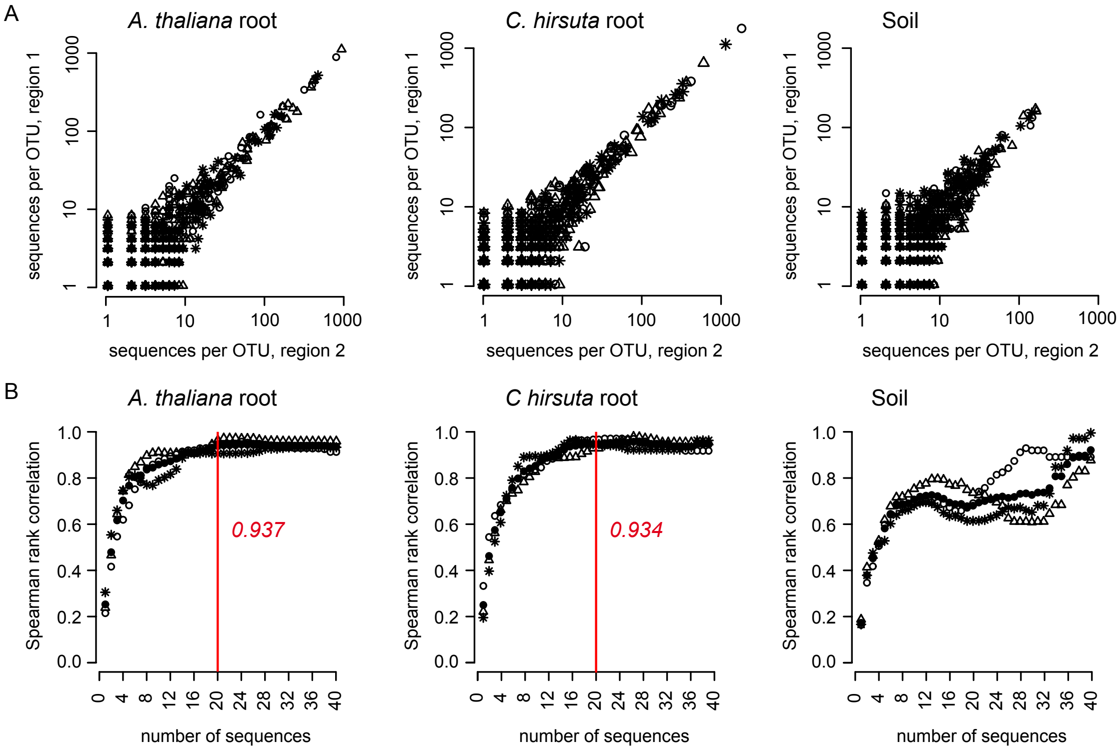

Fig. S1. Technical reproducibility of repeated library sequencing. Parallel emPCRs were conducted for the amplicon library of the first greenhouse replicate experiment (Library L28) and sequenced on separate regions of the 454 pico titer plate. Regions 1 and 2 generated 368,675 and 416,352 of quality sequences with a minimum of 8,615 sequences per sample. We defined OTUs on co-clustered quality sequences (chimera and plant-DNA sequence derived OTUs were removed) and corrected for differences in sequencing depths by rarefaction to 6,000 sequences per sample. (A) The number of quality sequences per OTU as retrieved from each region is plotted for the three *A. thaliana* root samples (GHrep1AtCol1, circles; GHrep1AtCol2, triangles; GHrep1AtCol3, snowflakes), the three *C. hirsuta* root samples (GHrep1ChOx1, circles; GHrep1ChOx2, triangles; GHrep1ChOx3, snowflake) and the three soil samples (GHrep1Soil1, circles; GHrep1Soil2, triangles; GHrep1Soil3, snowflakes). (B) Non-parametric Spearman rank correlation of OTU abundances in the samples shown in (A) and their mean correlation values (filled circles) are plotted as a function of progressive thresholds (1 to 40) for the minimal number of sequences per OTU in a sample of the dataset. For the root samples, the red line indicates the threshold of 20 sequences per OTU and the corresponding Spearman rank correlation value (mean of 3 samples) is given in the plots. The threshold we identified is similar to previous studies (3, 16, 17).

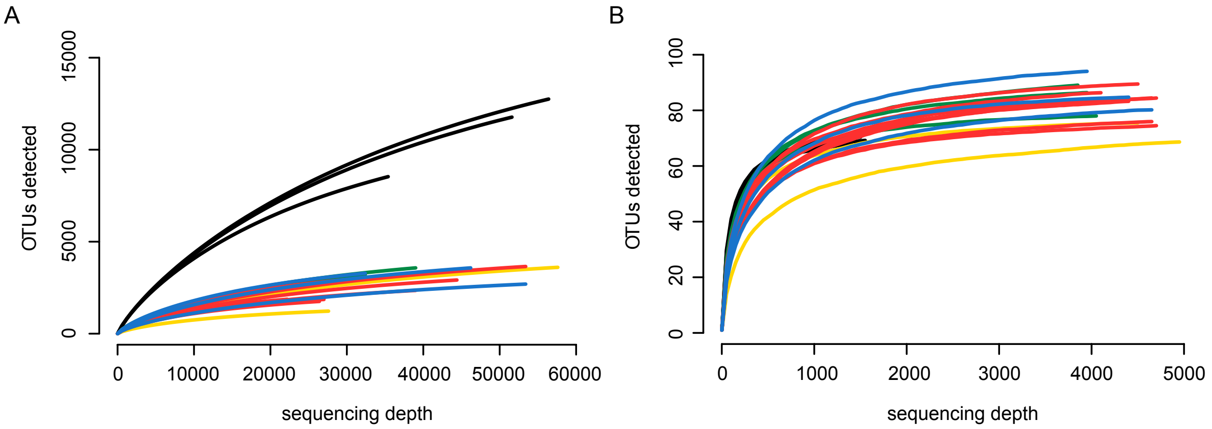

Fig. S2. Rarefaction analysis. The pooling of the sequences from both pico titer plate regions of the first greenhouse replicate experiment (Library L28) permitted to display the bacterial communities with a sequencing depth of 17,441 – 58,150 quality sequences per sample (Dataset S1). We defined OTUs on co-clustered quality sequences and performed rarefaction analysis for the soil (black) and root (colored) samples. Root samples include *A. halleri* (green), *A. lyrata* (yellow), *A. thaliana* (red) and *C. hirsuta* (blue). Rarefaction curves are based on all quality sequences obtained (A) and on the ACM dataset (B).

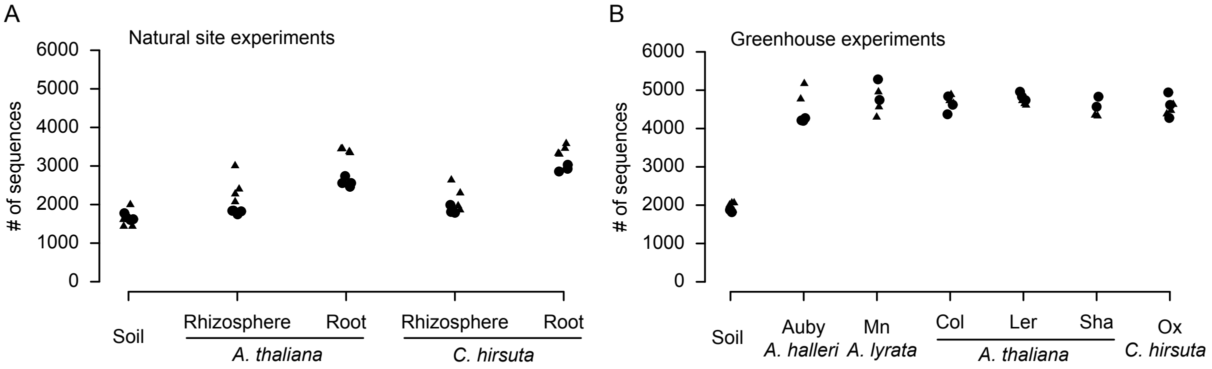

Fig. S3. Quality sequences after application of the threshold. Number of quality sequences in the ACM after thresholding the dataset to OTUs, which reach in at least one sample a minimum of 20 quality sequences. (A) Circles depict the Eifel site samples whereas triangles show Cologne site samples. (B) In the greenhouse experiments, the circles and triangles refer to the first and the second replicate experiment, respectively.

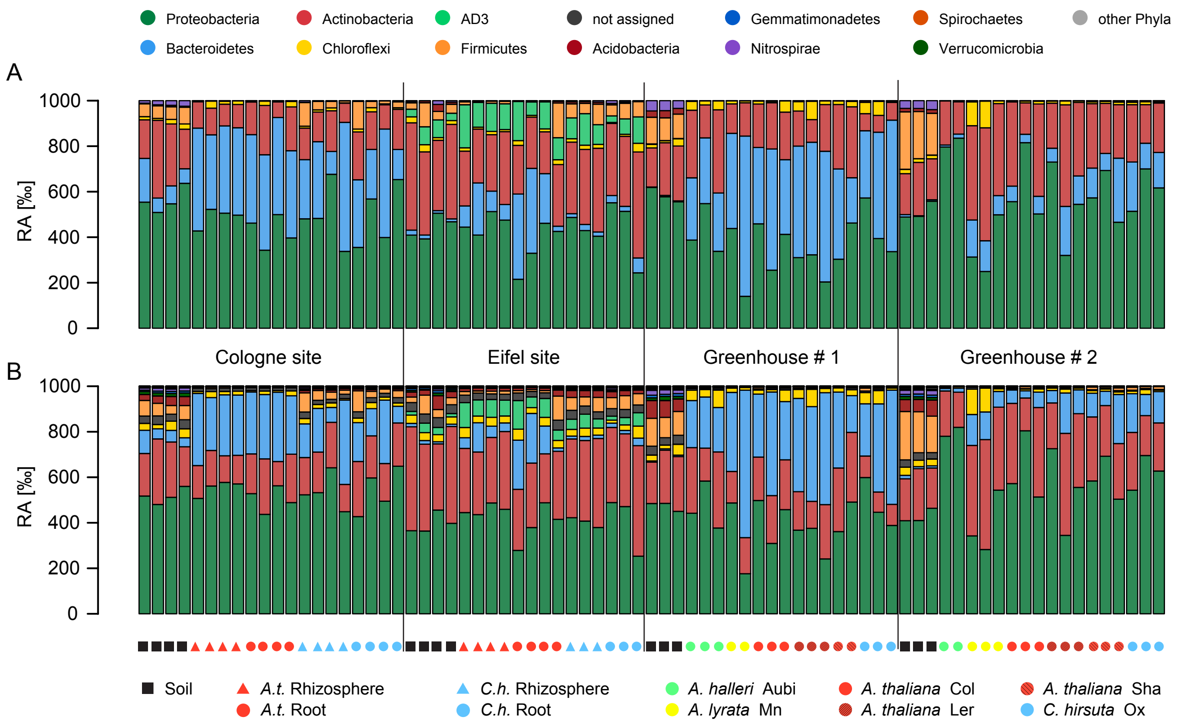

Fig. S4. Taxonomy overview. Taxonomic structure at the phylum rank of the ACM (A) and the TIC (B). Soil samples are marked with black squares and rhizosphere (triangles) and root samples (circles) are colored by plant species: *A. halleri* (green), *A. lyrata* (yellow), *A. thaliana* (red) and *C. hirsuta* (blue). The *A. thaliana* ecotypes discriminate by shading.

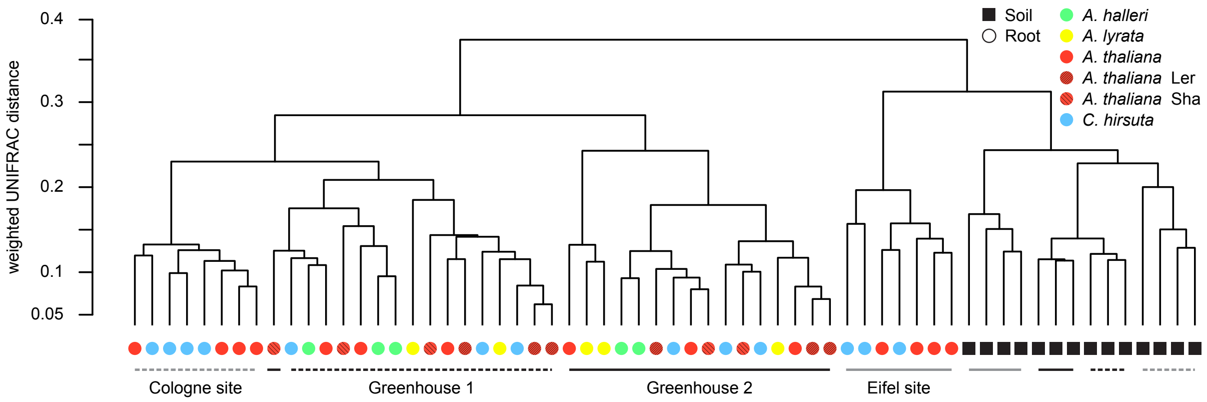

Fig. S5. Beta diversity. Between-sample diversity was calculated for TICs using weighted UniFrac distance metric (Phylogeny-based and sensitive to the sequence abundances) on 6,000 sequences per sample. The *A. thaliana* ecotype Col (non-shaded red) was used in the greenhouse experiments.

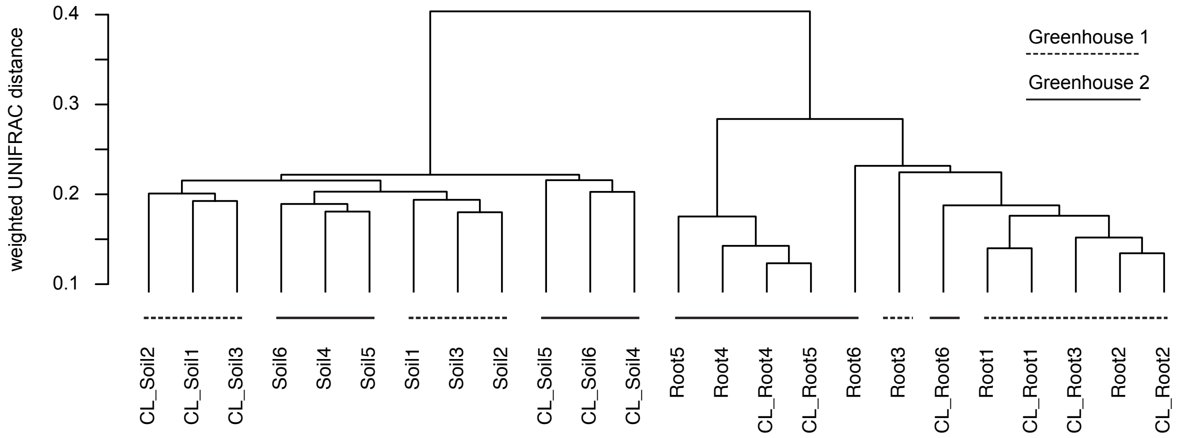

Fig. S6. Biological versus technical variation. DNA samples of the *3 A. thaliana* root (Ecotype Col-0) and 3 soil samples, from the greenhouse replicate #1 (indexes 1 to 3) and #2 (indexes 4 to 6), were utilized for community profiles generated in a common control library (CL, library L40). The weighted UniFrac distance was calculated based on communities containing 530 sequences per sample.

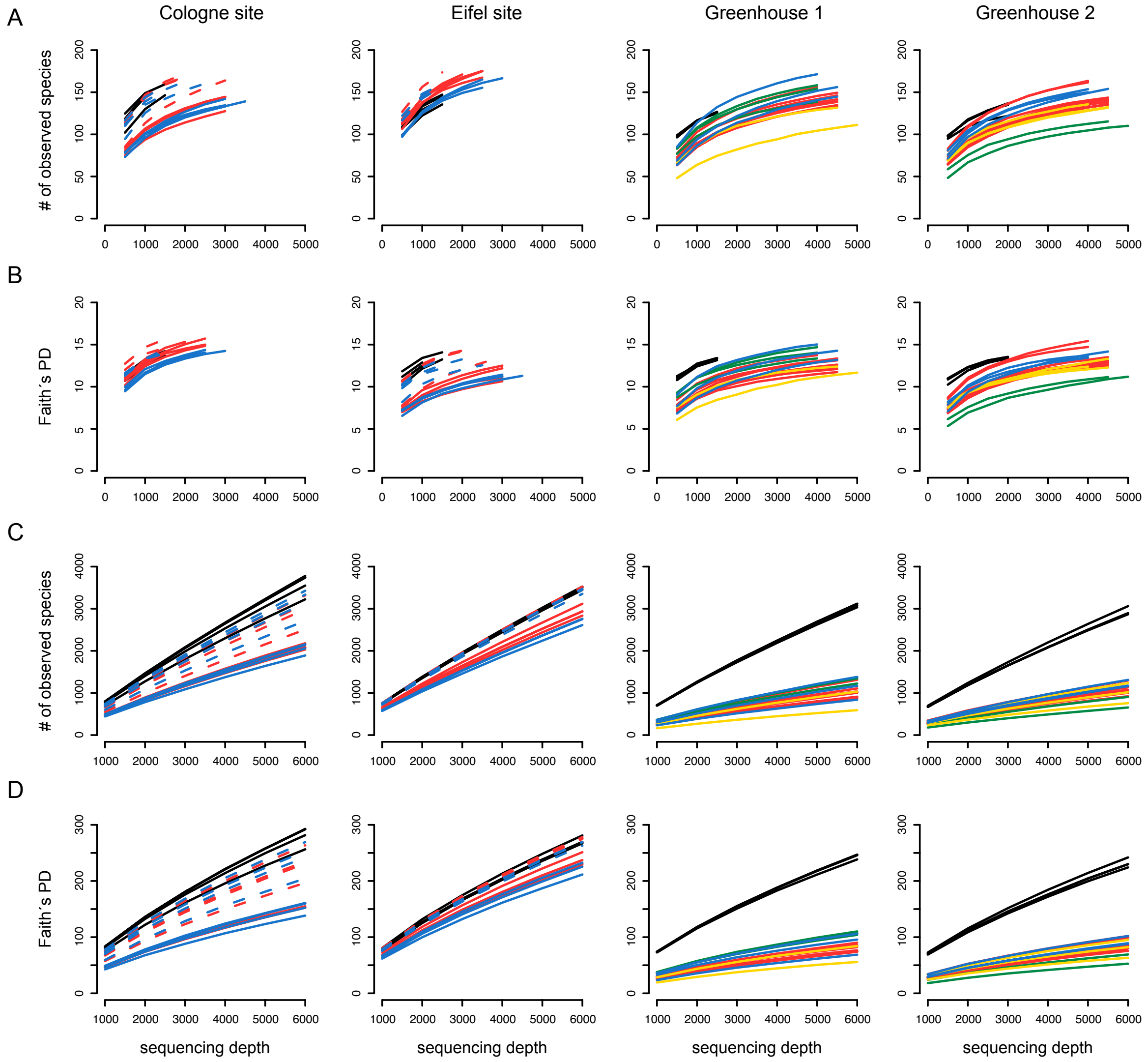

Fig. S7. Alpha diversity analyses. Within-sample diversity of the ACM (A and B) and TICs (C and D) was measured by OTU richness (A and C) and with Faith´s Phylogenetic Diversity (PD) metric (B and D). Alpha diversity is plotted as a function of the sequencing depth in the samples of the natural sites Cologne and Eifel and from the two replicate greenhouse experiments. Samples are colored as follows: Soil (black), *A. halleri* (green), *A. lyrata* (yellow), *A. thaliana* (red) and *C. hirsuta* (blue) root and rhizosphere samples with solid and hashed lines, respectively. Mean values of 100 rarefactions at each sampling depth are shown.

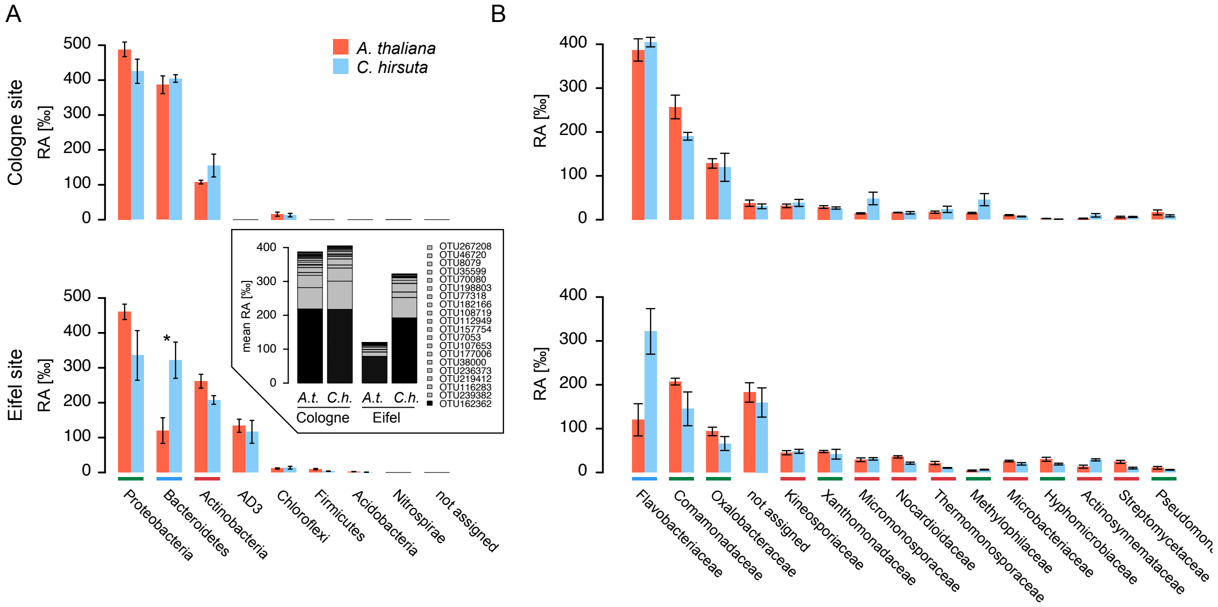

Fig. S8. Taxonomical profiles of *A. thaliana* and *C. hirsuta* root communities from Cologne and Eifel sites. Mean relative abundance (RA, ±s.e.m.) of taxa detected in root communities (color-coded by species) at the phylum (A) and the family rank (B). The affiliation of each family taxon (B) is color-coded to its corresponding phylum (A). Asterisks indicate significant differences between *A. thaliana* and *C. hirsuta* root communities (Tukey, $P < 0.1$ (FDR)). The inset in A reports the stacked abundances of individual OTUs assigned to the phylum Bacteroidetes in roots of *A. thaliana* (*A.t.*) and *C. hirsuta* (*C.h.*) for both sites and the dominant Flavobacterium OTU (OTU162362) is marked in black.

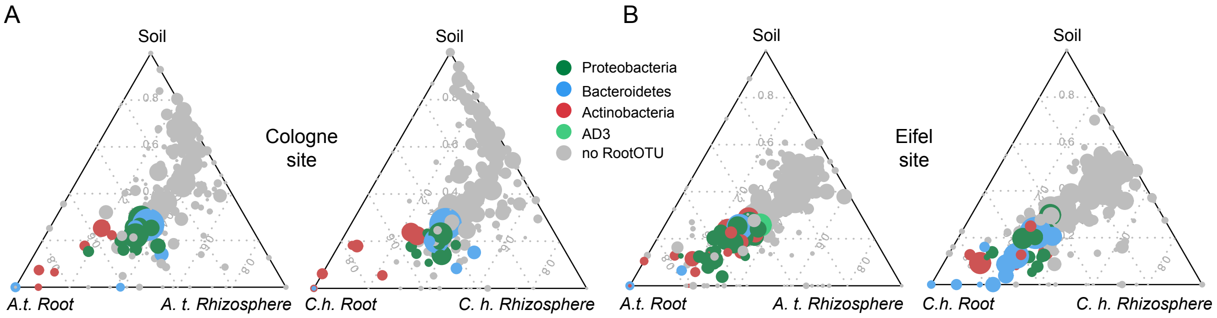

Fig. S9. RootOTUs per species and site. RootOTUs are OTUs that are enriched in roots compared to the corresponding soil communities (Tukey, $P < 0.1$ (FDR)). RootOTUs are colored by taxonomy and OTUs, which are not enriched in root communities, are plotted in grey. The ternary plots depict the relative occurrence of individual OTU (circles) in the indicated sample types of *A. thaliana* (*A.t.*) and *C. hirsuta* (*C.h.*) compared to soil for the Cologne (A) and the Eifel site (B). The size of the circles is proportional to the mean abundance in the community.

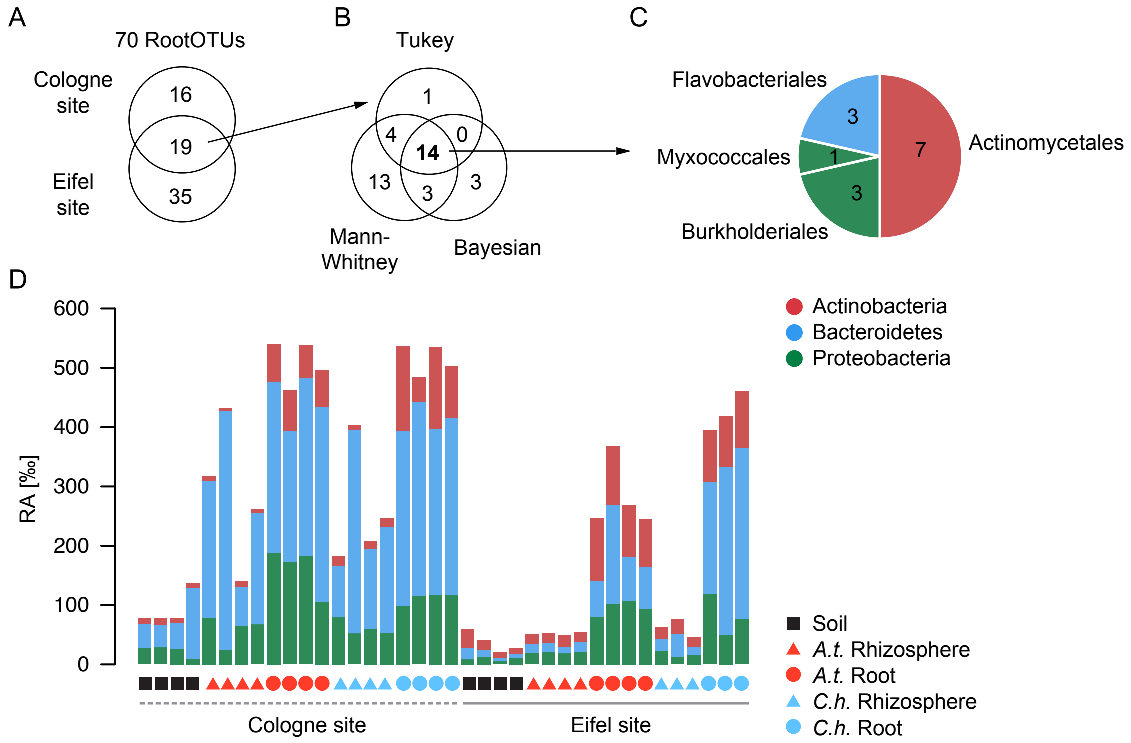

Fig. S10. Identification of shared RootOTUs of the natural site experiments. (A) The Venn diagram identifies 19 shared RootOTUs of the 70 RootOTUs of both sites Cologne and Eifel based on parametric statistics, (Tukey, $P < 0.1$ (FDR)). (B) RootOTUs shared between the two sites were identified with non-parametric Mann-Whitney and Bayesian statistics and we defined the ´shared RootOTUs´ from the validation by all three different statistical methods. (C) The pie chart reports the taxonomic composition of the 14 shared RootOTUs at the order rank. (D) Stacked relative abundance (RA) of the shared RootOTUs of the Cologne and Eifel sites. Each segment in the bar corresponds to one of the 14 shared RootOTUs.

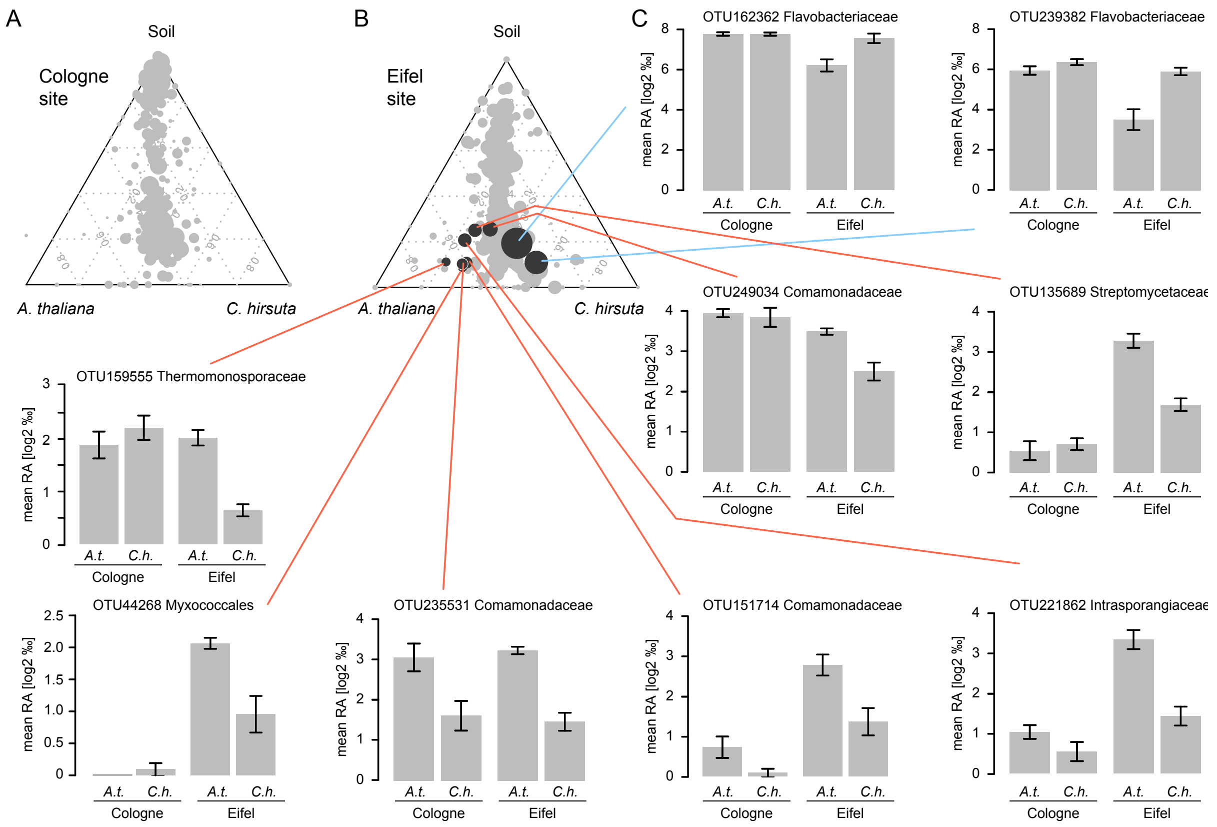

Fig. S11. Species-specific accumulation of OTUs. The ternary plots (corresponding to Figure 3) depict the relative occurrence of individual OTUs (circles) in root communities of *A. thaliana* and *C. hirsuta* compared to the respective soil microbiota for the Cologne (A) and the Eifel site (B). RootOTUs accumulating significantly different between *A. thaliana* and *C. hirsuta* are highlighted in dark grey (Tukey, *P* < 0.1 (FDR)) and are linked with their corresponding bargraphs (log2 abundance (±s.e.m.)). Red (*A. thaliana*) and blue (*C. hirsuta*) colored lines mark the species-specific enrichment of the RootOTUs. OTUs are labeled with OTU-ID and taxonomic assignments at family or order rank.

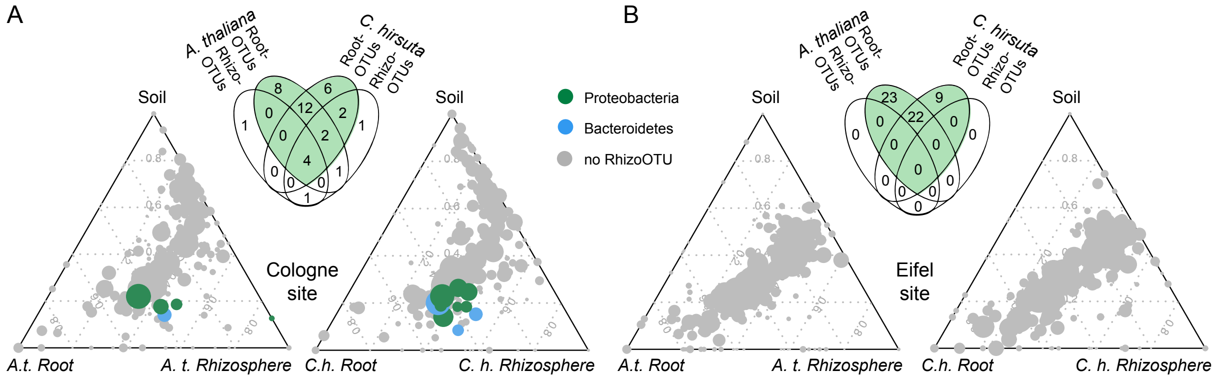

Fig. S12. Rhizosphere effect. RhizoOTUs are OTUs that are enriched in the rhizosphere samples compared to the corresponding soil communities of the natural sites (Tukey, $P < 0.1$ (FDR)). The RhizoOTUs are colored by taxonomy (Phylum rank) and OTUs, which are not enriched in rhizosphere communities, are plotted in grey. The ternary plots depict the relative occurrence of individual OTUs (circles) in the indicated sample types of *A. thaliana* (*A.t.*) and *C. hirsuta* (*C.h.*) compared to soil for the Cologne (A) and the Eifel site (B). The size of the circles is proportional to the mean abundance in the community. The Venn diagramm for each site compares the RhizoOTUs by species to the RootOTUs by species. The RootOTUs of both species are underlayed with green.

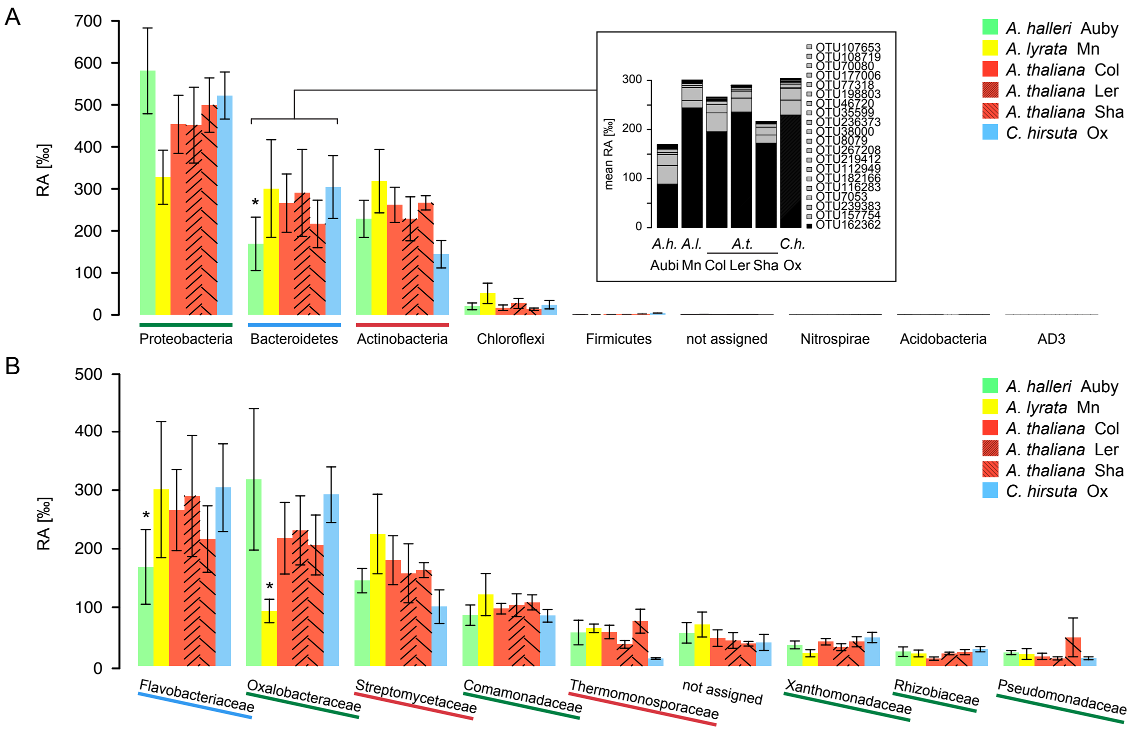

Fig. S13. Taxonomic profiles of *A. thaliana* and relative species grown under controlled conditions. Mean abundance (±s.e.m.) of taxa detected in root communities (colorcoded by species) of the ACM at the phylum (A) and the family rank (B). The 9 most abundant families are shown. The affiliation of each family taxon (B) is color-coded corresponding to its phylum (A). Asterisks indicate species-specific differences between the indicated species (Tukey; $P < 0.1$ (FDR)). The inset in A reports the stacked abundances of individual OTUs (OTU-IDs in the legend) assigned to the phylum Bacteroidetes in roots of *A. halleri* (*A.h.*), *A. lyrata* (*A.l.*), *A. thaliana* (*A.t.*) and *C. hirsuta* (*C.h.*) and the dominant Flavobacterium OTU (OTU162362) is marked in black.

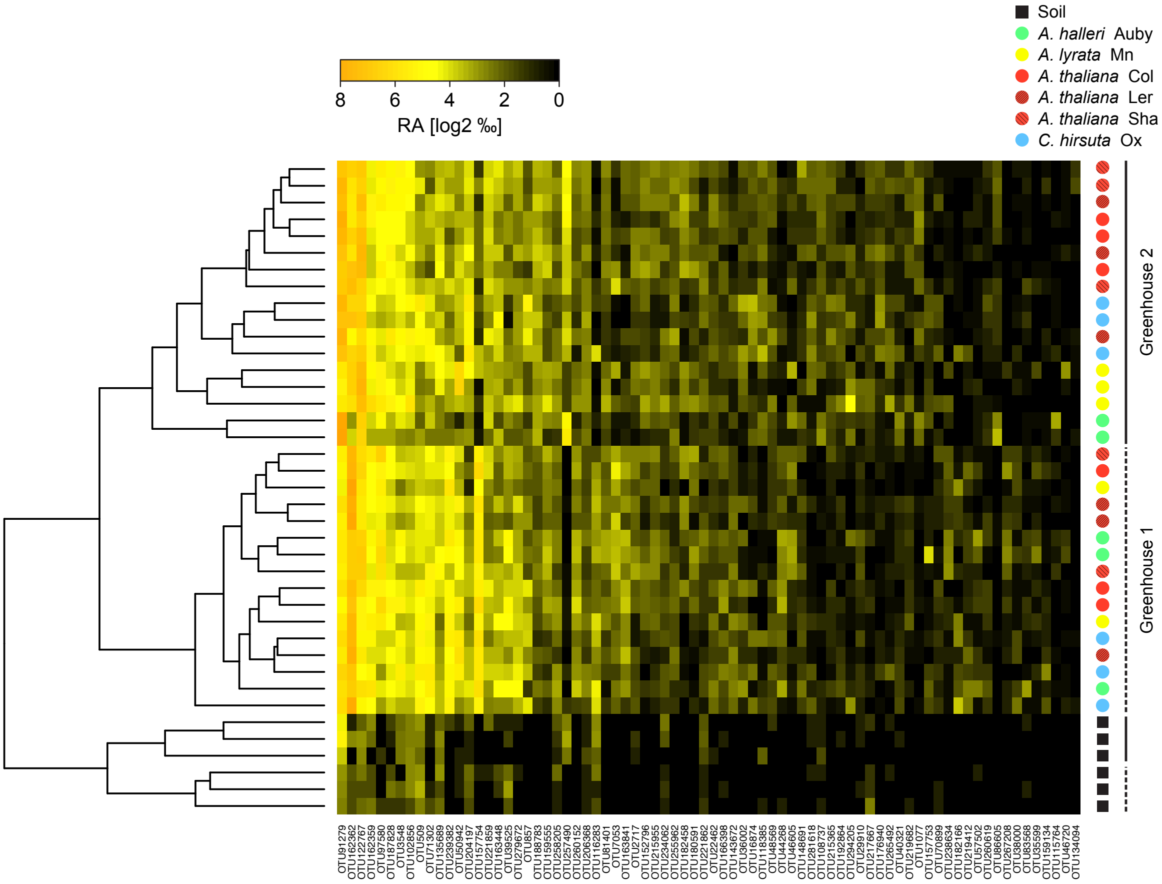

Fig. S14. The root-enriched microbiota of *A. halleri, A. lyrata, A. thaliana* and *C. hirsuta*. The abundance of RootOTUs of all greenhouse samples is displayed and sorted by mean rank abundance (x-axis). Hierarchical clustering is based on average Pearson distances.

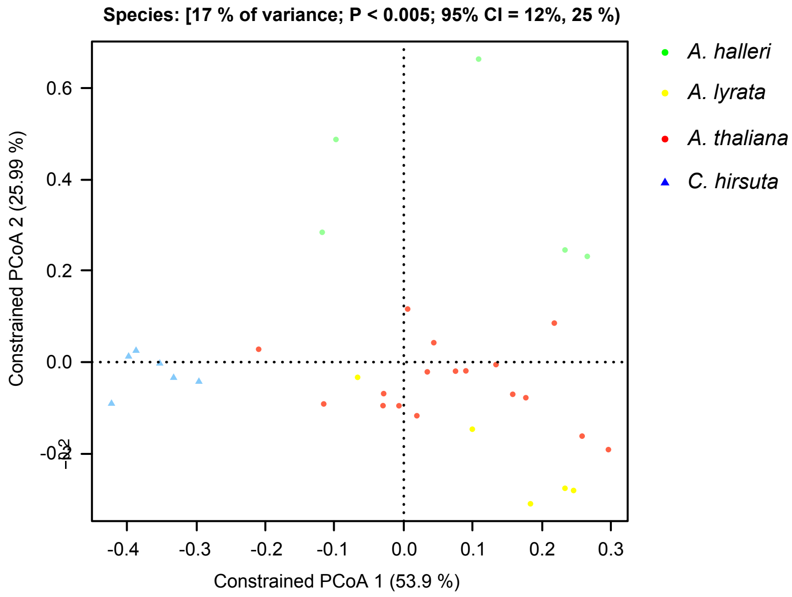

Fig. S15. Sample scores of RootOTU communities based on Bray-Curtis distances. The RootOTU communities of greenhouse samples were utilized for canonical analysis of principal coordinates, which was constrained for the variable host species. The corresponding OTU scores and sample arrows are presented in Fig. 4B.

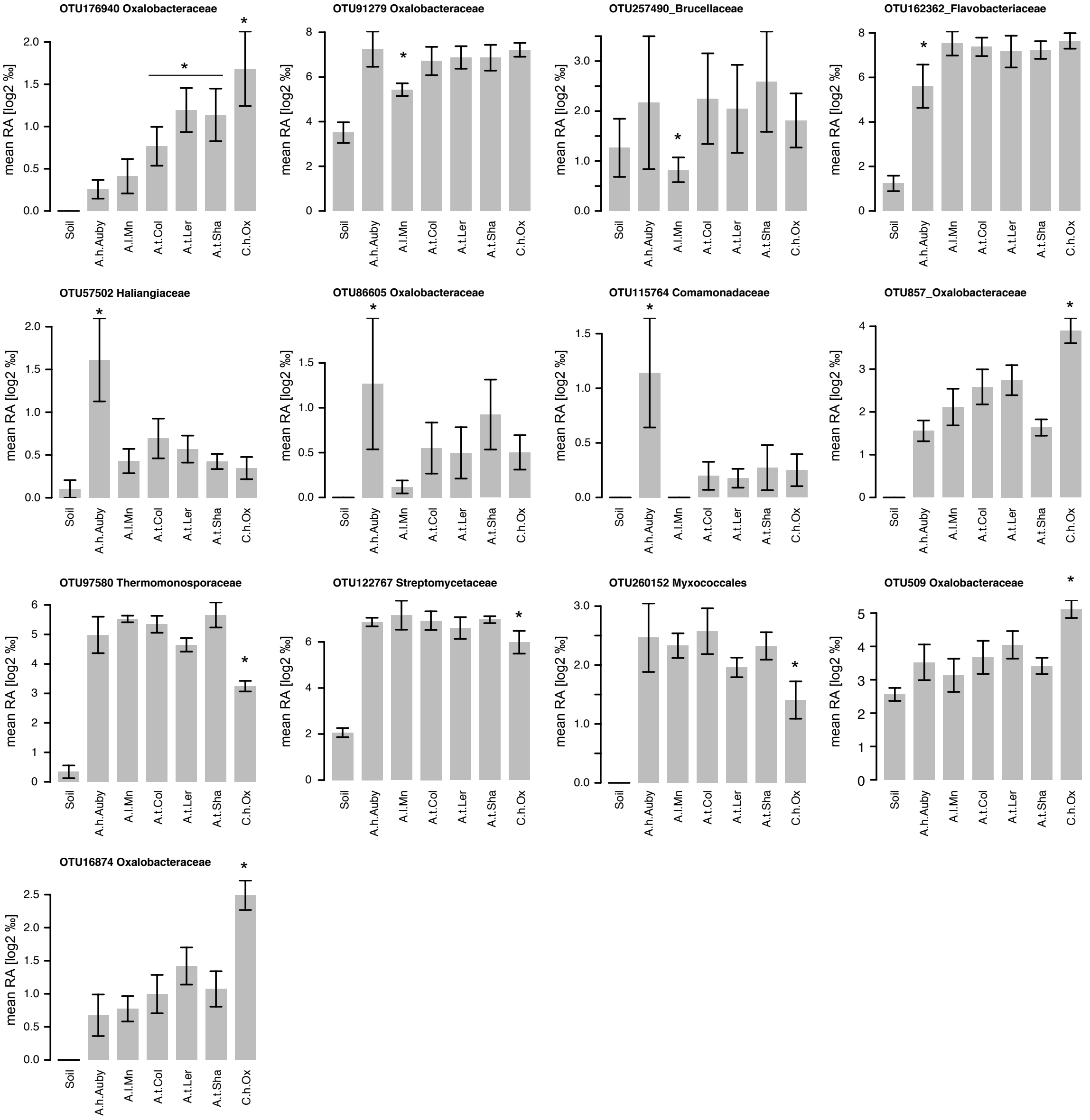

Fig. S16. Species-specific accumulation of RootOTUs. The 14 species-specific RootOTUs of the greenhouse experiments were identified based on Tukey posthoc test for the variable *species* (*P* < 0.1 (FDR)). The statistic tests revealed that OTU176940 discriminated *A. thaliana* and also *C. hirsuta* from the other species. The mean abundance of the species-specific RootOTUs is depicted in soil and root samples of the indicated species. The bargraphs report the variation (±s.e.m.) of the average log2 abundance over both replicate experiments. The asterisks are placed on the species as identified by the Tukey test. OTUs are marked with OTU-ID, and taxonomic assignments at family rank.

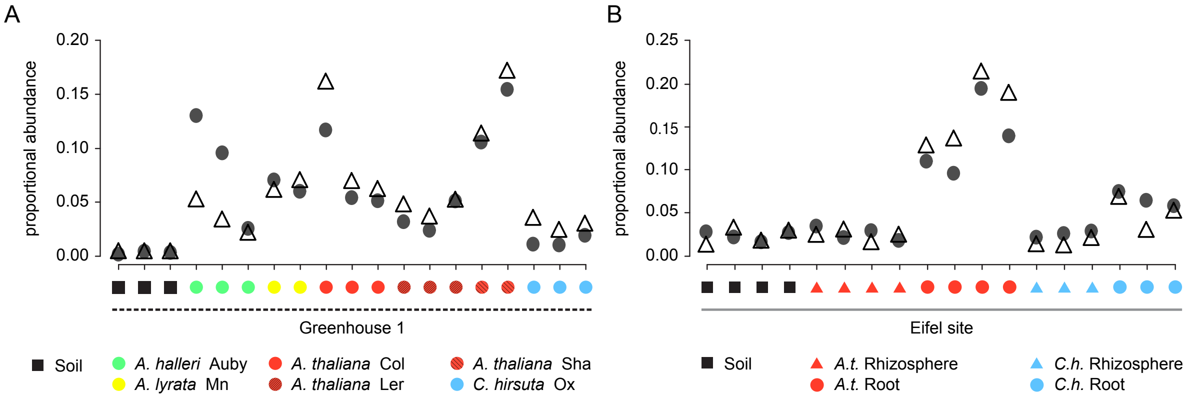

Fig. S17. Validation of Thermomonosporaceae OTU accumulation by quantitative PCR. The DNA samples of the greenhouse experiment #1 (A) and the Eifel site (B) were used as template for qPCR analysis of Thermomonosporaceae accumulation (open triangles). For the pyrosequencing quantification (solid circles), the relative abundance values [‰] of all OTUs of the ACM assigned to the family Thermomonosporaceae were summed to obtain the cumulative Thermomonosporaceae abundance per sample. These values were then transformed to express the proportional abundances across all samples. The qPCR protocol normalizes the abundance of Thermomonosporaceae amplicons with the abundance of a 16S rDNA community amplicon. These normalized values were then also transformed to express the proportional abundances across all samples. Overall correlation between 454 and qPCR quantifications of Thermomonosporaceae abundance is 0.82 and 0.96 for the greenhouse experiment #1 samples and the Eifel site samples, respectively.

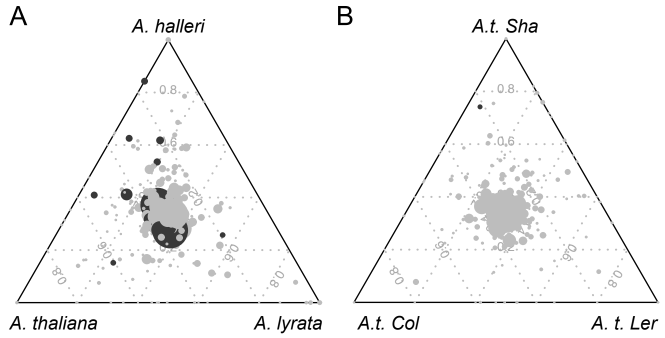

Fig. S18. Increased inter- compared to intraspecies variation in Arabidopsis root microbiota composition. Ternary plots depict relative OTU occurrence of the ACM in root communities of indicated Arabidopsis species (A) or *A. thaliana* ecotypes (B). Each circle represents an OTU and the size of the circle is proportional to the OTU´s abundance. The black colored OTUs in (A) refer to the 13 RootOTUs that vary by *species* and (B) an OTU that varies by the factor *genotype* (ANOVA, *F* < 0.1 (FDR)).

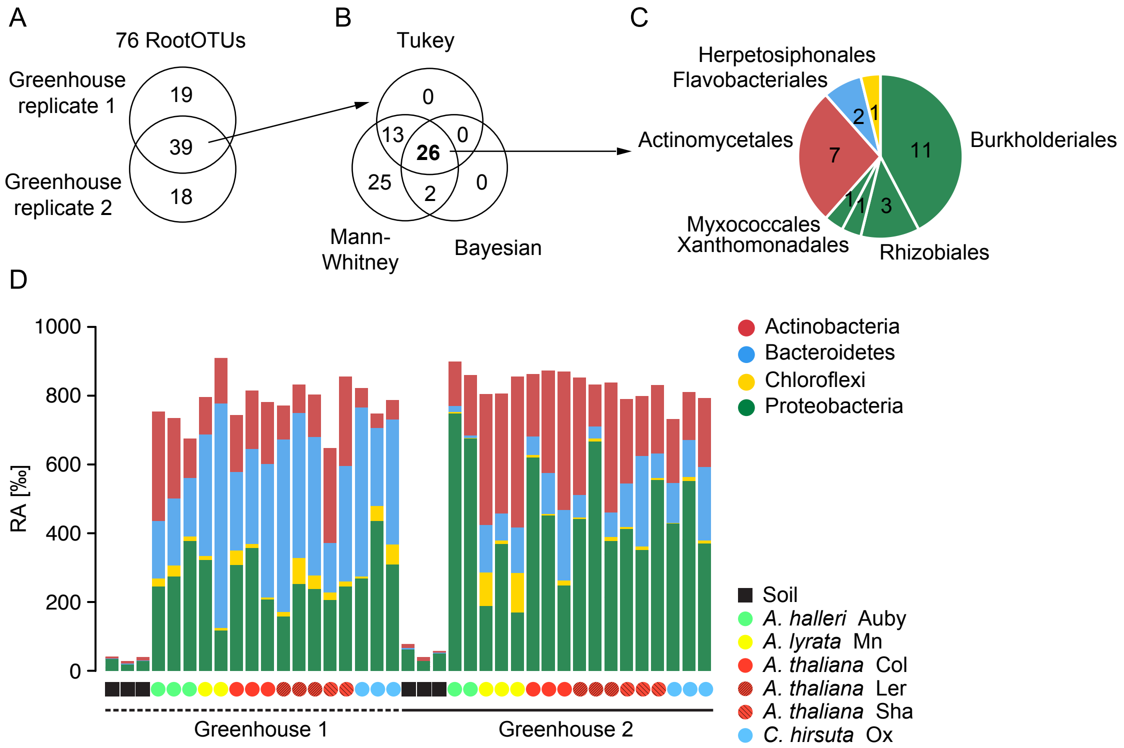

Fig. S19. Identification of the shared RootOTUs from the greenhouse experiments. (A) The Venn diagram determines 39 shared RootOTUs from the 76 RootOTUs of all species in the replicate experiments based on parametric statistics (Tukey, $P < 0.1$ (FDR)). (B) RootOTUs shared between the two replicates were identified with non-parametric Mann-Whitney and Bayesian statistics and we defined the ´shared RootOTUs´ from the validation by the three different statistical methods. (C) The pie chart reports the taxonomic composition of the 26 shared RootOTUs at the order rank. (D) Stacked relative abundance (RA) of the shared RootOTUs of both greenhouse experiments. Each segment in the bar corresponds to one of the 26 shared RootOTUs.

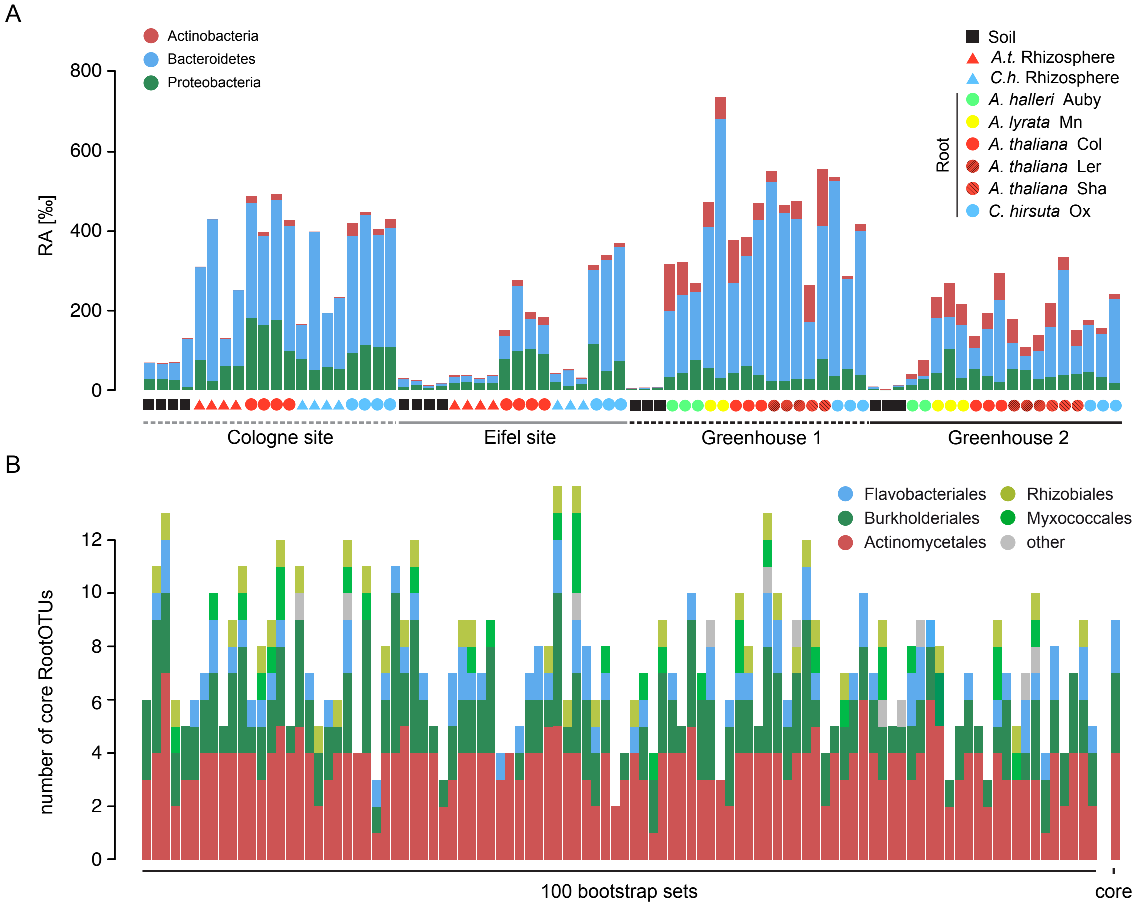

Fig. S20. Core RootOTUs. Stacked OTU abundance of the core RootOTUs identified between that natural site and the replicate greenhouse experiments (A). Each segment in the bar corresponds to an OTU color-coded by phylum. (B) Core RootOTUs detected in 100 bootstrap sets are plotted in stacked columns per bootstrap set and colored according to their taxonomic assignment at the order rank. The column to the right (marked with ′core′) represents the triad core RootOTUs as identified for the original data set.

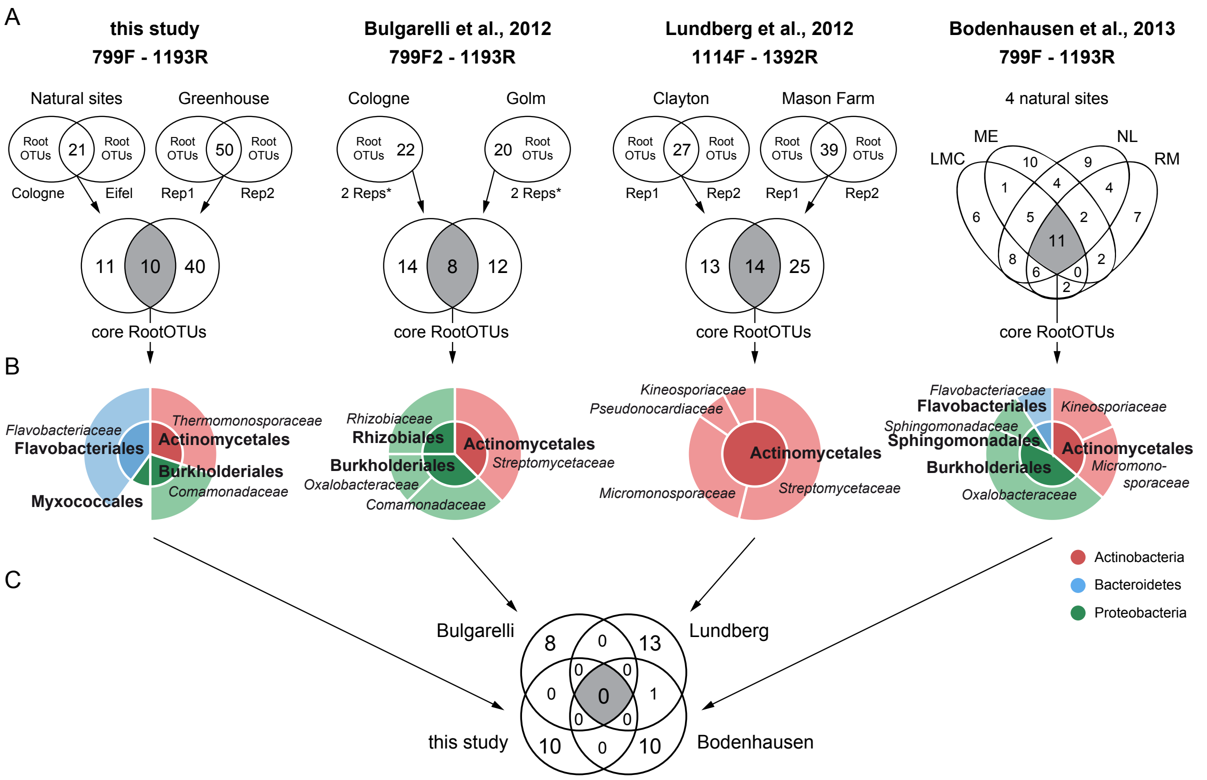

Fig. S21: Core root microbiota comparison across studies. Analysis of the core root microbiota using the datasets from the Bulgarelli et *al*. (2012), Lundberg et *al*. (2012) and Bodenhausen et *al*. (2013) studies and comparison the core RootOTUs of this study. These studies have the examination of *A. thaliana* root endophyte communities across different soil types and environments in common, but based on the following PCR primer combinations: 799F – 1193R (this study and reference 18), 799F2 – 1193R (3) and 1114F – 1392R (17). Using QIIME, the sequences of the soil and root samples of these four studies were co-clustered into OTUs at 97 % sequence similarity using the Greengenes reference dataset. For each study, a corresponding OTU table was extracted and its core root microbiota was determined. The common origin of the OTU table permits the direct comparison of OTU IDs between studies. The bacterial community profiles of this study, the Bodenhausen, Bulgarelli and Lundberg studies were covered by 6,000, 4,500, 1,000 and 1,000 sequences per sample, respectively. For each study the ACM were defined and examined with the same statistic analysis pipeline as used for the main analysis of this study. For each dataset, we followed the same procedure to identify the core microbiota: (A) OTUs that are enriched in root compared to soil samples (RootOTUs) were calculated within each replicate (Rep#) experiment, the shared RootOTUs (number is given) between replicated experiments were determined and the core defined from the overlap between soil types. Replicate samples of the Bulgarelli study were combined (*) within each soil type due to low sample number. We compared the root

endophyte communities of the Bodenhausen study since the same PCR primer combination was used. The root endophyte samples were collected from four natural sites in the US (Lake Michigan College, LMC; Michigan Extension, ME; North Liberty, NL and Route Marker, RM). (B) Taxonomic assignments of the core root OTUs are reported at the order (bold) and family (italic) ranks in the center and the outer ring of the pie charts, respectively. The segments of the pie charts are colored by the bacterial phyla of the corresponding OTU. No reliable taxonomy assignment at family rank was obtained for the Myxococcales OTU in the core root microbiota of this study. (C) The Venn diagram dissects the relative membership of core OTUs between the four tested studies.

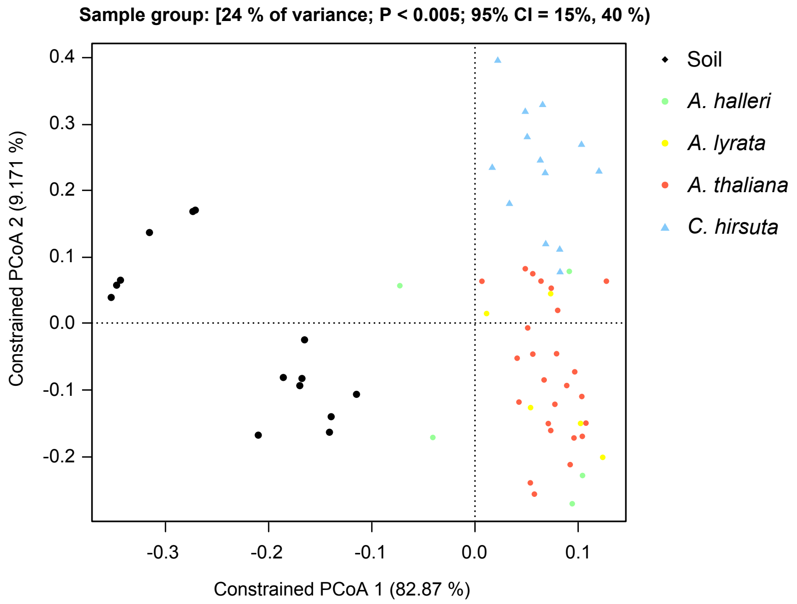

Fig. S22. Sample scores of RootOTU communities based on Bray-Curtis distances. The RootOTU communities of soil and root samples both of the natural site and greenhouse experiments were utilized for canonical analysis of principal coordinates, which was constrained for *sample group*. Sample groups included all root samples by species and the soil samples as additional group. The corresponding OTU scores and sample arrows are presented in Fig. 5*B*.

Table S1. Soil parameters. Geochemical characterization of the natural experimental ´Cologne soil´ (CS) batches ‘CS-4’ and ‘CS-5’ and the soils from the natural sites. Soil parameters (A) and macronutrients (B).

A

Experiment	Soil	[1]C. org. (%)	Clay (%)	Silt (%)	Sand (%)	pH	[2]Classification
Greenhouse	**CS-4**	1.4	13.4	37.3	49.3	6.95	sandy loam
Greenhouse	**CS-5**	4.0	21.0	31.0	48.0	6.94	Loam
Natural site	**Cologne**	3.5	21.0	31.0	48.0	6.98	Loam
Natural site	**Eifel**	4.0	16.0	31.0	53.0	6.14	sandy loam

[1] organic carbon

[2] Soil texture classification according FAO

B

Experiment	Soil	Extract	[3]N	[3]P	[3]K	[3]Ca	[3]Mg
Greenhouse	**CS-4**	[1]H2O	14.90	8.56	27.76	52.29	8.23
		[2]AAE		96.27	146.09	1572.70	118.40
Greenhouse	**CS-5**	[1]H2O	22.20	6.28	22.58	52.08	8.90
		[2]AAE		85.30	124.80	1604.10	118.50
Natural site	**Cologne**	[1]H2O	4.70	5.59	31.97	166.50	9.46
		[2]AAE		71.60	123.80	12021.50	224.60
Natural site	**Eifel**	[1]H2O	1.50	9.28	43.09	56.08	26.34
		[2]AAE		17.70	178.00	2693.70	506.90

[1] determined with 1:10 (w/v) H2O extract as a proxy for plant-available nutrients

[2] determined with 1:10 (w/v) ammonium-acetate-EDTA (AAE) extract as a proxy for reserve-nutrients

[3] mg/kg

Table S2. Touch-down PCR program. Thermal cycling conditions utilized to generate barcoded amplicon libraries.

Step #	Temperature [°C]	Time [seconds]	# of cycles
1	94	120	
2	94	30	5x
3	58	60	
4	72	15	
5	94	30	5x
6	57	60	
7	72	30	
8	94	30	5x
9	56	60	
10	72	45	
11	94	30	20x
12	55	60	
13	72	60	
14	72	600	
15	15	pause	

Table S3. Shapiro-Wilk analysis. For each of the data subsets, which we have examined with ANOVA (Supporting information), we tested normal distribution of data points for each taxon/OTU using the Shapiro-Wilk test. All ANOVA tests are listed in the table with the corresponding Figure number in the manuscript and the experiments (sites/replicates) and sample types included in the analysis. For each test the analysis level and the number of taxa/OTUs per level that were examined are indicated. The results of the Shapiro-Wilk tests are given with number of taxa/OTUs per analysis level (also as percentage) for which the null hypothesis (data points are normally distributed) was not rejected.

Figure	**Sites/ Replicates**	**Samples**	**Analysis**	**Taxa/OTUs**	**[1]SW**	**%**
S8A	both natural sites	root	Phylum	9	**3**	**33.3**
S8B	both natural sites	root	Family	51	**28**	**54.9**
3A	Cologne site	soil, root & rhizosphere	OTUs	*227	**89**	**39.2**
3B	Eifel site	soil, root & rhizosphere	OTUs	*229	**112**	**48.9**
S11	both natural sites	root	RootOTUs	70	**47**	**67.1**
S13A	both replicates	root	Phylum	*8	**3**	**37.5**
S13B	both replicates	root	Family	*50	**15**	**30.0**
S19A	replicate 1	soil, root & rhizosphere	OTUs	*225	**59**	**26.2**
S19A	replicate 2	soil, root & rhizosphere	OTUs	*226	**68**	**30.1**
4C	both replicates	root	RootOTUs	76	**46**	**60.5**
S18	both replicates	root (A.h., A.l. & A.t.)	OTUs	*225	**72**	**32.0**

[1] Shapiro-Wilk statistics: the number of taxa/OTUs with $P > 0.05$ is reported

*Note, not all of the 9 Phyla, 51 Families or 237 ACM OTUs were present in the individual data subsets

Supplementary Dataset Legends

Dataset S1. Experimental design. This excel file *Dataset_S1.xlsx* contains for each sample the detailed experimental information about the type of experiment, replicate, type of sample, plant species and plant genotype. Further it contains sequencing related information such as library-ID, barcode sequences, number of generated raw sequences, number of quality sequences, number of quality sequences in the ACM data subset and the column *SRA_filename* lists the name of the raw fasta file, as stored at the short read archive (SRA).

Dataset S2. OTU table of the experiment 1 by regions.

The excel document *Dataset_S2.xlsx* documents the analysis of technical reproducibility of community profiles. The samples of the greenhouse replicate #1 (L28) were used for parallel sequencing and the resulting raw data was co-clustered using QIIME. A first work sheet includes the QIIME mapping file and the second work sheet contains the OTU table where plant-sequence-derived OTUs were removed.

Dataset S3. OTU table of the ACM.

The tab-delimited text file *Dataset_S3.txt* contains the data matrix of the abundant community members (ACM). The data matrix contains OTU counts per sample, the Greengenes taxonomy and the OTU representative fasta sequences.

Dataset S4. OTU table of the TICs.

This tab-delimited text file *Dataset_S4.txt* presents the data matrix of the threshold-independent community (TIC). This OTU table was rarefied at a sampling depth of 6,000 sequences per sample and was utilized for OTU-based and beta diversity analyses. The data matrix contains OTU counts per sample, the Greengenes taxonomy and the OTU representative fasta sequences.

Dataset S5. OTU table of the common sequencing library analysis.

The tab-delimited text file *Dataset_S5.txt* contains the data file used to examine the common sequencing library L40. The file contains the rarefied OTU table with a sampling depth of 530 sequences per sample, which was the QIIME output utilized for beta-diversity analysis.

Dataset S6. Summary table with the statistic test results.

The excel file *Dataset_S6.xlsx* contains the results of all statistic tests performed organized in separate worksheets. Worksheets are alphabetically indexed and the name of the worksheet contains the type of data (Phylum, Family, OTU, RootOTUs) analyzed and the statistical test (ANOVA, Tukey, Mann-Whitney (MW)) used. Natural site (NS) and greenhouse (GH) experiments are marked. The worksheet contains the model utilized for the analysis. BH: Benjamini and Hochberg method for adjusting *P* values for multiple hypothesis testing.

Dataset S7. Design, rarefied OTU counts and taxonomy of data subsets from the *A. thaliana* root microbiome comparison. The excel document *Dataset_S7.xlsx* contains a first worksheet with the detailed experimental information about the soil type, replicate number, type of sample, plant species and plant genotype for the soil and root samples from this study and the Bulgarelli et al. (2012), Lundberg et al. (2012) and Bodenhausen et al. (2013) studies. Additional worksheets contain the rarefied OTU tables for this study, the Bodenhausen, the Bulgarelli and the Lundberg studies with sampling depths of 6,000, 4,500, 1,000 and 1,000 sequences per sample, respectively. Finally, a worksheet containing the corresponding Greengenes taxonomy assignments is provided. Note, the OTU representative fasta sequences can be retrieved from Dataset S10.

Dataset S8. Raw OTU table. The file *Dataset_S8.biom* corresponds to the OTU table built using the clustered sequence/OTU information per sample and the corresponding Greengenes taxonomy assignments. Plant-sequence-derived OTUs were removed from this OTU table. The file *Dataset_S8.biom* is available at our homepage together with the R scripts used for data analyses (http://www.mpipz.mpg.de/R_scripts).

Dataset S9. Raw OTU table from the *A. thaliana* root microbiome comparison. The file *Dataset_S9.biom* corresponds to the OTU table as resulted from the co-clustering of the sequences from this study with the sequences from Bulgarelli et al. (2012), Lundberg et al. (2012) and Bodenhausen et al. (2013). This common OTU table does not contain chimeric and plant-sequence-derived OTUs. The provided OTU taxonomies were obtained from the Greengenes database (release gg_otus_13_05). The file *Dataset_S9.biom* is available at our homepage together with the R scripts used for data analyses (http://www.mpipz.mpg.de/R_scripts).

Dataset S10. OTU representative sequences. The file *Dataset_S10.fasta* ("gzipped") contains the OTU representative fasta sequences corresponding to the Dataset S9. The file *Dataset_S10.fasta* is available at our homepage together with the R scripts used for data analyses (http://www.mpipz.mpg.de/R_scripts).