

Functional overlap of the *Arabidopsis* leaf and root microbiota

Yang Bai^{1*}, Daniel B. Müller^{2*}, Girish Srinivas^{1*}, Ruben Garrido-Oter^{1,3,4*}, Eva Potthoff², Matthias Rott¹, Nina Dombrowski¹, Philipp C. Münch^{5,6,7}, Stijn Spaepen¹, Mitja Remus-Emsermann², Bruno Hüttel⁸, Alice C. McHardy^{4,5}, Julia A. Vorholt^{2*} & Paul Schulze-Lefert^{1,4*}

Roots and leaves of healthy plants host taxonomically structured bacterial assemblies, and members of these communities contribute to plant growth and health. We established *Arabidopsis* leaf- and root-derived microbiota culture collections representing the majority of bacterial species that are reproducibly detectable by culture-independent community sequencing. We found an extensive taxonomic overlap between the leaf and root microbiota. Genome drafts of 400 isolates revealed a large overlap of genome-encoded functional capabilities between leaf- and root-derived bacteria with few significant differences at the level of individual functional categories. Using defined bacterial communities and a gnotobiotic *Arabidopsis* plant system we show that the isolates form assemblies resembling natural microbiota on their cognate host organs, but are also capable of ectopic leaf or root colonization. While this raises the possibility of reciprocal relocation between root and leaf microbiota members, genome information and recolonization experiments also provide evidence for microbiota specialization to their respective niche.

Plants and animals harbour abundant and diverse bacterial microbiota¹. These taxonomically structured bacterial communities have important functions for the health of their multicellular eukaryotic hosts^{2–4}. The leaf and root microbiota of flowering plants have been extensively studied by culture-independent analyses, which have consistently revealed the co-occurrence of four main bacterial phyla: Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria^{5–15}. Determinants of microbiota composition at lower taxonomic ranks, that is, at genus and species level, are host compartment, environmental factors and host genotype^{6,7,12,16}.

Soil harbours an extraordinary rich diversity of bacteria and these define the start inoculum of the *Arabidopsis thaliana* root microbiota^{6,7}. The inoculum source of the leaf microbiota is thought to be more variable owing to the inherently open nature of the leaf ecosystem, probably involving bacteria transmitted by aerosols, insects, or soil^{8,9,17}. A recent study of the grapevine (*Vitis vinifera*) microbiota showed that the root-associated bacterial assemblies differed significantly from aboveground communities, but that microbiota of leaves, flowers, and grapes shared a greater proportion of taxa with soil communities than with each other, suggesting that soil may serve as a common bacterial reservoir for belowground and aboveground plant microbiota¹⁸.

A major limitation of current plant microbiota research is the lack of systematic microbiota culture collections that can be employed in microbiota reconstitution experiments with germ-free plants to address principles underlying community assembly and proposed microbiota functions for plant health under laboratory conditions¹⁹.

Bacterial culture collections from roots and leaves

We employed three bacterial isolation procedures to establish taxonomically diverse culture collections of the *A. thaliana* root and leaf microbiota. Bacterial isolates were recovered from pooled or individual

roots or leaves of healthy plants using colony picking from agar plates, limiting dilution in liquid media in 96-well microtitre plates, or microbial cell sorting (see Methods). We adopted a two-step bar-coded pyrosequencing protocol²⁰ for taxonomic classification of the cultured bacteria by determining ≥ 550 base pairs (bp) 16S ribosomal RNA (rRNA) gene sequences (Supplementary Fig. 1; Methods). In parallel, parts of the root and leaf material was used for cultivation-independent 16S rRNA gene community sequencing to cross-reference Operational Taxonomic Unit (OTU)-defined taxa from the microbiota with individual colony forming units (CFUs) in the culture collections.

A total of 5,812 CFUs were recovered from 59 independently pooled *A. thaliana* root samples of plants mainly grown in Cologne soil, Germany, whereas 2,131 CFUs were retrieved from leaf washes of individual leaves collected from *A. thaliana* populations at six locations near Tübingen, Germany, or Zurich, Switzerland (Supplementary Data 1). Recovery estimates for root-associated OTUs were calculated using the culture-independent community profiles of the present and two earlier studies^{6,12} and varied for the top 100 OTUs (70% of sequencing reads) between 54–65% and at $\geq 0.1\%$ relative abundance (RA) between 52–64% (Methods; Extended Data Fig. 1a–c; Supplementary Data 2). For leaf samples, the culture-independent 16S rRNA gene analyses from individual and pooled leaves (60 samples from six sites) revealed similar community profiles at all tested geographic sites and high leaf-to-leaf consistency (Extended Data Fig. 2). Recovery estimates of the top 100 leaf-associated bacterial OTUs (86% of all sequencing reads) were 54% and at $\geq 0.1\%$ RA 47% (Extended Data Fig. 1d). The root-derived CFUs correspond to 23 of 38 and the leaf-derived CFUs belong to 28 of 45 detectable bacterial families. Root- and leaf-derived CFUs each represent all four bacterial phyla typically associated with *A. thaliana* roots and leaves. Thus, most bacterial families that are reproducibly associated with *A. thaliana* roots and leaves have culturable members.

¹Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. ²Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland.

³Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. ⁴Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. ⁵Computational Biology of Infection Research, Helmholtz Center for Infection Research, 38124 Braunschweig, Germany. ⁶Max-von-Pettenkofer Institute, Ludwig Maximilian University, German Center for Infection Research (DZIF), partner site LMU Munich, 80336 Munich, Germany. ⁷German Center for Infection Research (DZIF), partner site Hannover-Braunschweig, 38124 Braunschweig, Germany. ⁸Max Planck Genome Center, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.

*These authors contributed equally to this work.

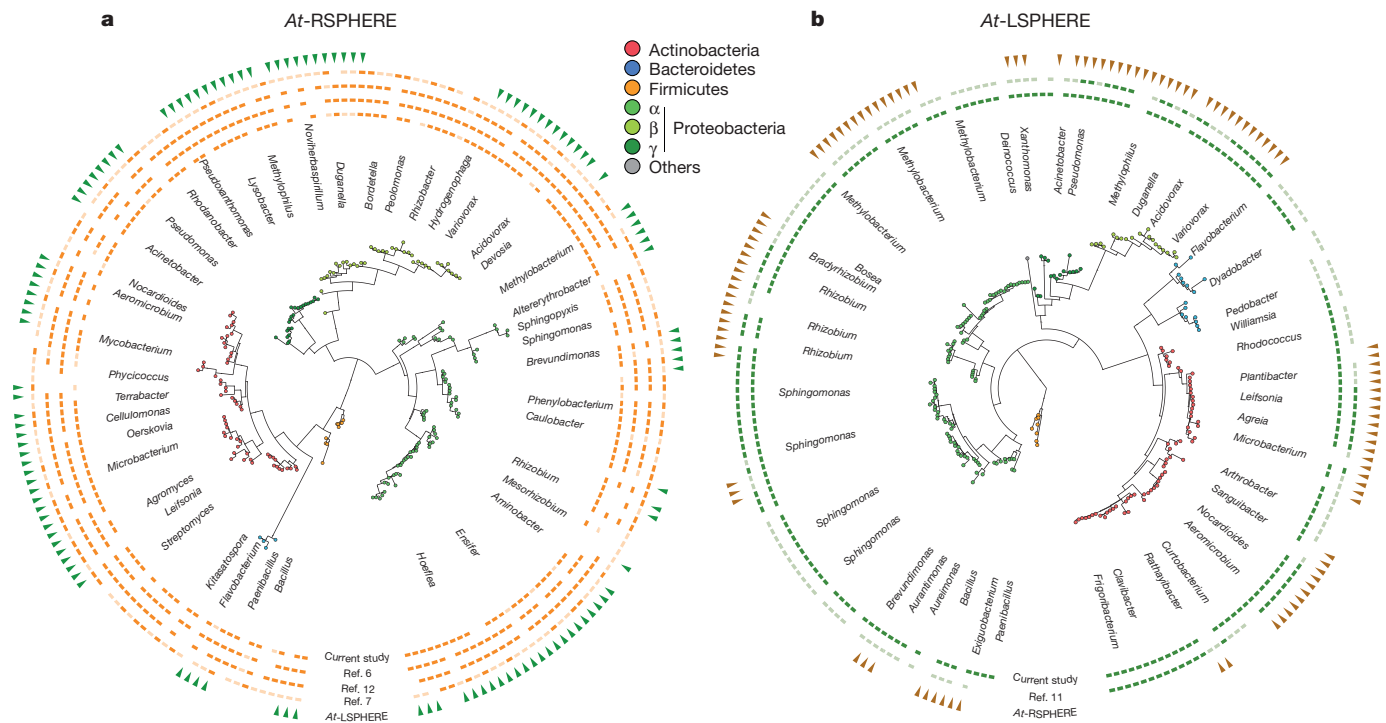


Figure 1 | Taxonomic overlap between *At*-RSPHERE and *At*-LSPHERE isolates and their representation in culture-independent microbiota profiling studies. **a, **b**, Phylogenetic trees of *At*-RSPHERE (**a**; $n = 206$ isolates) and *At*-LSPHERE (**b**; $n = 224$ isolates) bacteria. Their taxonomic overlap is shown in the outermost ring (green or brown triangles). **a**, Representation of *At*-RSPHERE bacteria in each of four indicated culture-independent profiling studies of the *A. thaliana* root microbiota;**

At-RSPHERE and *At*-LSPHERE culture collections

We selected from the aforementioned culture collections a taxonomically representative core set of bacterial strains after Sanger sequencing of a ≥ 550 bp fragment of the 16S rRNA gene and additional strain purification (Methods). To increase the intra-species genetic diversity of the culture collections, and because the quantitative contribution of a single isolate to its corresponding OTU cannot be estimated, we included bacterial strains sharing $\geq 97\%$ 16S rRNA gene sequence identity (widely used for bacterial species definition), but representing independent host colonization events, that is, recovered from different plant roots or leaves. In total we selected 206 root-derived isolates that comprise 28 bacterial families belonging to four phyla (designated *At*-RSPHERE) and 224 leaf-derived isolates that comprise 29 bacterial families belonging to five phyla (designated *At*-LSPHERE) (Extended Data Fig. 3a, b; Supplementary Data 1; Methods). Additionally, to represent abundant soil OTUs ($\geq 0.1\%$ RA) we selected 33 bacterial isolates encompassing eight bacterial families belonging to three phyla from unplanted Cologne soil (Extended Data Fig. 3c).

Notably, the majority of the *At*-RSPHERE isolates share $\geq 97\%$ 16S rRNA gene sequence identity matches with root-associated OTUs reported in four independent studies in which *A. thaliana* plants had been grown in Cologne soil^{6,12} or other European^{6,12} or US soils⁷ (inner four circles in Fig. 1a; Methods). Similarly, the bulk of *At*-LSPHERE isolates match leaf-derived OTUs detected in *A. thaliana* populations at the Tübingen/Zurich locations or US-grown plants (innermost two circles in Fig. 1b). This indicates that representatives of the majority of *At*-RSPHERE and *At*-LSPHERE members co-populate the corresponding *A. thaliana* organs in multiple tested environments, including two continents, Europe and North America.

Phylogenetic analysis based on 16S rRNA gene Sanger sequences revealed that 119 out of 206 *At*-RSPHERE isolates (58%) share $\geq 97\%$ sequence identity matches with corresponding 16S rRNA gene

root-associated OTUs with RAs $\geq 0.1\%$ (dark orange) or $\leq 0.1\%$ (light orange). **b**, Representation of *At*-LSPHERE bacteria in the two indicated culture-independent phyllosphere profiling studies; leaf-associated OTUs with RAs $\geq 0.1\%$ (dark green) or $< 0.1\%$ (light green). Taxonomic assignment and phylogenetic tree inference were based on partial 16S rRNA gene Sanger sequences.

fragments of *At*-LSPHERE members (outermost circle in Fig. 1a). Similarly, 108 out of 224 *At*-LSPHERE isolates (48%) share $\geq 97\%$ sequence identity matches with *At*-RSPHERE members (outermost circle in Fig. 1b). This extensive overlap both at the rank of bacterial genera and bacterial families (20 out of 38 detectable families) between leaf- and root-derived bacteria is notable because we collected leaf and root specimen from environments that are geographically widely separated (> 500 km) and is consistent with a previous report on leaf and root microbiota overlap in *V. vinifera*¹⁸. This overlap is corroborated by the corresponding culture-independent leaf and root community profiles (Extended Data Fig. 4). As essentially all *A. thaliana* root-associated bacteria are recruited from the surrounding soil biome^{6,7,12}, this raises the possibility that unplanted soil also defines the start inoculum for a substantial proportion of the leaf microbiota with subsequent selection for niche-adapted organisms.

Comparative genome analysis of the culture collections

To characterize the functional capabilities of the core culture collections we subjected each isolate to whole-genome sequencing and generated a total of 432 high-quality draft genomes (206 from leaf, 194 from root and 32 from soil; Supplementary Data 3). Taxonomic assignment of the whole-genome sequences confirmed that these isolates span a broad taxonomic range, belonging to 35 different bacterial families distributed across five phyla (Supplementary Data 4).

Based on the whole-genome taxonomic information, we grouped the isolates into family-level clusters. We found that clusters of genomes are characterized by a relatively large core-genome, with an average of 33.6% of the annotated proteins present in each member and a smaller fraction of singleton genes identified in only one genome per cluster (14.0%). Detailed analysis of phylogenetic diversity of each cluster revealed a substantial overlap between leaf, root and soil isolates (Supplementary Data 5). Many clusters showed no clear separation of

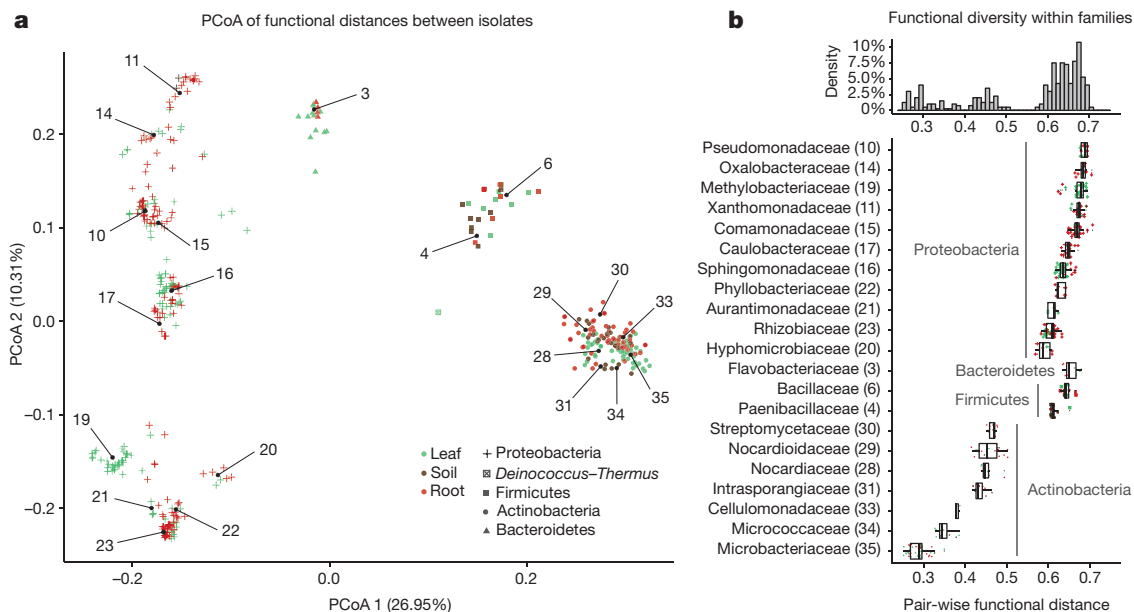


Figure 2 | Analysis of functional diversity between sequenced isolates. **a**, Principal coordinate analysis (PCoA) plot depicting functional distances between sequenced genomes ($n = 432$) based on the KEGG Orthology (KO) database annotation. Each point represents a genome. Colours represent the organ of isolation and shapes correspond to their taxonomy. Numbers inside the plot refer to bacterial families listed in **b**. **b**, Analysis of functional diversity within bacterial families as measured by pair-wise

isolates based on their ecological niche, suggesting shared core functions. However, other clusters contained isolates of one organ or showed clear separation among them, suggesting niche specialization within some clusters (Supplementary Data 5). We then examined the functional diversity between the sequenced isolates in order to determine whether the observed phylogenetic overlap corresponded with functional similarities between leaf and root isolates. Principal coordinates analysis (PCoA) of functional distances (Fig. 2a; Methods) revealed a clear clustering of genomes on the basis of their taxonomy, but only limited separation of genomes on the basis of their ecological compartment. Taken together, both phylogenetic and functional diversification of the genomes is strongly driven by their taxonomic affiliation and weakly by the ecological niche.

We examined the functional diversity within each bacterial family (Fig. 2b) in order to identify bacterial taxa with varying degrees of functional versatility. Families belonging to Actinobacteria show a lower functional diversity (average distance 0.37) compared to those belonging to Bacteroidetes, Firmicutes and especially Proteobacteria (0.65 average pair-wise distance), which exhibit a higher degree of within-family functional diversification, even though all family-level groups have a comparable degree of phylogenetic relatedness. Among these groups, Pseudomonadaceae, Oxalobacteraceae and Methylobacteriaceae members show the highest functional heterogeneity, compared to Microbacteriaceae strains, which we identified as the least functionally diverse family (Fig. 2b).

We searched for signatures of niche specialization at individual functional categories using enrichment analysis to identify functional categories over-represented in genomes from root and leaf or soil isolates (Fig. 3; Methods). Specifically, we found the category ‘carbohydrate metabolism’ to be enriched in the leaf and soil genomes compared to those isolated from roots (Mann–Whitney test, $P = 1.29 \times 10^{-7}$; Fig. 3b). We speculate that this differential enrichment could reflect the availability of simple carbon sources in roots through the process of root exudation (sugars, amino acids, aliphatic acids)^{21,22}, whereas bacteria associated with leaves or unplanted soil might rely on a more diverse repertoire of carbohydrate metabolism genes to access scarce

functional distances between genomes (bottom panel; $n = 432$). Higher pairwise distances between members of a family indicate a larger degree of functional diversity. Only families with at least five members are shown. The histogram (top panel) was calculated for the entire data set and the y-axis corresponds to the percentage of data points in each bin. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.

and complex organic carbon, for example, polysaccharides and leaf cuticular waxes. The category ‘xenobiotics biodegradation and catabolism’ is enriched in the root genomes with respect to those isolated from leaves ($P = 2.60 \times 10^{-11}$; Fig. 3b), which is consistent with previous evidence that genes for aromatic compound utilization are expressed in the rhizosphere²³. No single taxon is responsible for these significant differences, but this seems to be a general feature across the sequenced bacterial genomes of the respective ecological niche (Extended Data Figs 5 and 6). Interestingly, we observed the same trends of differential abundance of functional categories in *V. vinifera* root metagenome samples¹⁸ compared to their respective unplanted soil controls (Extended Data Fig. 7).

Together, these findings indicate a substantial overlap of functional capabilities in the genomes of the *Arabidopsis* leaf- and root-derived culture collections and differences at the level of individual functional categories that may reflect specialization of the leaf and root microbiota to their respective niche. Additional genomic signatures for niche-specific colonization are likely to be hidden in genes for which a functional annotation is currently unavailable (~57%).

Synthetic community colonization of germ-free plants

We colonized germ-free *A. thaliana* plants with synthetic communities (SynComs) consisting of bacterial isolates from our culture collections to assess their potential for host colonization in a gnotobiotic system containing calcined clay as inert soil substitute (Methods). To mimic the taxonomic diversity of leaf and root microbiota in natural environments we employed mainly two SynComs: ‘L’ comprising 218 leaf-derived bacteria and ‘R+S’ consisting of 188 members of which 158 are root-derived and 30 are soil-derived bacteria (Supplementary Data 6). Input SynComs were either inoculated directly before sowing of surface-sterilized seeds in calcined clay and/or spray-inoculated on leaves of three-week-old germ-free plants. For all defined communities we examined three independent SynCom preparations, each tested in three closed containers containing four plants. We employed 16S rRNA gene community profiling with a method validated for defined communities²⁴ to detect potential community shifts between input and output

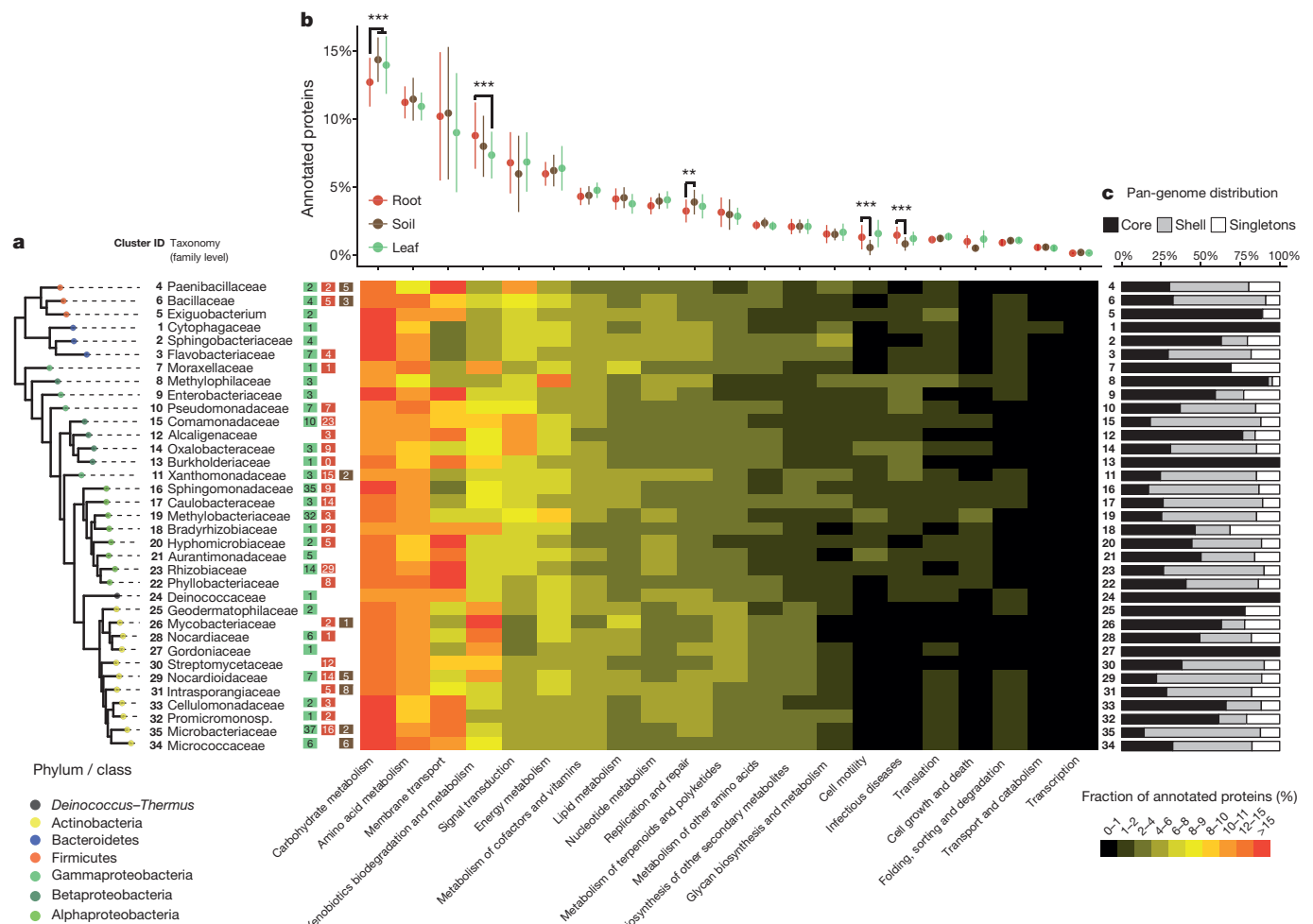


Figure 3 | Functional analysis of sequenced isolates. a, Phylogeny of family-level clusters of bacterial isolates. The tips of the tree are annotated, from left to right, with the cluster ID, taxonomic classification, followed by the number of sequenced isolates from leaf, root or soil that constitute each cluster. The heat map depicts the average percentage of annotated proteins of each cluster belonging to each functional category. **b**, Functional enrichment analysis between leaf ($n = 206$), root ($n = 194$)

and soil ($n = 32$) genomes. Points and bars correspond to the mean abundance and standard deviation of each functional category. P values were obtained using the non-parametric Mann–Whitney test corrected by the Bonferroni approach. **c**, Analysis of pan-genome distribution for each cluster of genomes, indicating the percentage of annotated proteins found in only one isolate (singletons), in more than one but not all (shell) or in all genomes within the cluster (core).

SynComs in samples of seven week-old roots, leaves, or unplanted clay. In this community analysis, ‘indicator OTUs’ either represent a single strain or a known group of isolates (Supplementary Data 6).

Upon application of the input R+S SynCom to clay (‘R+S in clay’) and co-cultivation with *A. thaliana* plants for seven weeks we retrieved reproducible R+S output communities from clay (without host), root, and leaf compartments (Supplementary Fig. 2). These output SynCom profiles were robust against a 75% reduction in RA of Proteobacteria compared to Actinobacteria, Bacteroidetes and Firmicutes in the input R+S SynCom (input ratios 1:1:1:1 or 1:1:1:0.25, respectively), which was confirmed by PCoA (Fig. 4a). PCoA also revealed distinct output communities in each of the three tested compartments (Fig. 4a; $P < 0.001$ Extended Data Fig. 8a, b). This indicates that a marked host-independent community change occurred in clay (without host) as well as host-dependent community shifts that are specific for leaves and roots. Next, we tested the ‘L SynCom of leaf-derived bacteria by spray inoculation on leaves of three week-old plants. After four weeks of L SynCom co-incubation with plants, output communities were detected in leaves and roots (Supplementary Fig. 3). PCoA revealed that these two output communities were different between each other, but robust against a 75% reduction in RA of input Proteobacteria (Fig. 4b; Supplementary Fig. 3; $P < 0.001$;

Extended Data Fig. 8c, d). The converging output communities despite varying RAs of input SynComs suggest that the communities have reached a steady state. These experiments also reveal that both R+S and L SynCom members not only colonize cognate host organs, but are capable of ectopic colonization of leaves and roots, which might be linked to the extensive species overlap of *A. thaliana* leaf and root microbiota in natural environments (Fig. 1a, b). Additionally, this provides experimental support for the hypothesis that a subset of leaf-colonizing bacteria originates from unplanted soil and raises the possibility for reciprocal bacterial colonization events between roots and leaves during and/or after the establishment of the respective microbiota, for example, by ascending migration of rhizobacteria from roots to leaves²⁵. Upon leaf spray application of SynComs, a small amount of leaf bacteria is likely to land on the clay surface and thereafter colonize roots, which is not fundamentally different from processes occurring in natural environments, for example, during rain showers and/or leaf dehiscence.

A comparison of rank abundance profiles between indicator OTUs for all root- and leaf-derived isolates and corresponding OTUs identified in the environmental root and leaf samples revealed similar trends at phylum, class and family levels (Extended Data Fig. 9). This validates the gnotobiotic plant system as a tool for microbiota reconstitution experiments.

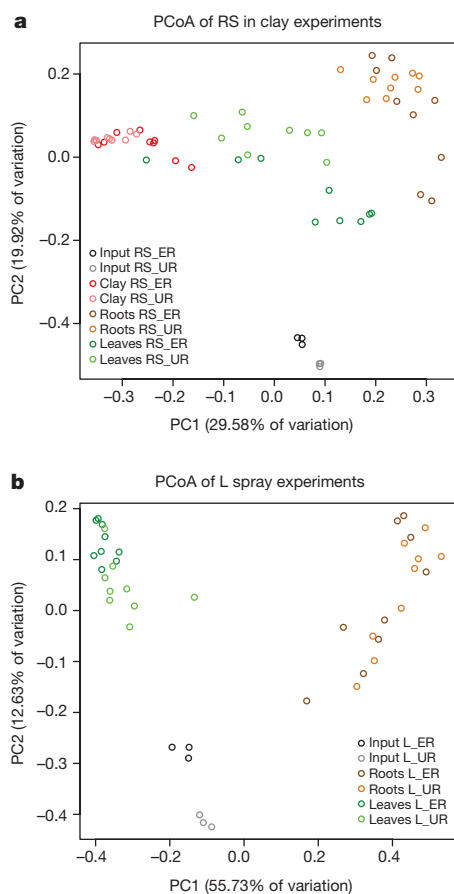


Figure 4 | SynCom colonization of germ-free *A. thaliana* plants. **a, b,** Principal coordinate analysis (PCoA) of Bray–Curtis distances of input and output SynCom profiles of RS in clay (**a**; $n = 60$) and L spray (**b**; $n = 42$) experiments. Each condition was tested with 6 independently prepared SynComs; each preparation was used for 3 independent inoculations. L, leaf-derived strains; RS, root- and soil-derived strains; ER, equal strain ratio; UR, unequal strain ratio.

Niche-specific microbiota establishment with SynComs

The species overlap between root and leaf microbiota and their corresponding culture collections (Fig. 1a, b; Extended Data Fig. 4) prompted us to test whether R+S and L SynComs equally contribute to root and leaf microbiota establishment. Both SynComs were pooled and inoculated in clay together with surface-sterilized *A. thaliana* seeds (designated ‘RSL in clay’, Fig. 5a). We also tested whether a preformed root-associated community can interfere with leaf-associated community establishment. After three weeks of co-cultivation, half of the plants grown with the ‘RSL in clay’ SynCom were treated by leaf-spray inoculation with the L SynCom supplemented with 15 root-derived strains (designated ‘RSL in clay & L+15R spray’). Plant organ-specific output communities were determined after a further four weeks of co-incubation. We also inoculated the L SynCom alone in clay and determined output SynComs (designated ‘L in clay’, Fig. 5a).

We found significant differences between leaf-associated output communities of the ‘RSL in clay’ and ‘RS in clay’ experiments (Fig. 5b; $P < 0.001$, Extended Data Fig. 8f; Supplementary Figs 2 and 4) and that the output community on leaves after ‘L in clay’ inoculation is similar to the leaf outputs of ‘RSL in clay’ inoculation (Fig. 5b; $P < 0.001$, Extended Data Fig. 8f; Supplementary Figs 4 and 5), indicating that in this comparison the leaf-derived SynCom has a stronger influence on leaf microbiota structure than root- and soil-derived bacteria. However, both ‘RSL in clay’ and ‘L in clay’ leaf outputs are significantly different from the leaf output of the ‘L spray’ experiment (Fig. 5b; $P < 0.001$, Extended Data Fig. 8e; Supplementary Figs 3–5), showing that many leaf-derived isolates do not successfully colonize leaves when only

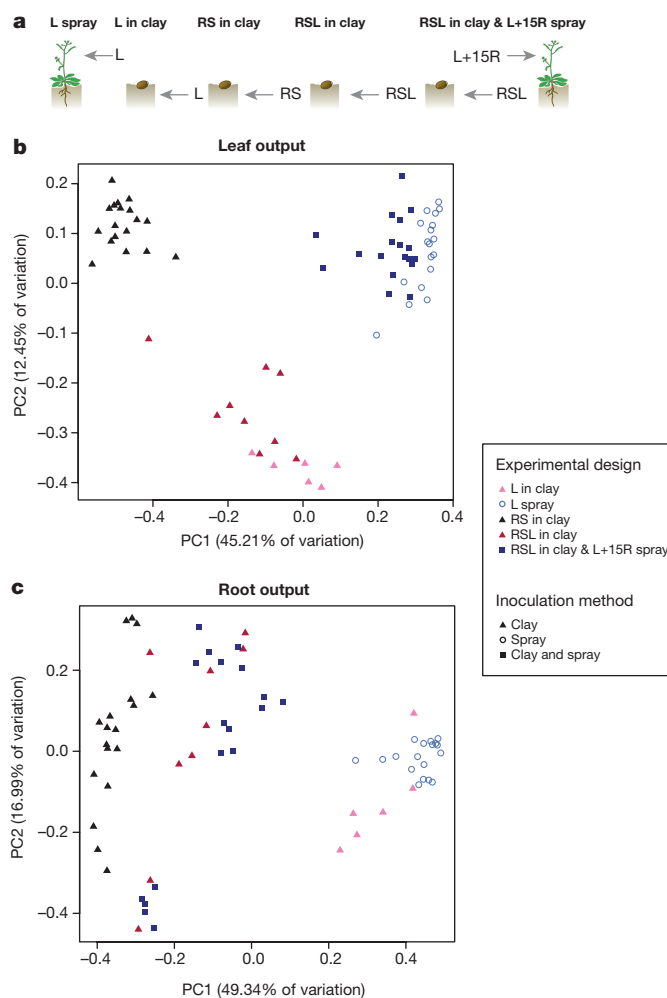


Figure 5 | SynCom competition supports host-organ-specific community assemblies. **a,** Pictograms illustrating ‘L spray’, ‘L in clay’, ‘RS in clay’, ‘RSL in clay’, and ‘RSL in clay & L+15R spray’ SynCom experiments. **b, c,** PCoA of Bray–Curtis distances of leaf (**b**; $n = 69$) and root (**c**; $n = 69$) outputs of the five experiments illustrated in **a**. R, root-derived isolates; S, soil-derived isolates; L, leaf-derived isolates. L in clay was tested with 6 independently prepared SynComs; RSL in clay experiment was tested with 3 independently prepared SynComs, each used for 3 independent inoculations. All other experiments were tested with 6 independently prepared SynComs and each preparation was used for 3 independent inoculations.

inoculated in the clay environment. For example, of the top 16 genera a total of three are grossly underrepresented in leaf outputs of the ‘RSL in clay’ compared to the ‘RSL in clay & L+15R spray’ experiment (*Chryseobacterium*, *Sphingomonas* and *Variovorax*; Supplementary Fig. 6) and these three genera are abundant in the natural leaf microbiota (Extended Data Fig. 4). Finally, leaf outputs were strikingly similar between ‘RSL in clay & L+15R spray’ and ‘L spray’ only experiments (Fig. 5b; Supplementary Figs 3 and 7), indicating that the L+15R SynCom, leaf spray-inoculated three weeks after RSL application to clay, can displace the RSL leaf output. Collectively, these results support the hypothesis that leaf microbiota establishment benefits from air- and soil-borne inoculations^{8,17}, although we note that our single application of bacteria to leaves does not mimic the continuous exposure of plant leaves to airborne microorganisms in nature.

A comparison of the root-associated community outputs of the experiments described above revealed that the ‘RSL in clay’ experiment is more similar to root outputs of the ‘RS in clay’ than ‘L in clay’ experiments (Fig. 5c; $P < 0.001$ Extended Data Fig. 8g), suggesting that the root- and soil-derived SynCom has a stronger influence on root

microbiota structure than the leaf-derived SynCom. In this experiment the fractional contribution of root-specific indicator OTUs increases in the output, but decreases for leaf-specific indicator OTUs, relative to their input, pointing to a potential adaptation of root-derived bacteria for root colonization (Extended Data Fig. 10a; Mann–Whitney; $P < 0.05$). This is further supported by the observation that in the ‘RSL in clay’ experiment root colonization rates for root-specific indicator OTUs are higher compared to those specific for leaves when applying a 0.1% relative abundance threshold in at least one biological replicate (69% and 33%, respectively). Taken together, this suggests that root-derived bacteria are better adapted to colonize their cognate host niche than leaf-derived bacteria. Further comparisons of the root-associated output communities of the ‘L in clay’ and ‘L spray’ experiments (Fig. 5c; Supplementary Figs 3 and 5) revealed similar community composition, indicating convergence of ectopic root-associated community outputs despite different inoculation time points or sites of application. Additional reciprocal transplantation experiments using a ‘R’ (root strains only) SynCom either applied to clay (‘R in clay’) or by spray inoculation (‘R spray’) confirmed the convergence of ectopic community outputs also for root-derived bacteria on leaves (Extended Data Fig. 10 b, c; Supplementary Figs 8 and 9). Convergence of ectopic SynCom outputs is consistent with the hypothesis that a subset of leaf and root colonizing bacteria has the potential to relocate between leaves and roots.

Conclusions

By employing systematic bacterial isolation approaches, we established expandable culture collections of the *A. thaliana* leaf- and root-associated microbiota, which capture the majority of the species found reproducibly in their respective natural communities ($\geq 0.1\%$ relative abundance). The sequenced bacterial genomes as well as any future updates are available at <http://www.at-sphere.com>. These resources together with the remarkable reproducibility of the gnotobiotic reconstitution system enable future studies on bacterial community establishment and functions under laboratory conditions.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 June; accepted 9 November 2015.

Published online 2 December 2015.

- Rosenberg, E. & Xilber-Rosenberg, I. *The Hologenome Concept: Human, Animal and Plant Microbiota* (Springer, 2013).
- Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Rev. Microbiol.* **9**, 279–290 (2011).
- Berendsen, R. L., Pieterse, C. M. & Bakker, P. A. The rhizosphere microbiome and plant health. *Trends Plant Sci.* **17**, 478–486 (2012).
- Subramanian, S. *et al.* Cultivating healthy growth and nutrition through the gut microbiota. *Cell* **161**, 36–48 (2015).
- Delmotte, N. *et al.* Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc. Natl Acad. Sci. USA* **106**, 16428–16433 (2009).
- Bulgarelli, D. *et al.* Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* **488**, 91–95 (2012).
- Lundberg, D. S. *et al.* Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**, 86–90 (2012).
- Vorholt, J. A. Microbial life in the phyllosphere. *Nature Rev. Microbiol.* **10**, 828–840 (2012).
- Bodenhausen, N., Horton, M. W. & Bergelson, J. Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PLoS One* **8**, e56329 (2013).
- Guttman, D. S., McHardy, A. C. & Schulze-Lefert, P. Microbial genome-enabled insights into plant-microorganism interactions. *Nature Rev. Genet.* **15**, 797–813 (2014).
- Horton, M. W. *et al.* Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. *Nat. Commun.* **5**, 5320 (2014).

- Schlaeppli, K., Dombrowski, N., Oter, R. G., Ver Loren van Themaat, E. & Schulze-Lefert, P. Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. *Proc. Natl Acad. Sci. USA* **111**, 585–592 (2014).
- Edwards, J. *et al.* Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc. Natl Acad. Sci. USA* **112**, E911–E920 (2015).
- Hacquard, S. *et al.* Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* **17**, 603–616 (2015).
- Bulgarelli, D. *et al.* Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* **17**, 392–403 (2015).
- Lebeis, S. L. *et al.* Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science* **349**, 860–864 (2015).
- Maignien, L., DeForce, E. A., Chafee, M. E., Eren, A. M. & Simmons, S. L. Ecological succession and stochastic variation in the assembly of *Arabidopsis thaliana* phyllosphere communities. *MBio* **5**, e00682–e13 (2014).
- Zarraonaindia, I. *et al.* The soil microbiome influences grapevine-associated microbiota. *MBio* **6**, e02527–14 (2015).
- Lebeis, S. L., Rott, M., Dangl, J. L. & Schulze-Lefert, P. Culturing a plant microbiome community at the cross-Rhodes. *New Phytol.* **196**, 341–344 (2012).
- Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108**, 6252–6257 (2011).
- Faure, D., Vereecke, D. & Leveau, J. J. Molecular communication in the rhizosphere. *Plant Soil* **321**, 279–303 (2009).
- Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S. & Vivanco, J. M. The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu. Rev. Plant Biol.* **57**, 233–266 (2006).
- Ramachandran, V. K., East, A. K., Karunakaran, R., Downie, J. A. & Poole, P. S. Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol.* **12**, R106 (2011).
- Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10**, 996–998 (2013).
- Chi, F. *et al.* Ascending migration of endophytic rhizobia, from roots to leaves, inside rice plants and assessment of benefits to rice growth physiology. *Appl. Environ. Microbiol.* **71**, 7271–7278 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank D. Lundberg, S. Lebeis, S. Herrera-Paredes, S. Biswas and J. Dangl for sharing the calcined clay utilization protocol before publication; M. Kisielow of the ETH Zurich Flow Cytometry Core Facility for help with bacterial cell sorting as well as M. Baltisberger, D. Jolic and D. Weigel for their help in finding natural *Arabidopsis* populations; E. Kemen and M. Agler for sharing the Illumina Mi-Seq protocol for profiling of defined communities before publication and A. Szczyrba for his advice with the genome assembly. This work was supported by funds to P.S.-L. from the Max Planck Society, a European Research Council advanced grant (ROOTMICROBIOTA), the ‘Cluster of Excellence on Plant Sciences’ program funded by the Deutsche Forschungsgemeinschaft, the German Center for Infection Research (DZIF), by funds to J.A.V. from ETH Zurich (ETH Research Grant ETH-41 14-2), a grant from the Swiss National Research Foundation (310030B_152835), and a European Research Council advanced grant (PhyMo).

Author Contributions J.A.V. and P.S.-L. initiated, coordinated and supervised the project. Y.B., M.R., N.D. and S.S. isolated root and soil bacteria strains. Y.B. collected root material and performed culture-independent community profiling. D.B.M., E.P. and M.R.-E. collected environmental leaf material. D.B.M. and E.P. isolated leaf strains and performed culture-independent community profiling. G.S. and R.G.-O. analysed culture-independent 16S rRNA amplicon sequencing data. Y.B., D.B.M. isolated DNA and prepared samples for genome sequencing. R.G.-O., P.C.M., B.H. and A.C.M. organized the genome sequencing data. R.G.-O. assembled and annotated draft genomes and performed comparative genome analyses. Y.B. and D.B.M. performed recolonization experiments; G.S. and R.G.-O. analysed the recolonization data. Y.B., D.B.M., R.G.-O., J.A.V. and P.S.-L. wrote the manuscript.

Author Information Sequencing reads (454 16S rRNA, MiSeq 16S rRNA and WGS HiSeq reads) have been deposited in the European Nucleotide Archive (ENA) under accession numbers PRJEB11545, PRJEB11583 and PRJEB11584, and genome assemblies and annotations corresponding to the leaf, root and soil culture collections have been deposited in the BioProject database under accession numbers PRJNA297956, PRJNA297942 and PRJNA298127. Isolates have been deposited at the Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures (<https://www.dsmz.de/>). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence should be addressed to J.A.V. (jvorholt@ethz.ch) or P.S.-L. (schlef@mpipz.mpg.de).

METHODS

Sampling of *A. thaliana* plants and isolation of root-, leaf- and soil-derived bacteria.

A. thaliana plants were either harvested from natural populations or grown in different natural soils and used for bacterial isolations by colony picking, limiting dilution or bacterial cell sorting as well as 16S rRNA gene-based community profiling. To obtain a library of representative root colonizing bacteria, *A. thaliana* plants were grown in different soils (50.958 N, 6.856 E, Cologne, Germany; 52.416 N, 12.968 E, Golm, Germany; 50.982 N, 6.827 E, Widdersdorf, Germany; 47.941 N, 04.012 W, Saint-Evarzec, France; 48.725 N, 3.989 W, Roscoff, France) and harvested before bolting. Briefly, *Arabidopsis* roots were washed twice in washing buffers (10 mM MgCl₂ for limiting dilution and PBS for colony picking⁶) on a shaking platform for 20 min at 180 rpm and then homogenized twice by Precellys24 tissue lyser (Bertin Technologies) using 3 mM metal beads at 5,600 rpm for 30 s. Homogenates were diluted and used for isolation approaches on several bacterial growth media (Supplementary Data 7). For isolations based on colony picking, diluted cell suspensions were plated on solidified media and incubated, before isolates of plates containing less than 20 colony-forming units (CFUs) were picked after a maximum of two weeks of incubation. For limiting dilution, homogenized roots from each root pool were sedimented for 15 min and the supernatant was empirically diluted, distributed and cultivated in 96-well microtitre plates²⁰. In parallel to the isolation of root-derived bacteria, roots of plants grown in Cologne soil were harvested and used to assess bacterial diversity by culture-independent 16S rRNA gene sequencing. Additionally, soil-derived bacteria were extracted from unplanted Cologne soil by washing soil with PBS buffer, supplemented with 0.02% Silwet L-77 and subjected to bacterial isolation as well as 16S rRNA gene community profiling. For the isolation of representative phyllosphere strains, naturally grown *Arabidopsis* plants were collected at eight different sites in southern Germany and Switzerland (six main sampling sites used for bacterial isolations and community profiling: 47.4090306 N, 8.470169444 E, Hoegg, Switzerland; 47.474825 N, 8.305008333 E, Baden, Switzerland; 47.4816806 N, 8.217547222 E, Brugg, Switzerland; 48.5560194 N, 9.134944444 E, Farm, Tuebingen, Germany; 48.5989861 N, 9.201655556 E, Haeslach, Germany; 48.602682 N, 9.213247258 E, Haeslach, Germany; and two additional sites only used for bacterial isolation: 47.4074722 N, 8.50825 E, Zurich, Switzerland; 47.4227222 N, 8.548666667 E, Seebach, Switzerland) during spring and autumn of 2013 and used for bacterial isolations as well as 16S rRNA gene profiling. Leaf-colonizing bacteria of individual leaves were washed off by alternating steps of intense mixing and sonication. The suspension was subsequently filtered (CellTrics filters, 10 µm, Partec GmbH, Görlitz, Germany) in order to remove remaining plant or debris particles as well as cell aggregates and applied to cell sorting on a BD FACS Aria III (BD Biosciences) as well as to plating on different media (Supplementary Data 1 and 7). All isolates were subsequently stored in 30% or 40% glycerol at -80 °C.

Culture-independent bacterial 16S rRNA gene profiling of *A. thaliana* leaf, root and corresponding soil samples.

Parts of *A. thaliana* leaves, roots and corresponding unplanted soil samples used for bacterial isolation were also processed for bacterial 16S rRNA gene community profiling using 454 pyrosequencing. Frozen root and corresponding soil samples were homogenized, DNA was extracted with Lysing Matrix E (MP Biomedicals) at 5,600 rpm for 30 s, and DNA was extracted from all samples using the FastDNA SPIN Kit for soil (MP Biomedicals) according to the manufacturer's instructions. Lyophilized leaf samples were transferred into 2 ml microcentrifuge tubes containing one metal bead and subsequently homogenized twice for 2 min at 25 Hz using a Retsch tissue lyser (Retsch, Haan, Germany). Homogenized leaf material was resuspended in lysis buffer of the MO BIO PowerSoil DNA isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA, USA), transferred into lysis tubes, provided by the supplier, and DNA extraction was performed following the manufacturer's protocol. DNA concentrations were measured by PicoGreen dsDNA Assay Kit (Life technologies), and subsequently diluted to 3.5 ng µl⁻¹. Bacterial 16S rRNA genes were subsequently amplified⁶ using primers targeting the variable regions V5-V7 (799F²⁶ and 1193R⁶, Supplementary Data 7). Each sample was amplified in triplicate by two independent PCR mixtures (a total of 6 replicates per sample plus respective no template controls). PCR products of triplicate were subsequently combined, purified and subjected to 454 sequencing. Obtained sequences were demultiplexed as well as quality and length filtered (average quality score ≥25, minimum length 319 bp with no ambiguous bases and no errors in the barcode sequences allowed)²⁷. High-quality sequences were subsequently processed using the UPARSE²⁴ pipeline and OTUs were taxonomically classified using the Greengenes database²⁸ and the PyNASt²⁹ method.

High-throughput identification of leaf-, root- and soil-derived bacterial isolates by 454 pyrosequencing.

We adopted a two-step barcoded PCR protocol²⁰ in combination with 454 pyrosequencing to define V5-V8 sequences of bacterial 16S rRNA genes of all leaf, root- and soil-derived bacterial (Supplementary Fig. 1). DNA of isolates was extracted by lysis of 6 µl of bacterial cultures in 10 µl of buffer I containing 25 mM NaOH, 0.2 mM EDTA, pH 12 at 95 °C for 30 min, before the

pH value was lowered by addition of 10 µl of buffer II containing 40 mM Tris-HCl at pH 7.5. Position and taxonomy of isolates in 96-well microtitre plates were indexed by a two-step PCR protocol using the degenerate primers 799F and 1392R containing well- and plate-specific barcodes (Supplementary Data 7) to amplify the variable regions V5 to V8. During the first step of PCR amplification, DNA from 1.5 µl of lysed cells was amplified using 2 U DSF-Taq DNA polymerase, 1 × complete buffer (both Bioron GmbH), 0.2 mM dNTPs (Life technologies), 0.2 µM of 1 of 96 barcoded forward primer with a 18-bp linker sequence (for example, A1_454_799F1_PCR1_wells; Supplementary Data 7) and 0.2 µM reverse primer (454B_1392R) in a 25 µl reaction. PCR amplification was performed under the following conditions: DNA was initially denaturated at 95 °C for 2 min, followed by 40 cycles of 95 °C for 30 s, 50 °C for 30 s and 72 °C for 45 s, and a final elongation step at 72 °C for 10 min. PCR products of each 96-well microtitre plate were combined and subsequently purified in a two-step procedure using the Agencourt AMPure XP Kit (Beckman Coulter GmbH, Krefeld, Germany) first, then DNA fragments were excised from a 1% agarose gel using the QIAquick Gel Extraction Kit (Qiagen). DNA concentration was measured by Nanodrop and diluted to 1 ng µl⁻¹.

During the second PCR step, 1 ng of pooled DNA (each pool represents one 96-well microtitre plate) was amplified by 1.25 U PrimeSTAR HS DNA Polymerase, 1 × PrimeSTAR Buffer (both TaKaRa Bio S.A.S, Saint-Germain-en-Laye, France), 0.2 mM dNTPs (Thermo Fisher Scientific Inc.), 0.2 µM of 1 of 96 barcoded forward primer targeting the 18-bp linker sequence (for example, P1_454_PCR2; Supplementary Data 7) and 0.2 µM reverse primer (454B_1392R) in a 50 µl reaction. The PCR cycling conditions were as follows. First, denaturation at 98 °C for 30 s, followed by 25 cycles of 98 °C for 10 s, 58 °C for 15 s and 72 °C for 30 s, and a final elongation at 72 °C for 5 min. PCR products were purified using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and QIAquick Gel Extraction Kit (Qiagen) as described for the purification of first step PCR amplicons. DNA concentration was determined by PicoGreen dsDNA Assay Kit (Life technologies) and samples were pooled in equal amounts. The final PCR product libraries were sequenced on the Roche 454 Genome Sequencer GS FLX+. Each sequence contained a plate-barcode, a well-barcode and V5-V8 sequences.

The sequences were quality filtered, demultiplexed according to well and plate identifiers²⁷. OTUs were clustered at 97% similarity by UPARSE algorithm²⁴. A nucleotide-based blast (v. 2.2.29) was used to align representative sequences of isolated OTUs to culture-independent OTUs and only hits ≥97% sequence identity covering at least 99% of the length of the sequences were considered.

Preparation of *A. thaliana* leaf (At-LSPHERE), root (At-RSPHERE) and soil bacterial culture collections. Based on representative sequences of OTUs from this as well as previously published culture-independent community analysis, bacterial CFUs in the culture collections with ≥97% 16S rRNA gene identity to root-, leaf- and soil-derived OTUs were purified by three consecutive platings on the respective solidified media before an individual colony was used to inoculate liquid cultures. These liquid cultures were used for validation by Sanger sequencing with both 799F and 1392R primers as well as for the preparation of glycerol stocks for the culture collections and for the extraction of genomic DNA for whole-genome sequencing. A total of 21 leaf-derived strains, previously described as phyllosphere bacteria^{8,9}, were added to the At-LSPHERE collection although these were undetectable in the present culture-independent leaf community profiling.

Preparation of bacterial genomic DNA for whole-genome sequencing. To obtain high molecular weight genomic DNA of bacterial isolates in our culture collections, we used a modified DNA precipitation protocol and the Agencourt AMPure XP Kit (Beckman Coulter GmbH). For each bacterial liquid culture, cells were collected by centrifugation at 3,220g for 15 min, the supernatant removed and cells were resuspended in 5 ml SET buffer containing 75 mM NaCl, 25 mM EDTA, 20 mM Tris/HCl at pH 7.5. A total of 20 µl lysozyme solution (50 mg ml⁻¹, Sigma) was added before the mixture was incubated for 30 min at 37 °C. Subsequently, 100 µl 20 mg ml⁻¹ proteinase K (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany) and 10% SDS (Sigma-Aldrich Chemie GmbH) were added, mixed, and incubated by shaking every 15 min at 55 °C for 1 h. If bacterial cells were insufficiently lysed, remaining cells were collected at 3,220g for 10 min and homogenized using the Precellys24 tissue lyser in combination with lysing matrix E tubes (MP Biomedicals) at 6,300 rpm for 30 s. After cell lysis, 2 ml 5 M NaCl and 5 ml chloroform were added and mixed by inversion for 30 min at room temperature. After centrifugation at 3,220g for 15 min, 6 ml supernatant were transferred into fresh falcon tubes and 3.6 ml isopropanol were added and gently mixed. After precipitation at 4 °C for 30 min, genomic DNA was collected at 3,220g for 5 min, washed once with 1 ml 70% (v/v) ethanol, dried for 15 min at room temperature and finally dissolved in 250 µl elution buffer (Qiagen). 2 µl 4 mg ml⁻¹ RNase A (Sigma-Aldrich Chemie GmbH) was added to bacterial genomic DNA solution and incubated over night at 4 °C.

The genomic DNA was subsequently purified using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and analysed by agarose gel (1% (w/v))

electrophoresis. Concentrations were estimated based on loaded Lambda DNA Marker (GeneRuler 1kb Plus, Thermo Scientific) and approximately 1 µg of genomic DNA was transferred into micro TUBE Snap-Cap AFA Fibre vials (Covaris Inc., Woburn, MA, USA). DNA was sheared into 350 bp fragments by two consecutive cycles of 30 s (duty cycle: 10%, intensity: 4, cycle/burst: 200) on a Covaris S2 machine (Covaris, Inc.). The Illumina sequencing libraries were prepared according to the manual of NEBNext Ultra UltraTM DNA Library Prep Kit for Illumina (New England Biolabs, USA). Quality and quantity was assessed at all steps by capillary electrophoresis (Agilent Bioanalyser and Agilent TapeStation). Finally libraries were quantified by fluorometry, immobilized and processed onto a flow cell with a cBot (Illumina Inc., USA) followed by sequencing-by-synthesis with TruSeq v3 chemistry on a HiSeq2500 (Illumina Inc., USA).

Genome assembly and annotation. Paired-end Illumina reads were subjected to quality and length trimming using Trimmomatic v. 0.33³⁰ and assembled using two independent methods (A5³¹ and SOAPdenovo³² v. 20.1). In each case, the assembly with the smaller number of scaffolds was selected. Detailed assembly statistics for each sequenced isolate can be found in Supplementary Data 3 and 4. Identification of putative protein-encoding genes and annotation of the genomes were performed using GLIMMER v. 3.02³³. Functional annotation of genes was conducted using Prokka v. 1.11³⁴ and the SEED subsystems approach using the RAST server API³⁵. Additionally, annotation of KEGG Orthologue (KO) groups was performed by first generating HMM models for each KO in the database^{36,37} the HMMER toolkit (v. 3.1b2)³⁸. Next, we employed the HMM models to search all predicted ORFs using the hmsearch tool, with an *E* value threshold of 10×10^{-5} . Only hits covering at least 70% of the protein sequence were retained and for each gene and the match with the lowest *E* value was selected.

Analyses of phylogenetic diversity within sequenced isolates. Each proteome was searched for the presence of the 31 well-conserved, single-copy, bacterial AMPHORA genes³⁹, designed for the purpose of high-resolution phylogeny reconstruction of genomes. Subsequently, a concatenated alignment of these marker genes was performed using Clustal Omega⁴⁰ v. 1.2.1. Based on this multiple sequence alignment, a species tree was inferred using FastTree⁴¹ v. 2.1, a maximum likelihood tool for phylogeny inference. Whole-genome taxonomic classification of sequenced isolates was conducted using taxator-tk⁴², a homology-based tool for accurate classification of sequences. Analyses of phylogenetic diversity were performed independently for each cluster based on pairwise tree distances between all isolates (Supplementary Data 5).

Analyses of functional diversity between sequenced isolates. Analyses of functional diversity between sequenced isolates were conducted by generating, for each genome in the data set, a profile of presence/absence of each KO group (or phyletic pattern). Subsequently, a distance measure based on the Pearson correlation of each pair of phyletic patterns was calculated, which allowed us to embed each genome as a data point in a metric space. PCoA was performed on this space of functional distances using custom scripts written in R. Pairwise functional distances within each family-level cluster was performed by calculating the average distance between all pairs of genomes belonging to each cluster. Finally, we calculated RAs of each functional category based on the percentage of annotated KO terms assigned to each category. Enrichment tests were performed to identify differentially abundant categories between groups of genomes based on their origin (root versus leaf and root versus soil) using the non-parametric Mann–Whitney Test (MWT). *P* values were corrected for multiple testing using the Bonferroni method, with a significance threshold $\alpha = 0.05$.

Recolonization experiments of leaf-, root- and soil-derived bacteria on *Arabidopsis*. Calcined clay¹⁶, an inert soil substitute, was washed with water, sterilized twice by autoclaving and heat-incubated until being completely dehydrated. *A. thaliana* Col-0 seeds were surface-sterilized with ethanol and stratified overnight at 4°C. Leaf-, root- and soil-derived bacteria of the culture collections were cultivated in 96-deep-well plates and subsequently pooled (in equal or unequal ratios) in order to prepare synthetic bacterial communities (SynComs) for inoculations below the carrying capacity of leaves and roots^{43,44}. To inoculate SynComs into the calcined clay matrix, OD₆₀₀ was adjusted to 0.5 and 1 ml (~2.75 × 10⁸ cells) was added to 70 ml 0.5 × MS media (pH 7; including vitamins, without sucrose), and mixed with 100 g calcined clay in Magenta boxes (~2.75 × 10⁶ cells per gr calcined clay), directly before sowing of surface-sterilized seeds. Plants were grown at 22°C, 11 h light, and 54% humidity. Alive cell counts (CFUs) of root-associated bacteria by serial dilutions of root homogenates after seven weeks of co-incubation were $1.4 \times 10^8 \pm 8.4 \times 10^7$ cells per gram root tissue. For leaf spray-inoculation of *A. thaliana* plants, bacterial SynComs were prepared as described above and adjusted to OD₆₀₀ 0.2, before the solution was diluted tenfold and 170 µl (~1.87 × 10⁶ cells) were sprayed into each magenta box containing four three-week-old plants using a TLC chromatographic reagent sprayer (BS124.000, Biostep GmbH, Jahnsdorf, Germany). The average volume per spraying event was determined by spraying repeatedly into 50 ml tubes and weighing before and after. All plants and

corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. All plants and corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. During harvest, leaves and roots of individual plants were carefully separated using sterilized tweezers and scissors to avoid cross-contamination and processed separately thereafter. All leaves being obviously contaminated with clay particles or touching the ground were carefully removed and omitted from further processing. Remaining aerial parts of four plants collected from one magenta box were combined and transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at –80°C until used for DNA extraction. Roots from one Magenta box were pooled, washed twice in 5 ml PBS at 180 rpm for 20 min, dried on sterilized Whatman glass microfibre filters (GE Healthcare Life Sciences), transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at –80°C until further processing. The corresponding unplanted clay samples were washed in 100 ml PBS supplemented with 0.02% Silwet L-77 at 180 rpm for 10 min, before particles were allowed to settle down for 5 min. The supernatant was collected by centrifugation at 3,220g for 15 min. The pellet was subsequently resuspended in 1 ml water, transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at –80°C.

To prepare DNA for bacterial 16S rRNA gene-based community analysis, all samples were homogenized twice by Precellys24 tissue lyser (Bertin Technologies), DNA was extracted and concentrations were measured by PicoGreen dsDNA Assay Kit (Life technologies), before bacterial 16S rRNA genes were amplified by degenerate PCR primers (799F and 1193R) targeting the variable regions V5–V7 (Supplementary Data 7). Each sample was amplified in triplicate (plus respective no template control) in 25 µl reaction volume containing 2 U DFS-Taq DNA polymerase, 1 × incomplete buffer (both Bioron GmbH, Ludwigshafen, Germany), 2 mM MgCl₂, 0.3% BSA, 0.2 mM dNTPs (Life technologies GmbH, Darmstadt, Germany), 0.3 µM forward and reverse primer and 10 ng of template DNA. After an initial denaturation step at 94°C for 2 min, the targeted region was amplified by 25 cycles of 94°C for 30 s, 55°C for 30 s and 72°C for 60 s, followed by a final elongation step of 5 min at 72°C. The three independent PCR reactions were pooled and the remaining primers and nucleotides were removed by addition of 20 U exonuclease I and 5 U Antarctic phosphatase (both New England BioLabs GmbH, Frankfurt, Germany) and incubated for 30 min at 37°C in the corresponding 1 × Antarctic phosphatase buffer. Enzymes were heat-inactivated and the digested mixture was used as template for the 2nd step PCR using the Illumina compatible primers B5-F and 1 of 96 differentially barcoded reverse primers (B5-1 to B5-96, Supplementary Data 7). All samples were amplified in triplicate for 10 cycles using identical conditions of the first-step PCR. Technical replicates of each sample were combined, run on a 1.5% (w/v) agarose gel and the bacterial 16S rRNA gene amplicons were extracted using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. DNA concentration was subsequently measured using the PicoGreen dsDNA Assay Kit (Life technologies) and 100 ng of each sample were combined. Final amplicon libraries were cleaned twice using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and subjected to sequencing on the Illumina MiSeq platform using an MiSeq Reagent kit v3 following the 2 × 350 bp paired-end sequencing protocol (Illumina Inc. USA).

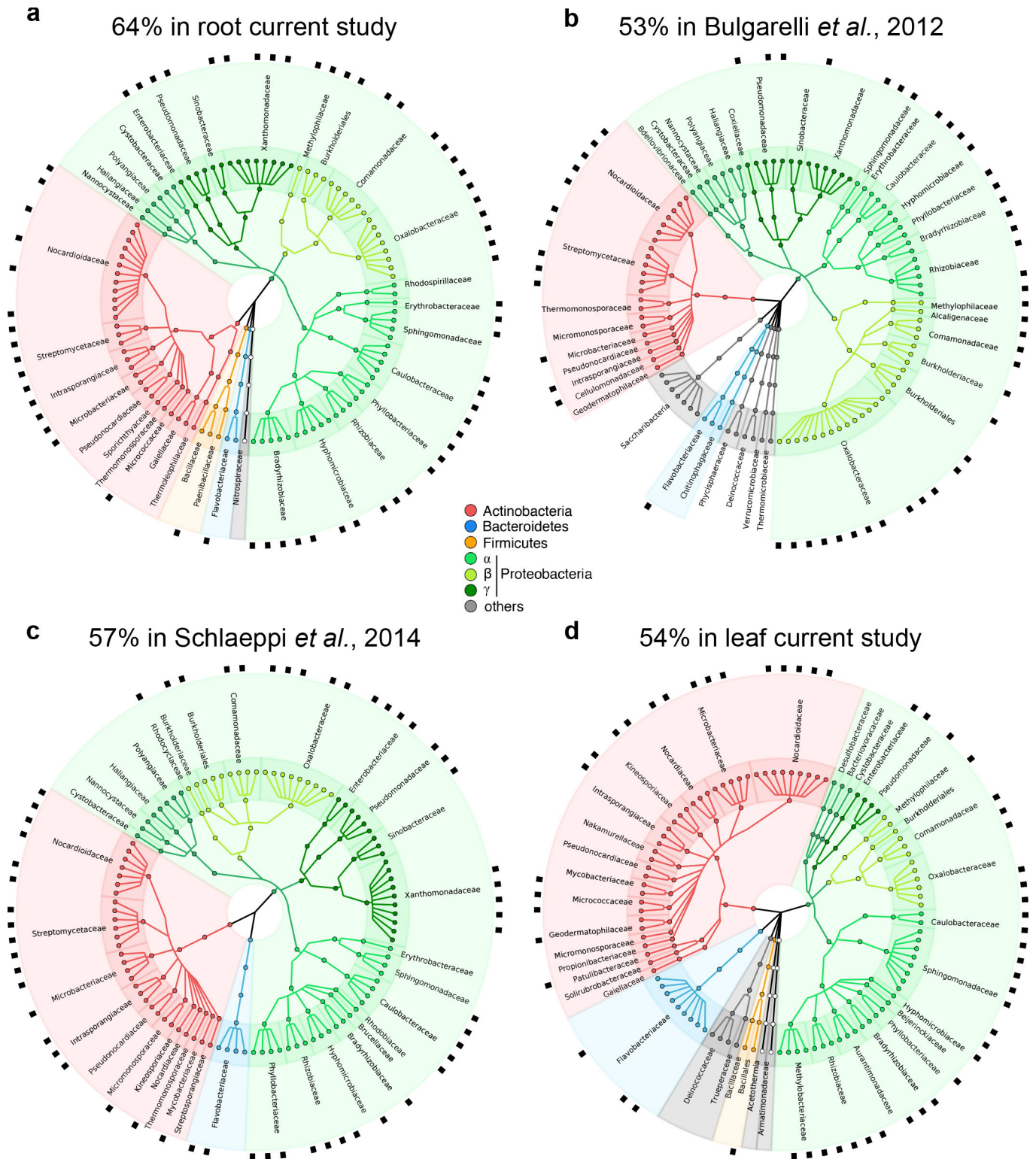
Forward and reverse reads were joined, demultiplexed and subjected to quality controls using scripts from the QIIME toolkit²⁷, v. 1.80 (Phred ≥ 20). The resulting high quality sequences were further clustered at 97% sequence identity together with Sanger sequences of leaf, root and soil isolates using the UPARSE²⁴ pipeline as described above. Taxonomic assignments of representative sequences were performed as explained in the previous sections. OTUs only corresponding to one or more Sanger 16S rRNA gene sequence(s) of purified strains in the *At*-RSPHERE, *At*-LSPHERE or soil collection were selected and designated 'indicator OTUs'. The heat maps were generated using the ggplot2 R package.

Accession numbers. Sequencing reads (454 16S rRNA, MiSeq 16S rRNA and WGS HiSeq reads) have been deposited in the European Nucleotide Archive (ENA) under accession numbers PRJEB11545, PRJEB11583 and PRJEB11584. Genome assemblies and annotations corresponding to the leaf, root and soil culture collections have been deposited in the National Center for Biotechnology Information (NCBI) BioProject database under accession numbers PRJNA297956, PRJNA297942 and PRJNA298127, respectively.

Code availability. All scripts for computational analysis and corresponding raw data are available at http://www.mpipz.mpg.de/R_scripts. The sequenced bacterial genomes as well as any future updates are available at <http://www.at-sphere.com>.

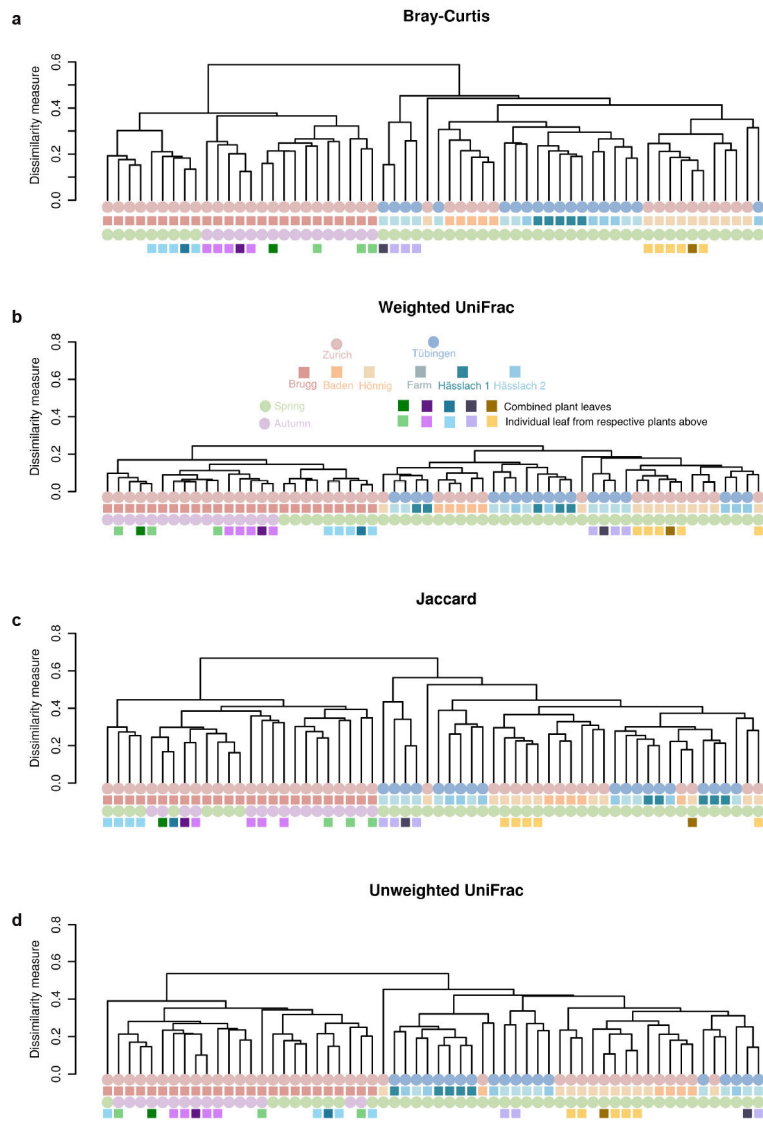
26. Chelius, M. K. & Triplett, E. W. The diversity of Archaea and Bacteria in association with the roots of *Zea mays* L. *Microb. Ecol.* **41**, 252–263 (2001).
27. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**, 335–336 (2010).

28. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72**, 5069–5072 (2006).
29. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* **26**, 266–267 (2010).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
31. Tritt, A., Eisen, J. A., Facciotti, M. T. & Darling, A. E. An integrated pipeline for de novo assembly of microbial genomes. *PLoS One* **7**, e42304 (2012).
32. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
33. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27**, 4636–4641 (1999).
34. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
35. Overbeek, R. *et al.* The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
36. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
37. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42**, D199–D205 (2014).
38. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
39. Wu, M. & Eisen, J. A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **9**, R151 (2008).
40. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539–539 (2011).
41. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
42. Dröge, J., Gregor, I. & McHardy, A. C. Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics* **31**, 817–824 (2015).
43. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
44. Bodenhausen, N., Bortfeld-Miller, M., Ackermann, M. & Vorholt, J. A. A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. *PLoS Genet.* **10**, e1004283 (2014).



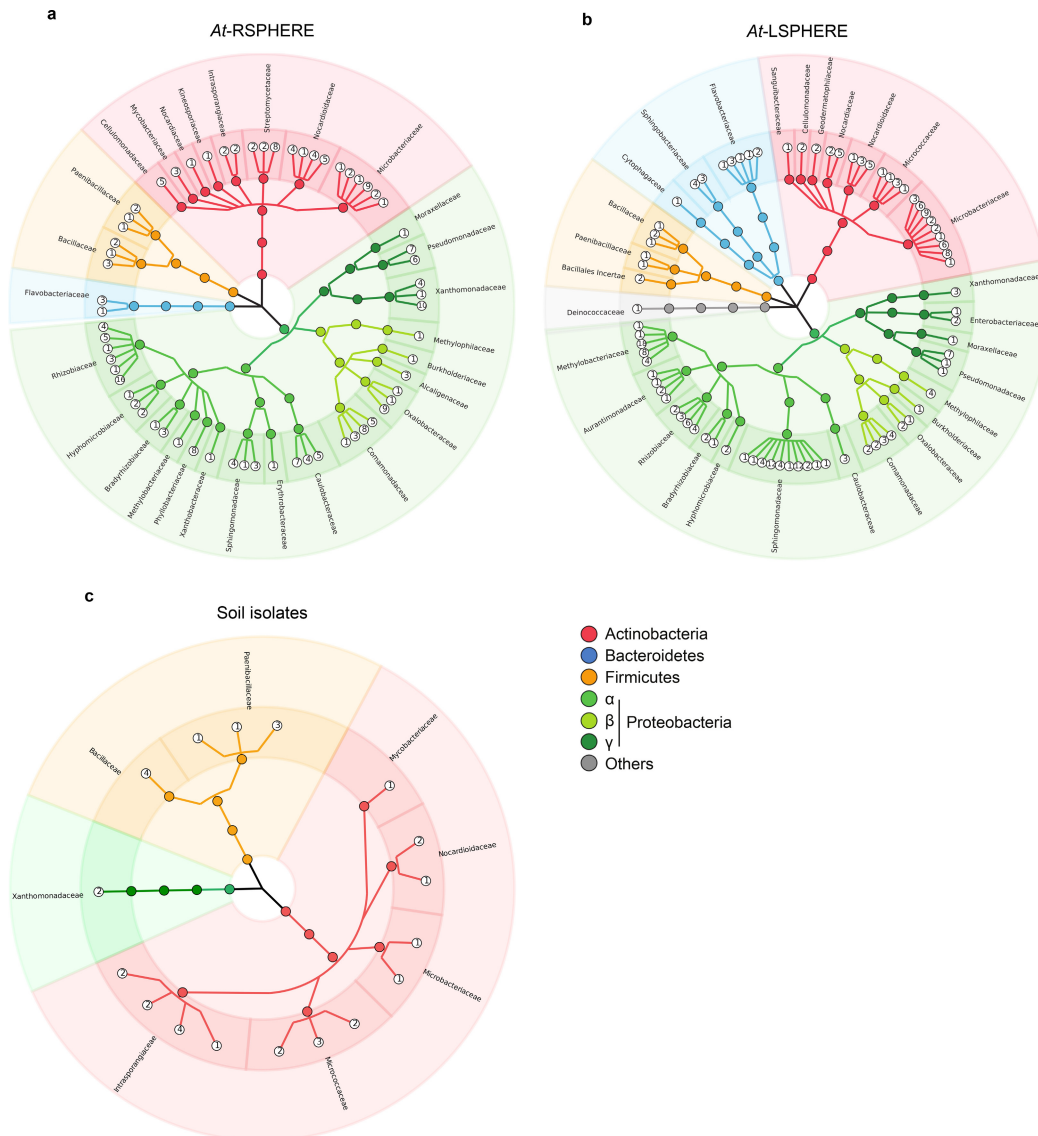
Extended Data Figure 1 | Culture-dependent coverage of *A. thaliana* root- and leaf-associated OTUs identified in several cultivation-independent studies. a–d. The inner circle depicts taxonomic assignments of top 100 root-associated OTUs (filled dots) for the indicated phyla and families that were identified in the current (a), ref. 6 (b) and

ref. 12 (c) studies with Cologne-soil-grown plants, and current leaf (d) study at locations around Tübingen and Zurich. Black squares of the outer ring highlight OTUs sharing $\geq 97\%$ 16S rRNA gene similarity to *Arabidopsis* root or leaf bacterial culture collection.



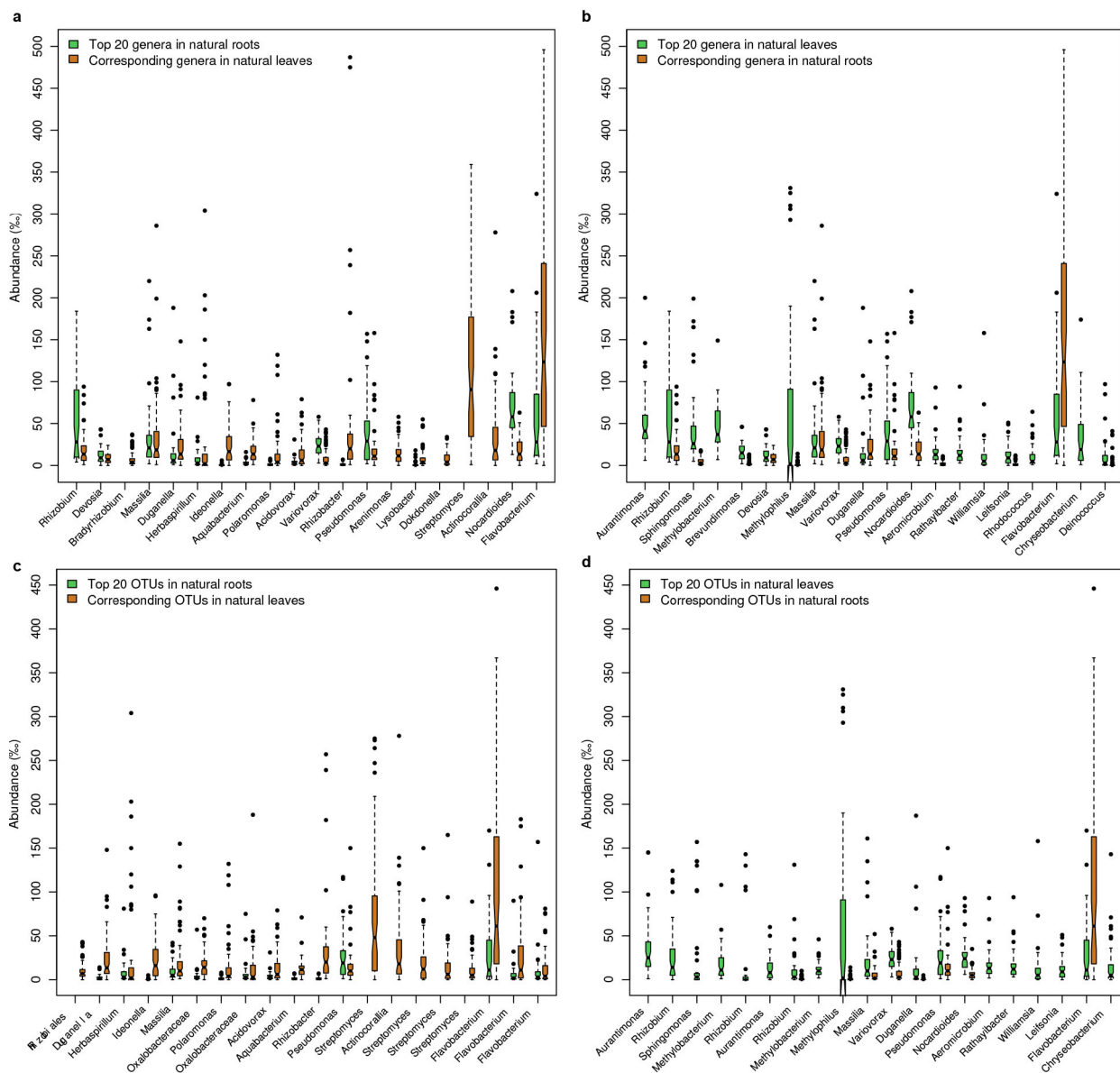
Extended Data Figure 2 | 16S rRNA gene community profiling of phyllosphere samples from different locations. a–d, The indicated Beta-diversity indices were calculated from leaf samples ($n = 60$) collected

from natural *A. thaliana* populations growing in the areas around Tübingen and Zurich. The indicated colour code refers to sampling locations, sampling sites, sampling season, and combined or individual leaves of respective plants.



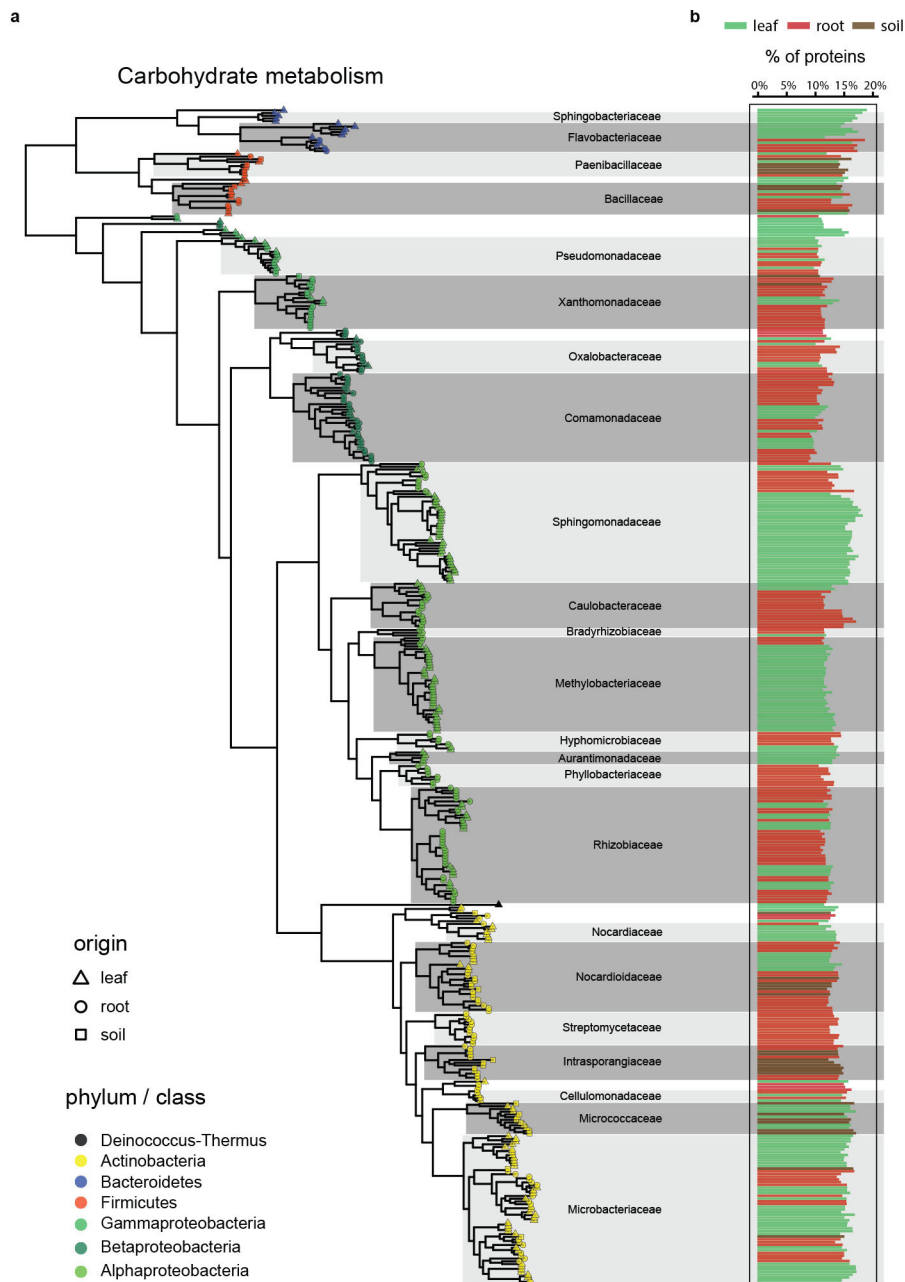
Extended Data Figure 3 | *At*-RSPHERE, *At*-LSPHERE and soil bacterial culture collections. **a**, *At*-RSPHERE ($n = 206$ isolates), a culture collection of the *A. thaliana* root microbiota. **b**, *At*-LSPHERE ($n = 224$ isolates), a culture collection of the *A. thaliana* leaf microbiota. **c**, Bacteria isolated

from Cologne soil ($n = 33$ isolates). Numbers inside white circles indicate the number of bacterial isolates sharing $\geq 97\%$ sequence identity, but isolated from independent roots, leaves and soil batches.



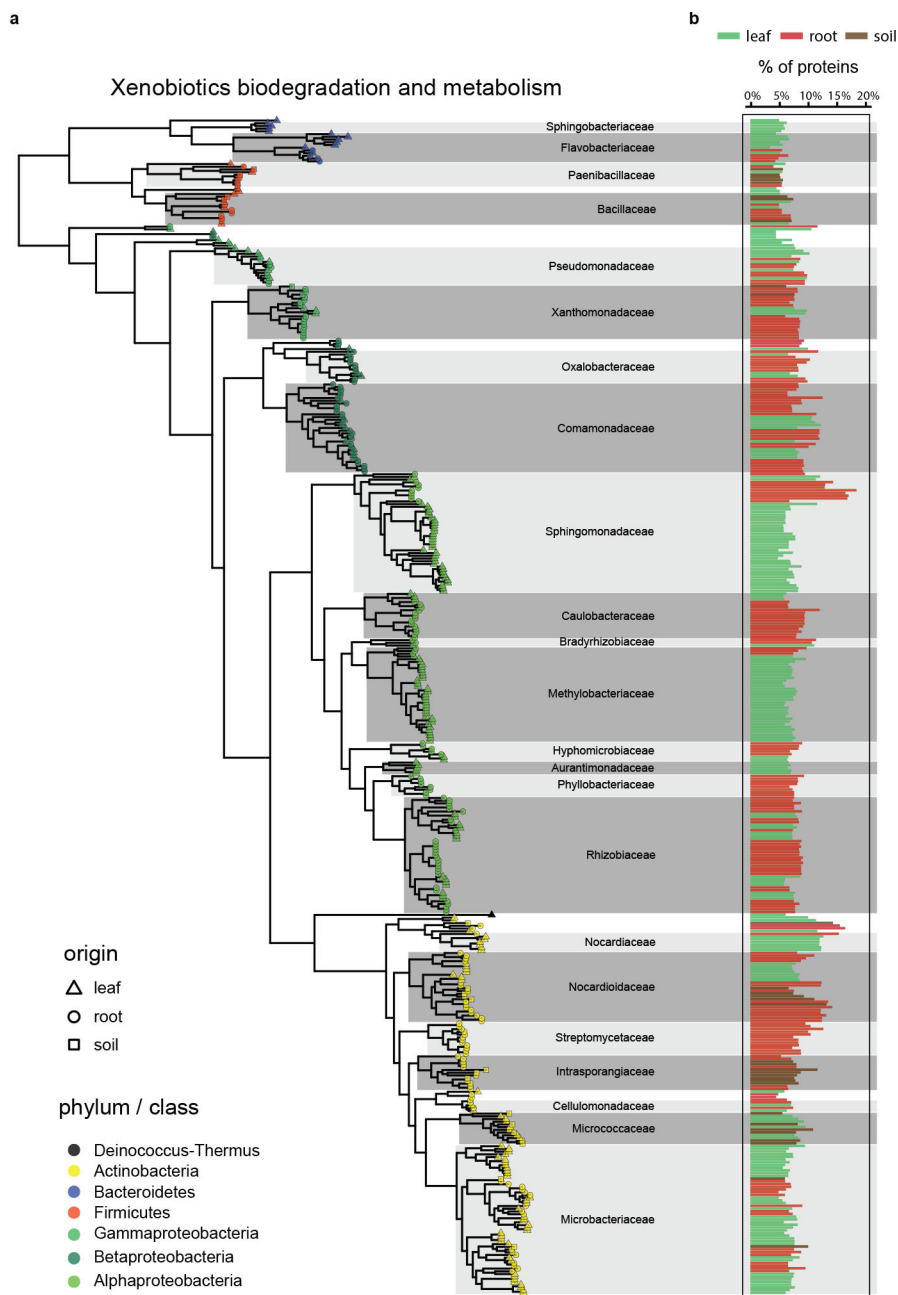
Extended Data Figure 4 | Taxonomy overlap between *A. thaliana* root- and leaf-associated bacterial community from plants grown in natural soils. a, b, Rank abundance plots of top 20 genera (a) and OTUs (b) in root bacterial communities ($n=8$) from Cologne with corresponding genera detected in leaf bacterial communities ($n=60$) from Zurich and

Tübingen. **c, d,** Rank abundance plots of top 20 genera (c) and OTUs (d) in leaf bacterial communities from Zurich and Tübingen with corresponding genera detected in root bacterial communities from Cologne. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.



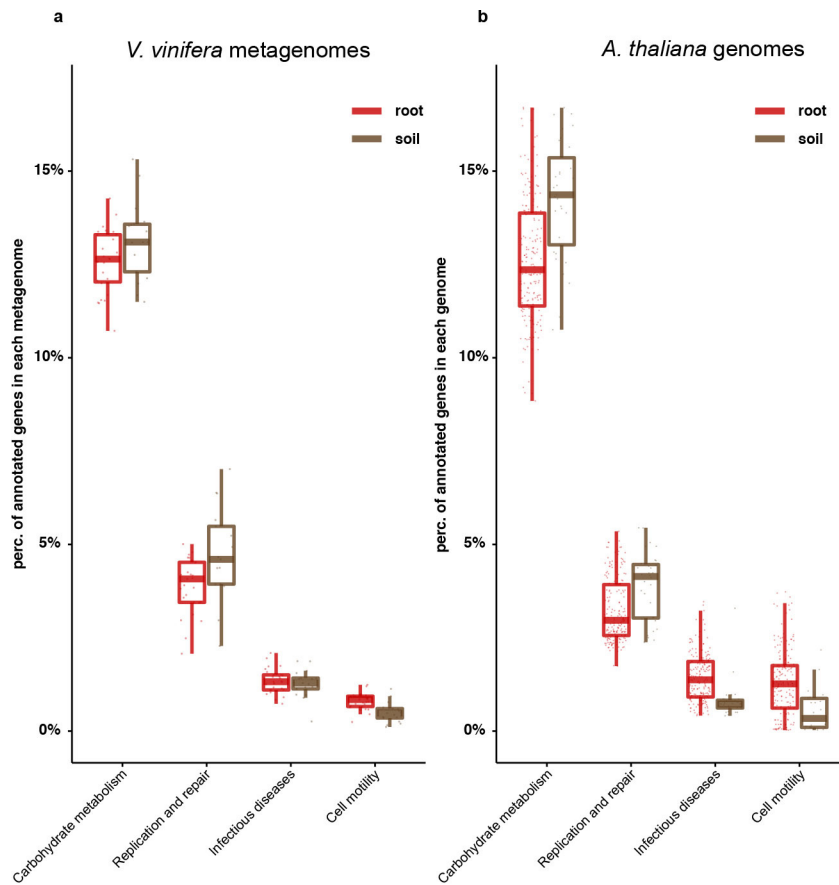
Extended Data Figure 5 | Phylogenetic distribution of ‘carbohydrate metabolism’ genes across sequenced isolates. **a**, Phylogeny of sequenced leaf ($n = 206$), root ($n = 194$) and soil ($n = 32$) isolates based on the concatenated alignment of the 31 conserved AMPHORA phylogenetic marker genes. The origin of each genome (leaf, root or soil) is shown by different shapes and their taxonomic affiliation (phylum level) is depicted

using various colours. Shaded areas correspond to the different clusters of genomes and are annotated with their consensus taxonomy (family level). **b**, Relative abundance of protein coding genes classified as belonging to the KEGG general term ‘carbohydrate metabolism’, measured as percentage of annotated proteins per genome.



Extended Data Figure 6 | Phylogenetic distribution of 'xenobiotic biodegradation and metabolism' genes across sequenced isolates.
a, Phylogeny of sequenced leaf ($n = 206$), root ($n = 194$) and soil ($n = 32$) isolates based on the concatenated alignment of the 31 conserved AMPHORA phylogenetic marker genes. The origin of each genome (leaf, root or soil) is shown by different shapes and their taxonomic

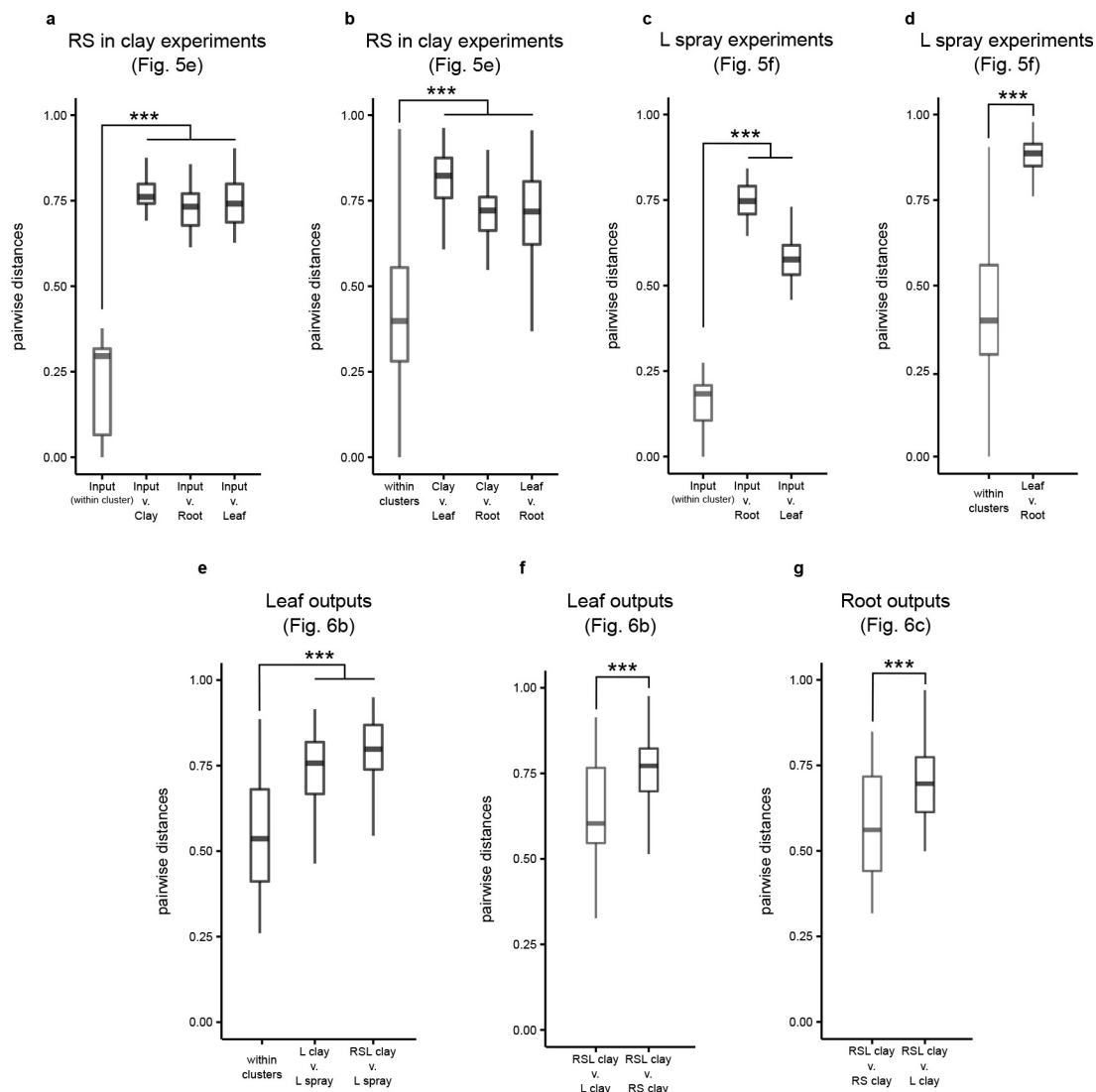
affiliation (phylum level; class level for Proteobacteria) is depicted using various colours. Shaded areas correspond to the different clusters of genomes and are annotated with their consensus taxonomy (family level). **b**, Relative abundance of protein coding genes classified as belonging to the KEGG general term 'xenobiotics biodegradation and metabolism', measured as percentage of annotated proteins per genome.



Extended Data Figure 7 | *V. vinifera* metagenome comparison.

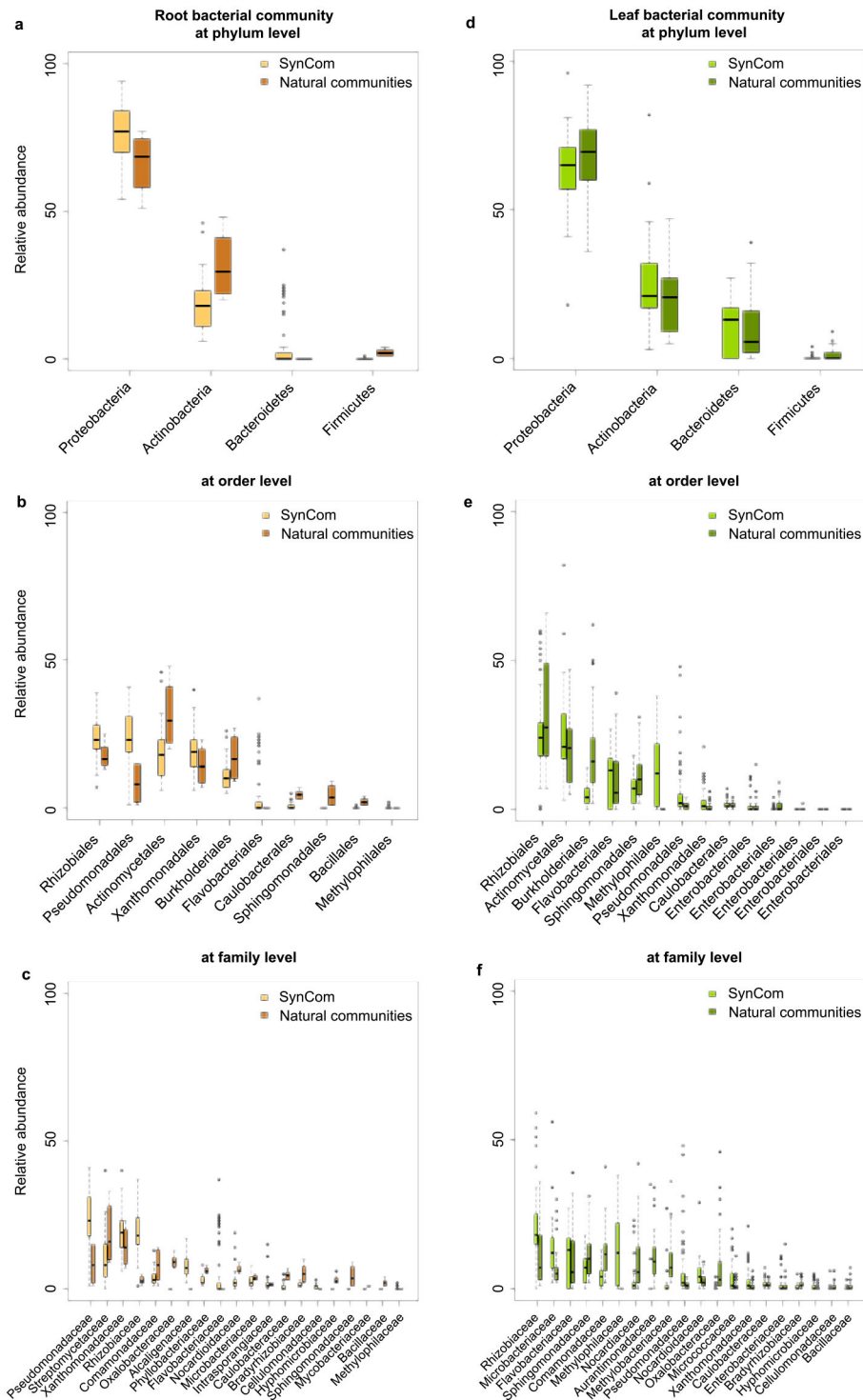
a, b, Functional enrichment analysis of *V. vinifera* root and soil shotgun metagenomes (**a**; $n = 47$) compared to *A. thaliana* culture collection root and soil genomes (**b**; $n = 432$). Functional category abundances

correspond to the percentage of annotated genes in each genome or metagenome sample. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.



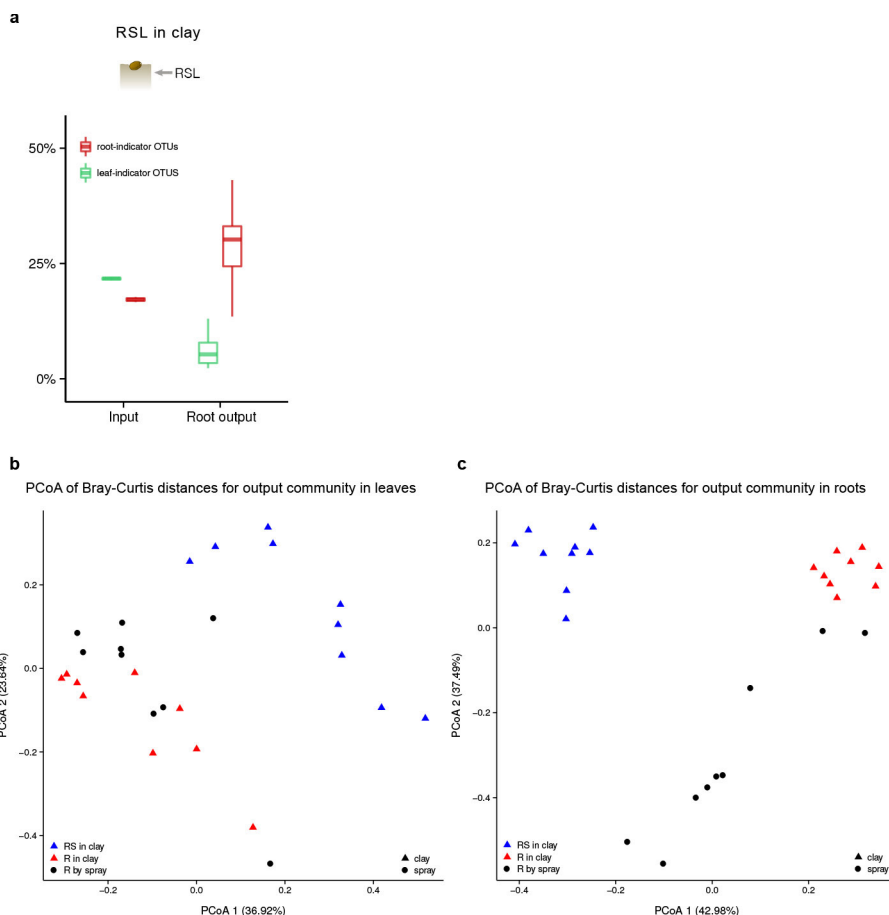
Extended Data Figure 8 | Cluster analysis of Bray-Curtis distances between groups of samples in the SynCom colonization of germ-free *A. thaliana* experiments. **a**, Comparison of pairwise distances within input samples and between input and output samples of the RS in clay experiments. **b**, Comparison of pairwise distances between samples within the same cluster and between different clusters of the RS in clay experiments. **c**, Comparison of pairwise distances between input samples and between input and output samples of the L spray experiments. **d**, Comparison of pairwise distances within samples within the same cluster and between different clusters of the L spray experiments. **e**, Comparison of pairwise distances between samples within the same cluster and between different clusters of the leaf output across experiments.

f, Comparison of pairwise distances between leaf output samples in the RSL in clay experiments and leaf output samples in the L in clay and RS in clay experiments. **g**, Comparison of pairwise distances between root output samples in the RSL in clay experiments and root output samples in the L in clay and RS in clay experiments. All comparisons marked with asterisks were subjected to a Student's *t*-test ($P < 0.001$ in each case). L in clay was tested with 6 independently prepared SynComs ($n = 6$); RSL in clay experiment was tested with 3 independently prepared SynComs, each used for 3 independent inoculations ($n = 9$). All other experiments were tested with 6 independently prepared SynComs and each preparation was used for 3 independent inoculations ($n = 18$). L, leaf-derived strains; RS, root- and soil-derived strains.



Extended Data Figure 9 | Similarity of rank abundances of SynCom outputs with corresponding root- and leaf-associated OTUs of plants grown in natural environments. **a–c**, Rank abundance plots of SynCom root outputs ($n = 69$) with corresponding root-associated OTUs in natural communities ($n = 8$) from plants grown in the present study in Cologne soil at the taxonomic ranks of phylum (**a**), order (**b**) and

family (**c**). **d–f**, Rank abundance plots of SynCom leaf outputs ($n = 69$) with corresponding leaf-associated OTUs in natural communities ($n = 60$) from plants grown in the present study around Tuebingen or Zurich at the taxonomic ranks of phylum (**d**), order (**e**) and family (**f**). Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.



Extended Data Figure 10 | Fractional contribution of *At*-LSPHERE and *At*-RPSHERE-specific OTUs and SynCom competition supports host organ-specific community assemblies. a, Fractional contribution of *At*-LSPHERE and *At*-RPSHERE specific OTUs in the input, leaf and the root output communities in the ‘RSL in clay’ experiment ($n = 9$). **b**, **c**, PCoA of Bray–Curtis distances of root (**b**; $n = 21$) and leaf (**c**; $n = 21$) outputs of the ‘R in clay’, ‘RS in clay’, and ‘R spray’ SynCom experiments.

R, root-derived isolates; S, soil-derived isolates; L, leaf-derived isolates. RSL in clay experiment was tested with 3 independently prepared SynComs, each used for 3 independent inoculations. All other experiments were tested with 3 independently prepared SynComs and each preparation was used for 3 independent inoculations. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.