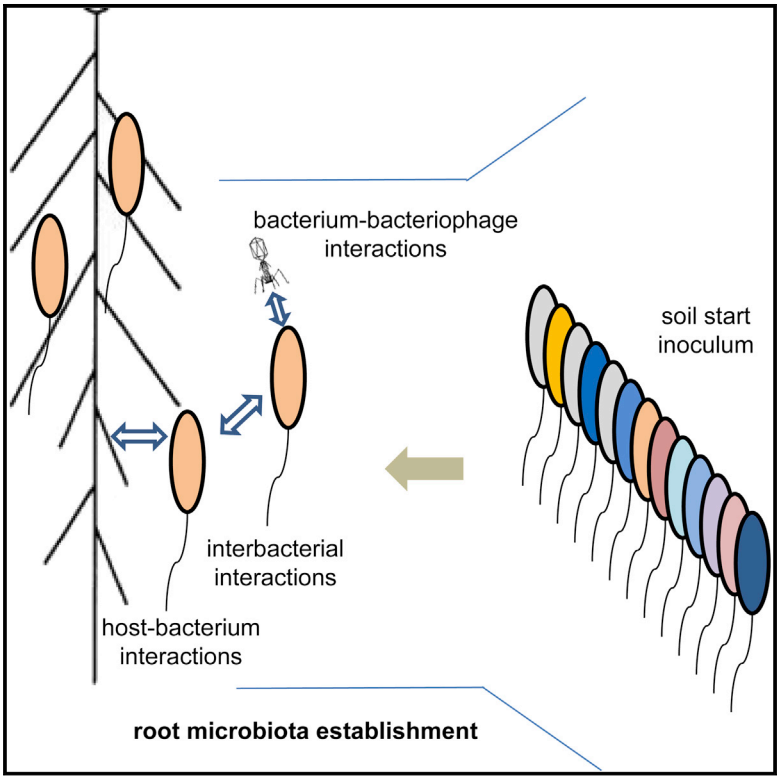
Resource

Cell Host & Microbe

Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley

Graphical Abstract



Authors

Davide Bulgarelli,
Ruben Garrido-Oter, ...,
Alice C. McHardy, Paul Schulze-Lefert

Correspondence

alice.mchardy@helmholtz-hzi.de (A.C.M.),
schlef@mpipz.mpg.de (P.S.-L.)

In Brief

Microbial communities inhabiting the root interior and surrounding soil contribute to plant growth. Bulgarelli et al. examine the microbiota that populates the roots of barley (*Hordeum vulgare*) and present evidence that integrated actions of microbe-microbe and host-microbe interactions drive root microbiota establishment through physiological processes occurring at the root-soil interface.

Highlights

- A small number of bacterial families dominate the root-enriched barley microbiota
- The host genotype determines the profile of a subset of community members
- Functions relevant for host interactions are enriched in root-associated taxa
- Genes mediating host, bacteria, and phage interactions show signs of positive selection

Bulgarelli et al., 2015, Cell Host & Microbe *17*, 392–403
March 11, 2015
http://dx.doi.org/10.1016/j.chom.2015.01.011

CellPress

CellPress

Cell Host & Microbe
Resource

Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley

Davide Bulgarelli,[1,4,6] Ruben Garrido-Oter,[1,2,3,6] Philipp C. Münch,[2] Aaron Weiman,[2] Johannes Dröge,[2] Yao Pan,[2,3] Alice C. McHardy,[2,3,5,7,*] and Paul Schulze-Lefert[1,3,7,*]

[1]Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany
[2]Department of Algorithmic Bioinformatics, Heinrich Heine University Duesseldorf, 40225 Duesseldorf, Germany
[3]Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany
[4]Division of Plant Sciences, College of Life Sciences, University of Dundee at The James Hutton Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK
[5]Computational Biology of Infection Research, Helmholtz Center for Infection Research, 38124 Braunschweig, Germany
[6]Co-first author
[7]Co-senior author
*Correspondence: alice.mchardy@helmholtz-hzi.de (A.C.M.), schlef@mpipz.mpg.de (P.S.-L.)
http://dx.doi.org/10.1016/j.chom.2015.01.011

SUMMARY

The microbial communities inhabiting the root interior of healthy plants, as well as the rhizosphere, which consists of soil particles firmly attached to roots, engage in symbiotic associations with their host. To investigate the structural and functional diversification among these communities, we employed a combination of 16S rRNA gene profiling and shotgun metagenome analysis of the microbiota associated with wild and domesticated accessions of barley (*Hordeum vulgare*). Bacterial families Comamonadaceae, Flavobacteriaceae, and Rhizobiaceae dominate the barley root-enriched microbiota. Host genotype has a small, but significant, effect on the diversity of root-associated bacterial communities, possibly representing a footprint of barley domestication. Traits related to pathogenesis, secretion, phage interactions, and nutrient mobilization are enriched in the barley root-associated microbiota. Strikingly, protein families assigned to these same traits showed evidence of positive selection. Our results indicate that the combined action of microbe-microbe and host-microbe interactions drives microbiota differentiation at the root-soil interface.

INTRODUCTION

Land plants host rich and diverse microbial communities in the thin layer of soil adhering to the roots, i.e., the rhizosphere, and within the root tissues, designated rhizosphere and root microbiota, respectively (Bulgarelli et al., 2013). Roots secrete a plethora of photosynthesis-derived organic compounds to the rhizosphere (Dakora and Phillips, 2002). This process, known as rhizodeposition, has been proposed as the major mechanism that enables plants to sustain their microbiota (Jones et al., 2009). In turn, members of the rhizosphere and root microbiota provide beneficial services to their host, such as indirect pathogen protection and enhanced mineral acquisition from surrounding soil for plant growth (Bulgarelli et al., 2013; Lugtenberg and Kamilova, 2009). Thus, the dissection of the molecular mechanisms underlying plant-microbe community associations at the root-soil interface will be a crucial step toward the rational exploitation of the microbiota for agricultural purposes. Recent studies performed using the model plant *Arabidopsis thaliana* revealed that the soil type and, to a minor extent, the host genotype shape root microbiota profiles (Bulgarelli et al., 2012; Lundberg et al., 2012). The structure of the microbial communities thriving at the root-soil interface appears to be resilient to host evolutionary changes, as indicated by a largely conserved composition of the root bacterial microbiota in *A. thaliana* and related species that spans 35 Ma of divergence within the family Brassicaceae (Schlaeppi et al., 2014). However, it is unclear whether microbiota divergence is greater in host species belonging to other plant families and whether the process of domestication, which gave rise to modern cultivated plants (Abbo et al., 2014) and which cannot be studied in *A. thaliana*, has left a human footprint of selection on crop-associated microbiota.

Barley (*Hordeum vulgare*) is the fourth-most cultivated cereal worldwide (Newton et al., 2011) and one of the earliest cereals consumed by humans, with evidence of presence of wild barley (*Hordeum vulgare* ssp. *spontaneum*) in human diets dating back to 17,000 BC (Kislev et al., 1992). Barley was one of the first plants subjected to domestication, which culminated ~10,000 years ago when the cultivation of domesticated barley (*Hordeum vulgare* ssp. *vulgare*) began in the Fertile Crescent. Anthropic pressure on barley evolution continued through diversification, which progressively differentiated early domesticated plants into several genetically distinct accessions whose area of cultivation radiated from the Middle East to the rest of the globe (Comadran et al., 2012). Nowadays, wild and cultivated barley accessions still coexist, providing an excellent experimental framework to investigate the structure and the evolution of the microbiota associated with a cultivated plant.

Here, we used an amplicon pyrosequencing survey of the bacterial 16S rRNA gene and combined it with state-of-the-art metagenomics and computational biology approaches to investigate

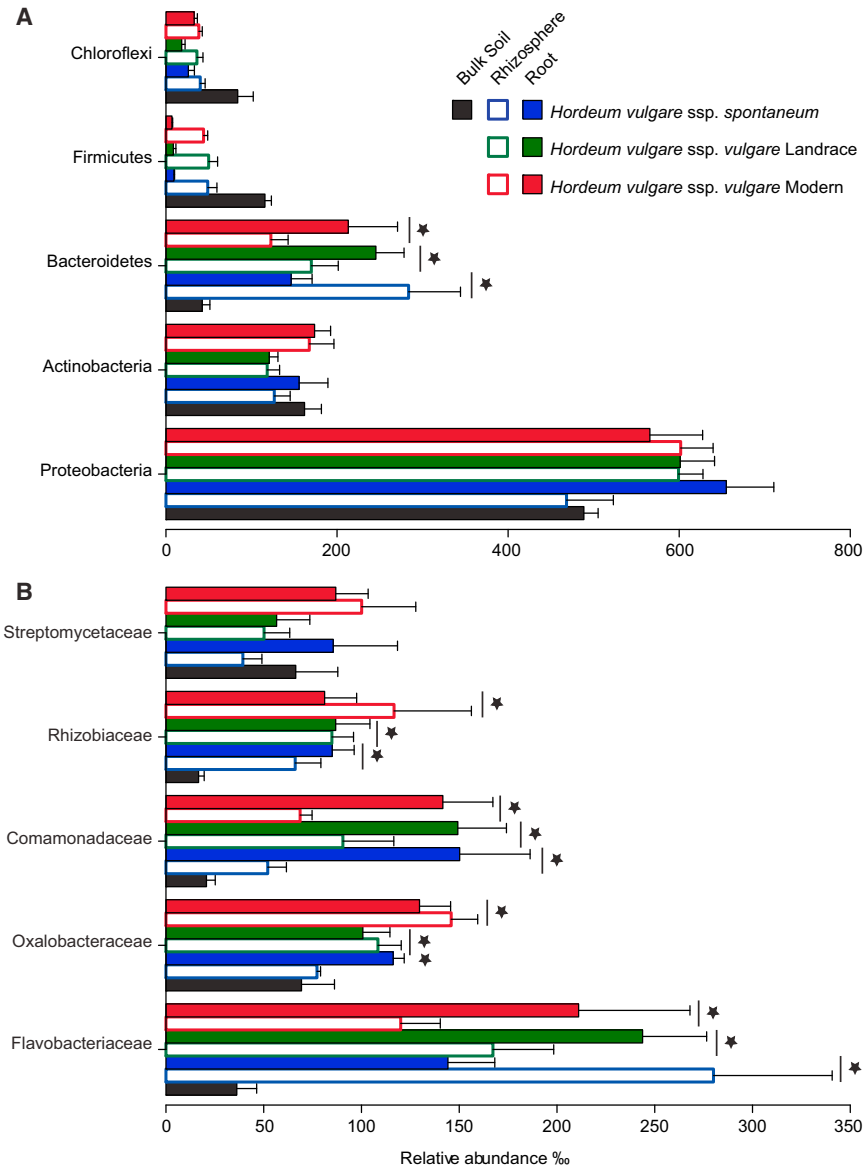


Figure 1. The Barley Rhizosphere and Root Microbiota Are Gated Communities

Average relative abundance (RA ± SEM) of the five most abundant (A) phyla and (B) families in soil, rhizosphere, and root samples as revealed by the 16S rRNA gene ribotyping. For each sample type, the number of replicates is n = 6. Stars indicate significant enrichment (FDR, $p < 0.05$) in the rhizosphere and root samples compared to bulk soil. Vertical lines denote a simultaneous enrichment of the given taxa in all three barley accessions. Only taxa with a RA > 0.5% in at least one sample were included in the analysis.

the structure and functions of the bacterial microbiota thriving at the barley root-soil interface. We found evidence for positive selection being exerted on a significant proportion of the proteins encoded by root-associated microbes, with a bias for cellular components mediating microbe-plant and microbe-microbe interactions.

RESULTS

The Structure of the Barley Bacterial Microbiota

We have grown barley accessions in soil substrates collected from a research field located in Golm, near Berlin (Bulgarelli et al., 2012), under controlled environmental conditions (Experimental Procedures). We subjected total DNA preparations from 6 bulk soil, 18 rhizosphere, and 18 root samples to selective amplification of the prokaryotic 16S rRNA gene with PCR primers encompassing the hypervariable regions V5-V6-V7 (Schlaeppi et al., 2014), and we generated 691,822 pyrosequencing reads. After in silico depletion of error-containing sequences, and chimeras as well as sequencing reads assigned to plant mitochondria, we identified 1,374 prokaryotic operational taxonomic units (OTUs) at 97% sequence similarity (Database S1; Experimental Procedures).

Taxonomic classification of the OTU-representative sequences to phylum level highlighted that Actinobacteria, Bacteroidetes, and Proteobacteria largely dominate the barley rhizosphere and root communities, where 88% and 96% of the pyrosequencing reads, respectively, were assigned to these three phyla. Of note, other members of the soil biota, such as Firmicutes and Chloroflexi, were virtually excluded from the plant-associated assemblages (Figure 1). The enrichment of members of the phylum Bacteroidetes significantly discriminated rhizosphere and root samples from bulk soil samples irrespective of the accession tested (moderated t test, false discovery rate-adjusted [FDR], p value < 0.05; Figure 1) At family level, Comamonadaceae, Flavobacteriaceae, and Rhizobiaceae designated a conserved barley microbiota whose enrichment differentiated the rhizosphere and root communities from bulk soil irrespective of the accessions tested (moderated t test, FDR, $p < 0.05$; Figure 1). Of note, the enrichment of a fourth family, Oxalobacteraceae, also significantly discriminated between root samples and unplanted soil in wild, landrace, and modern accessions (moderated t test, FDR < 0.05; Figure 1). Taken together, these results highlight a shift in community composition at the barley root-soil interface, which progressively differentiated the rhizosphere and root bacterial assemblages from the surrounding soil biota.

To gain insights into the richness of the barley microbiota we compared the total number of observed OTUs, Chao1, and the Shannon diversity indices of the communities retrieved from bulk soil and plant-associated microhabitats. All the indices revealed a significant reduction of the bacterial richness and

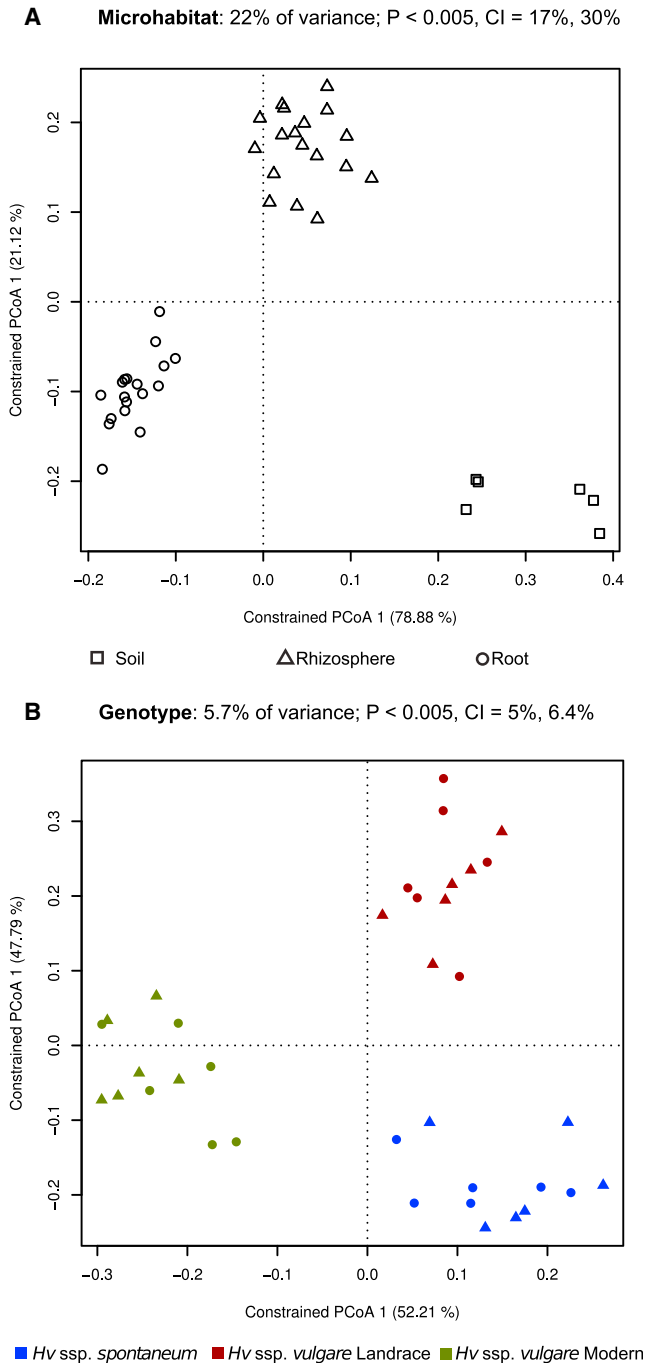


Figure 2. Constrained Principal Coordinate Analysis on the Soil and Barley Bacterial Microbiota

(A) Variation between samples in Bray-Curtis distances constrained by microhabitat (22% of the overall variance; p < 5.00E−2) and (B) by accession (5.7% of the overall variance; p < 5.00E−2). In both panels, triangles correspond to rhizosphere and circles to root samples. The percentage of variation explained by each axis refers to the fraction of the total variance of the data explained by the constrained factor. In (B) soil samples were not included.

diversity in the root samples (TukeyHSD, p < 0.05; Figure S1), while the rhizosphere microbiota displayed an intermediate composition between soil and root samples (Figure S1).

To elucidate whether the composition of the bacterial communities correlated or was independent of the sample type and the host genotype, we used the OTU count data to construct dissimilarity matrices with the UniFrac (Lozupone et al., 2011) and Bray-Curtis metrics. We applied a previously used relative abundances threshold (0.5%; Bulgarelli et al., 2012) to focus our analysis on PCR-reproducible OTUs. Permutational multivariate ANOVA based on distance matrices (ADONIS) revealed a marked contribution of the microhabitat (Bray-Curtis R2 = 0.11584; R2 Unweighted Unifrac R2 = 0.08851, p < 0.05) as well as phylogenetic-dependent contributions of the host genotype to the composition of the barley microbiota (Weighted Unifrac R2 = 0.24427; R2 Unweighted Unifrac R2 = 0.15262, p < 0.05). We used a canonical analysis of principal coordinates (CAP; Anderson and Willis, 2003) to better quantify the influence of these factors on the beta diversity. CAP analysis constrained by the environmental variables of interest revealed that the microhabitat explained 22% of the variance (p < 0.005; 95% confidence interval = 17%, 30%). Consistently, we observed a clear separation between plant-associated microhabitats and bulk soil samples followed by segregation of the rhizosphere and root samples (Figure 2A).

The host genotype alone could explain 5.7% of the overall variance of the data, and the constrained ordination showed a clear clustering of the samples corresponding to the wild, landrace, and modern accessions (Figure 2B). This proportion of the variation, albeit small, was found significant by permutation-based ANOVA (p < 0.005; Figure 2). Further exploration of these analyses revealed that the OTUs with the largest contribution to both constrained ordinations had a distinct taxonomic membership, mostly belonging to the phyla Proteobacteria and Bacteroidetes, and could explain most of the observed variation among microhabitats and genotypes (Figure S2A). Bootstrapping analysis of the constrained ordination (Experimental Procedures) indicated that the significance of the observed genotype effect could not be attributed to any individual OTUs. Only after randomly permuting the abundances of the 83 OTUs with the largest contribution (72.23% and 65.67% of the root and rhizosphere communities, respectively), the statistical significance was lost (Figure S2C). Consistently, CAP analyses generated using weighted UniFrac distance matrix, sensitive to OTU phylogenetic affiliations and OTU relative abundances, further supported the observed differentiation of the barley microbiota (Figure S2B). However, transformations based on unweighted UniFrac distance, which is sensitive to unique taxa, but not to OTU relative abundances, showed a drastic reduction of the variance explained by the microhabitat and failed to identify a significant host-genotype-dependent effect on the barley microbiota (Figure S2B). Together, these results further support the hypothesis that the barley rhizosphere and root are two microhabitats colonized by communities with taxonomically distinct profiles, which emerge from the soil biota through progressive differentiation.

To identify bacteria responsible for the diversification between the two root-associated microhabitats we employed a linear model analysis (Supplemental Experimental Procedures) to determine bacterial OTUs significantly enriched in root and rhizosphere compared to unplanted soil. With this approach we identified three distinct bacterial sub-communities thriving

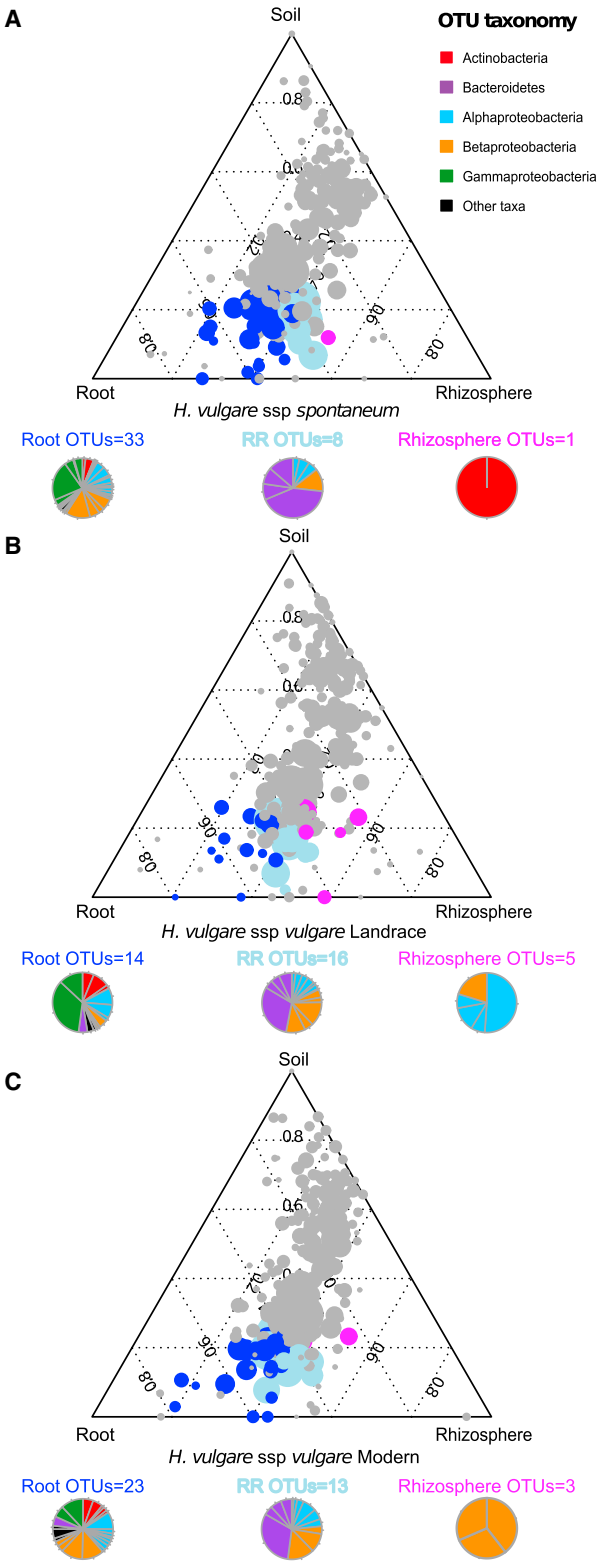


Figure 3. OTU Enrichment at the Barley Root/Soil Interface

Ternary plots of all OTUs detected in the data set with RA > 0.5% in at least one sample in (A) *Hordeum vulgare* ssp. *spontaneum*, (B) *H. vulgare* ssp. *vulgare* Landrace, and (C) *H. vulgare* ssp. *vulgare* Modern. Each circle represents one OTU. The size of each circle represents its relative abundance (weighted average). The position of each circle is determined by the contribution of the indicated compartments to the total relative abundance. Dark blue circles mark OTUs significantly enriched in the root microhabitat (*Root_OTUs*, FDR, $p < 0.05$), magenta circles mark OTUs significantly enriched in the rhizosphere microhabitat (*Rhizo_OTUs*, FDR, $p < 0.05$), and cyan circles mark OTUs significantly enriched in both microhabitats (*RR OTUs*, FDR, $p < 0.05$).

at the root-soil interface (Figure 3; Database S1). One sub-community, designated *Root_OTUs*, was defined by bacteria significantly enriched in the root samples and discriminating this sample type from bulk soil. *Root_OTUs* accounted for the largest fraction of the bacteria enriched in the barley microbiota in the wild and modern accessions (Database S1). A second sub-community was defined by bacteria enriched in both the rhizosphere and root samples and discriminating these samples from the bulk soil. This second sub-community, designated *RR_OTUs*, represented the largest fraction of the barley microbiota retrieved from the landrace accession (Database S1). Finally, a third sub-community defined by the bacteria discriminating the rhizosphere samples from bulk soil was identified. This sub-community, designated *Rhizo_OTUs*, represented the minor fraction of the barley microbiota irrespective of the accession tested (Database S1). Consistent with the constrained ordinations, taxonomic affiliations of the OTU-representative sequences assigned to *RR_OTUs* and *Root_OTUs* were largely represented by Bacteroidetes and Proteobacteria members (Database S1). We previously demonstrated that the root microbiota of the model plant *Arabidopsis thaliana* is dominated by members of Actinobacteria, Bacteroidetes, and Proteobacteria (Bulgarelli et al., 2012). We took advantage of the similar experimental platform used for the barley and *Arabidopsis* surveys, including the same soil type, to compare the bacterial microbiota retrieved from these monocotyledonous and dicotyledonous hosts. First, we re-processed the *A. thaliana* data set using exactly the same analysis pipeline we employed in the present study. Taxonomic classification using the representative sequences of the OTUs enriched in the root microbiota of barley and *A. thaliana* (Figure 4) revealed a similar taxonomic composition, with few bacterial taxa belonging to a limited number of bacterial families from different phyla, including members of Comamonadaceae, Flavobacteriaceae, Oxalobacteraceae, Rhizobiaceae, and Xanthomonadaceae. Notably, this analysis also revealed clear differences between the two host species. In particular, the enrichment in root samples of the families Pseudomonadaceae, Streptomycetaceae, and Thermomonosporaceae differentiated the *Arabidopsis* root-associated communities from barley. Conversely, the enrichment of members of the Microbacteriaceae family appears to be a distinctive feature of the barley root microbiota in the tested conditions. Excluding these qualitative differences, we found a very high correlation between the two sub-communities (0.90 Pearson correlation coefficient, $p = 0.005$).

The Barley Rhizosphere Microbiome

To gain further insights into the significance of the marked barley rhizosphere effect detected by the 16S rRNA gene survey, we reasoned that, unlike roots, where DNA is mostly plant derived, DNA isolated from the rhizosphere should mainly originate

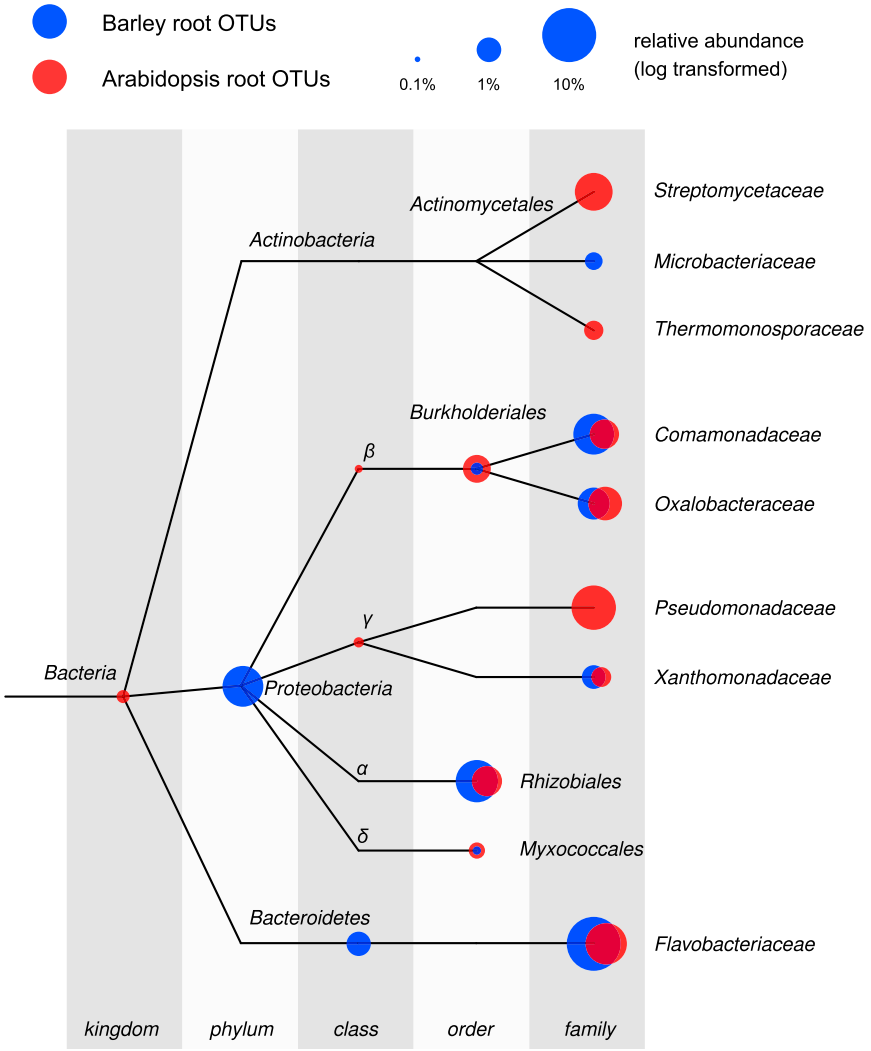


Figure 4. Taxonomic Representation of the Barley and *Arabidopsis* Root-Enriched Bacterial Taxa

The tree represents a subset of the NCBI taxonomy containing all OTUs found to be enriched in the barley and *Arabidopsis* root samples with respect to soil. The branches of the tree do not reflect evolutionary distances. The position of the dots corresponds to the taxonomic placement of each OTU-representative sequence in the taxonomy. The size of the dots illustrates the aggregated relative abundance of all OTUs assigned to a given taxon (log scale). OTUs enriched in *Arabidopsis* roots are depicted in red, whereas Barley root OTUs are shown in blue. Note that the relative abundance of each subset of root-enriched taxa with respect to its respective root community varies (Barley root OTUs, 45.44%; *Arabidopsis* root enriched OTUs, 59.02%).

from microbes, and we used the same rhizosphere DNA preparations for independent Illumina shotgun sequencing. We obtained two metagenome samples per host genotype, each corresponding to a different soil batch (Table S2) and generated an average of 75 million 100-bp paired-end reads per sample, adding up to a total of 44.90 Gb of sequence data. We then assembled the filtered reads of each sample independently using SOAPdenovo (Heger and Holm, 2000; Experimental Procedures). Despite the heterogeneity of the data, an average of 69.85% of the reads per sample were assembled into contigs (Table S2).

The partially assembled metagenome sequences (including unassembled singleton reads) were taxonomically classified with taxator-tk (Dröge et al., 2014), a tool for the taxonomic assignment of shotgun metagenomes (Experimental Procedures). Relative abundances were calculated by mapping the reads back to the assembled contigs and determining the number of reads assigned to each taxon. In total, 27.35% of all reads were assigned at least to the domain level. Of those, 94.04% and 0.054% corresponded to Bacteria and Archaea, respectively, and 5.90% to Eukaryotes (Database S1).

Comparison of SSU rRNA Genes and Metagenome Taxonomic Abundance Estimates

The availability of barley rhizosphere 16S rRNA gene amplicon and shotgun metagenome data provided an opportunity to compare both data sets. Toward this end, we classified the OTU-representative sequences onto the NCBI reference database (Sayers et al., 2009). This allowed us to cross-reference the relative abundances of each taxonomic bin from the rhizosphere metagenome with each OTU from the 16S rRNA gene analysis using the NCBI taxonomy and to directly compare the results of the two approaches (Figure 5). The analysis of the metagenome samples revealed the presence of Archaea (0.058% relative abundance) in the rhizosphere microhabitat, as well as members of bacterial phyla whose presence we did not detect in our 16S rRNA gene analysis, such as the Cyanobacteria (0.024% relative abundance). Our results also indicated an overrepresentation for Beta- and Gammaproteobacteria in the 16S rRNA gene taxonomic profiling, representing 10.12% and 9.64% of the whole community, respectively, compared with 7.73% and 5.50% as found in the metagenome samples. These quantitative differences can be at least partially attributed to the fact that Beta- and Gammaproteobacteria possess multiple ribosomal RNA operon copies (Case et al., 2007). The observed differences in detected taxa can furthermore be explained by known biases of 16S rRNA gene primers, in particular, the 799F primer was designed to avoid contamination from chloroplast 16S sequences, a side effect of which is a strong bias against Cyanobacteria (Chelius and Triplett, 2001).

We further assessed the variability in abundance estimates for bacterial taxa which could be detected in both analyses (excluding Cyanobacteria) and found several discrepancies, despite the overall high correlation (0.86 Pearson coefficient; $p < 1.75E-12$). The largest differences were found in taxonomic

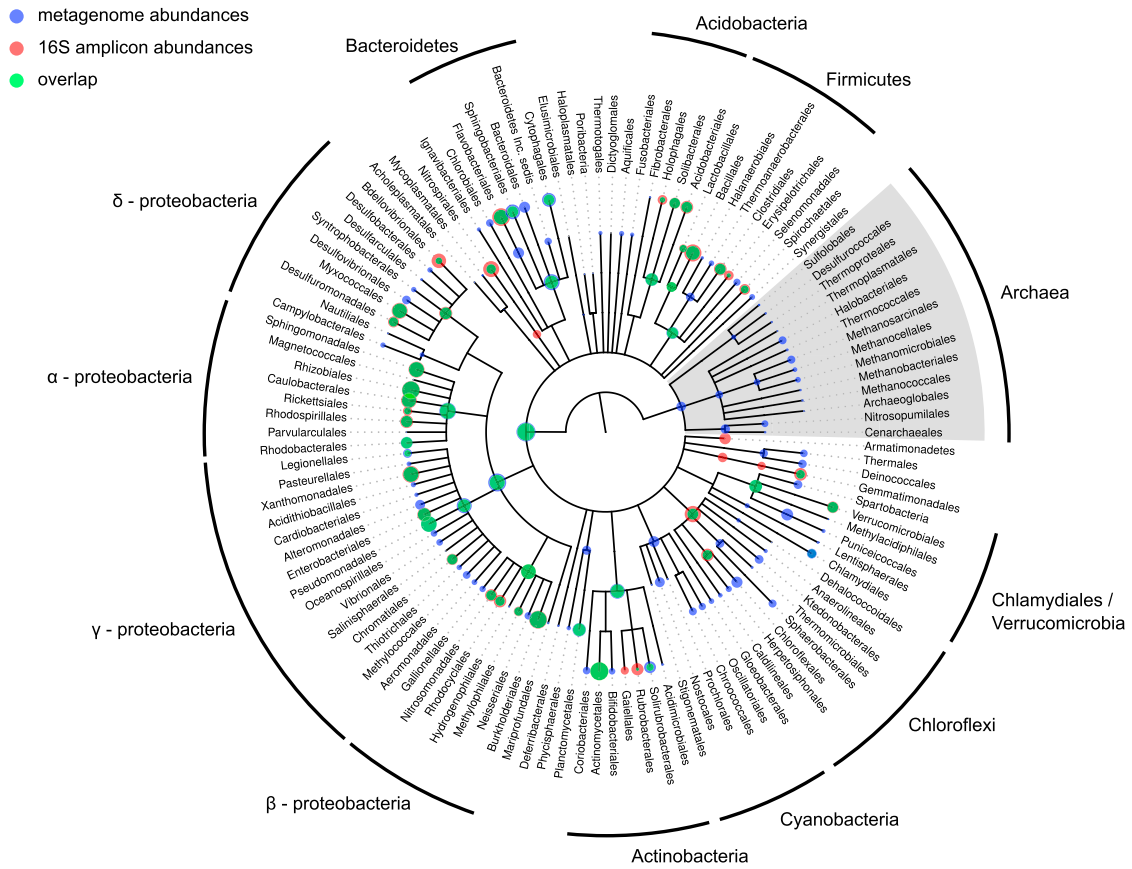


Figure 5. Comparison of 16S rRNA Amplicon and Metagenome Abundances

The tree represents the NCBI taxonomy for all taxonomically classified OTUs from the rhizosphere samples of the 16S rRNA survey as well as all metagenome bins, resolved down to the order rank. The branches of the tree do not reflect evolutionary distances. The position of the dots in the tree corresponds to the taxonomic placement of the representative sequences in the NCBI taxonomy. The size of the dots illustrates the average relative abundances per sample of each taxa (log scale). Blue dots represent abundances as found in the shotgun metagenome classification, red dots correspond to abundances from the 16S rRNA amplicon data, and green depicts an overlap.

groups for which 16S rRNA gene pyrotagging was reported to be either biased or lacking in resolution, due to either copy number variation or primer biases, especially for soil bacteria belonging to Chloroflexi, Deltaproteobacteria, and Bacteroidetes (Hong et al., 2009; Klindworth et al., 2013).

The taxonomic classification of fragments of 16S rRNA genes found in the metagenome shotgun reads allowed us to calculate the relative abundances of bacterial taxa not affected by primer biases. We found a high correlation between the results obtained for the two different 16S rRNA gene data sets (Figure 5; 0.89 Pearson correlation coefficient; $p < 21.55E{-}14$), indicating that the negative impact of the 799F primer bias on the beta-diversity estimates for the barley rhizosphere is only marginal, further validating the results reported above.

We also retrieved and analyzed 18S rRNA sequences following the same approach, which allowed us to compare eukaryotic and bacterial abundances in a quantifiable way. We found an increase in the relative abundance of eukaryotes (11.06%) when comparing 16S and 18S sequences relative to the estimate obtained from taxonomically classifying the metagenome sequences (5.90%), which could be partially explained by the high number and variability of rRNA operon copy number in eukaryotes (Amaral-Zettler et al., 2009). Furthermore, we were able to characterize the relative abundances of the major taxonomic groups found in the rhizosphere (Figure S3), revealing that fungi constitute the most abundant eukaryotic phylum in the barley rhizosphere (33.31% of all Eukaryotes).

Enrichment of Biological Functions in Root- and Rhizosphere-Associated Bacterial Taxa

The 16S rRNA gene survey revealed a clear dichotomy between the taxonomic composition of soil and root bacterial communities, a differentiation which, in barley, starts in the rhizosphere. Furthermore, a large fraction of bacterial taxa enriched in roots (*Root_OTUs*) was also enriched in the rhizosphere relative to unplanted soil (designated *RR_OTUs*). To determine if this differentiation process is linked to specific biological functions, we identified and annotated protein coding sequences (Experimental Procedures) and tested whether particular biological traits were significantly enriched in family-level taxonomic bins corresponding to RR_OTUs (containing 29.51% of all annotated protein coding sequences) with respect to

Table 1. Biological Functions in Root- and Rhizosphere-Associated Bacterial Taxa

Functional Category	p Value[a]
Protein secretion system type III	0.0013
Adhesion	0.0014
Regulation of virulence	0.0024
Siderophores	0.0024
Secretion	0.0072
Transposable elements	0.0177
Periplasmic stress	0.0188
Sugar phosphotransferase systems	0.0251
Bacteriophage integration excision lysogeny	0.0346
Invasion and intracellular resistance	0.0346
Protein secretion system type VI	0.0379
Detoxification	0.0379

Functional categories significantly enriched in taxonomic bins corresponding to *RR_OTUs* found in the barley rhizosphere metagenome.
[a]Calculated using a Mann-Whitney test, controlling for false discovery rate (FDR).

soil-associated bins, i.e., bins corresponding to OTUs which were not enriched in the root or in the rhizosphere (57.86% of the annotated sequences). Genes found in contigs that could not be taxonomically assigned, as well as those assigned to Cyanobacteria (12.81% of the total), were not included in this analysis.

We identified 12 functional categories which were significantly enriched in root and rhizosphere bacterial taxa (Table 1). These correspond to traits likely important for the survival or adaptation in the root-associated microhabitats, such as adhesion, stress response, and secretion. Importantly, categories relating to host-pathogen interactions (type III secretion system T3SS, regulation of virulence, invasion, and intracellular resistance) as well as microbe-microbe interactions (type VI secretion system; T6SS) and microbe-phage interactions (transposable elements, bacteriophage integration) were also significantly enriched. Interestingly, root- and rhizosphere-associated taxa were also significantly enriched in protein families related to iron mobilization (siderophore production) and sugar transport (sugar phosphotransferase systems).

To further assess the ecological significance of these functional enrichments, we performed a comparison with functional representation in sequenced isolates. We retrieved and analyzed 1,233 genomes from the NCBI database (Experimental Procedures; Supplemental Information) belonging to the soil- and root-associated bacterial taxa found in the barley rhizosphere and performed the same enrichment tests. We found only one functional category to be significantly enriched in the root-associated taxa with respect to the soil background taxa, namely, the T3SS ($p = 0.044$).

Positive Selection in the Barley Rhizosphere

To gain further insights on the molecular mechanisms driving the functional diversification of the barley rhizosphere microbiota, the gene families identified in the assembled barley metagenome were annotated based on matches to TIGRFAM (Haft et al., 2013) hidden Markov models (HMMs; Experimental Procedures), and we calculated, for each TIGRFAM, the ratio between the number of nonsynonymous (D_n) and synonymous (D_s) changes, a proxy for evolutionary pressure. Our analyses showed that 9% of the gene families had on average significantly higher D_n values and lower D_s values than the mean value calculated over all annotated sequences (one-sided Fisher test, FDR < 0.05), suggesting that they have been under positive (diversifying) selection. Interestingly, a closer investigation of these gene families revealed that positive selection signatures markedly characterize diverse proteins involved in pathogen-host interactions, including bacterial secretion, as well as proteins essential for phage defense (Figures 6A and S5). Strikingly, these proteins encode for a subset of the functions enriched in *RR_OTUs* and *Root_OTUs* (Table 1). Furthermore, we determined that 10.66% (115) of protein families encoded by the barley metagenome displayed a D_n/D_s ratio significantly greater than the metagenome mean D_n/D_s value in at least one of the barley genotypes tested (Table S3).

Of note, we identified significant signs of positive selection for a component of the T3SS, which is found in most Gram-negative bacteria and is used to suppress plant immune responses (Cornelis and Van Gijsegem, 2000; Table S6). Our findings are in line with previous studies, which reported evidence of positive selection for T3SS components in the bacterial phytopathogens *Pseudomonas syringae* (Guttman et al., 2006) and *Xanthomonas campestris* (Weber and Koebnik, 2006). Furthermore, we detected positive selection for components of the T6SS, a contact-dependent transport system mediating microbe-microbe interactions (Table S4; Russell et al., 2014). In particular, we found the forkhead-associated (FHA) domain to be under strong positive selection. This domain is a phosphopeptide recognition domain embedded in diverse bacterial regulatory proteins, which control various cellular processes including pathogenic and symbiotic interactions (Durocher and Jackson, 2002).

Microbial Elicitors and Effectors of Plant Immunity under Positive Selection

One branch of the plant immune system recognizes and is activated by a variety of evolutionary conserved microbial epitopes, designated microbe-associated molecular patterns (MAMPs) (Boller and Felix, 2009). The co-evolutionary arms race between the plant host and microbial pathogens leads to reciprocal selective pressure for the interacting proteins to change. To avoid activation of plant defenses, phytopathogens have evolved different mechanisms such as the diversifying evolution of elicitor epitopes by mutation or reassortment, and the injection of strain-specific pathogen effector proteins into host cells to intercept intracellular immune signaling (Shames and Finlay, 2012).

To identify putative elicitors of plant immune responses at the root-soil interface, we searched for genes that contained clusters of residues under positive selection using a sliding window approach (Figure 6B; Experimental Procedures). A total of 56 putative elicitors of plant immune responses were previously identified in the genomes of six plant pathogenic and a soil-dwelling bacterium using a similar approach (McCann et al., 2012). Remarkably, we found a semantic overlap of nine protein families

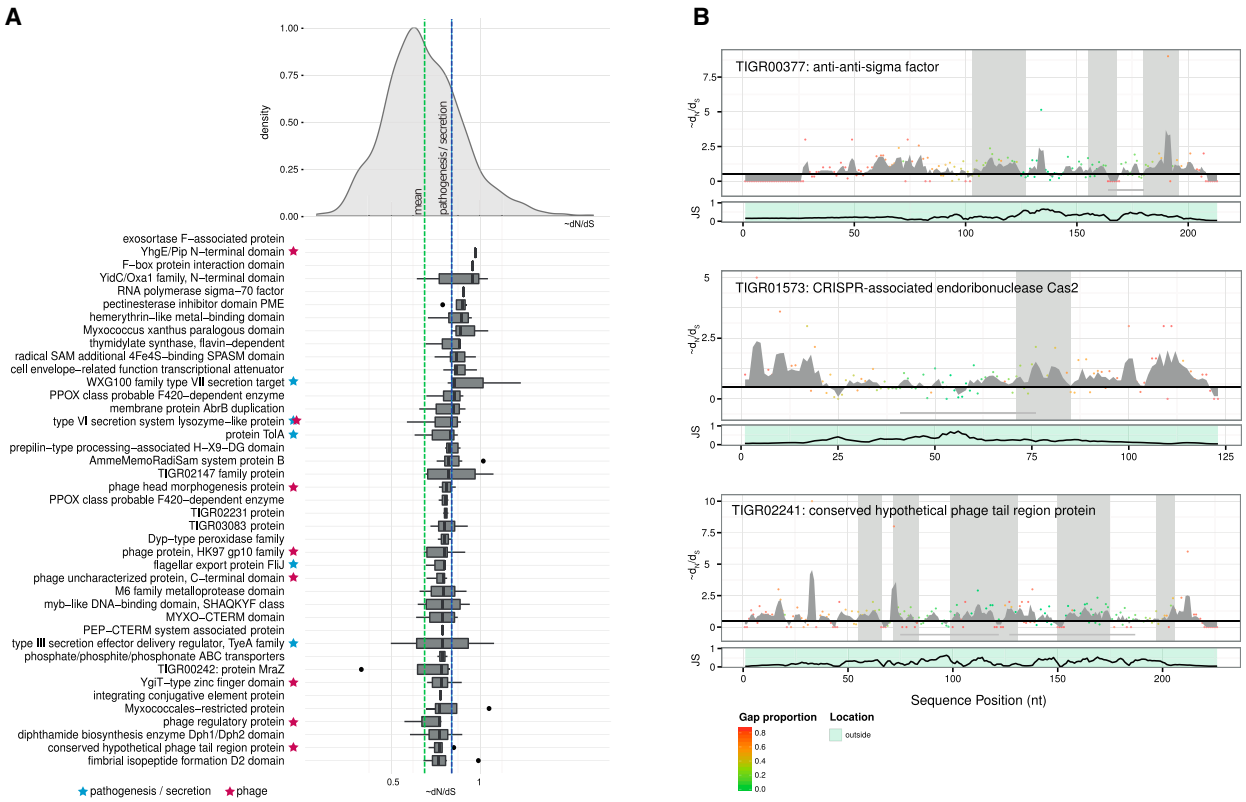


Figure 6. Proteins under Selection in the Barley Rhizosphere Microbiome

(A) Top-ranking protein families under positive selection with significantly increased D_n/D_s statistic. The distribution at the top shows the density function over all protein families smoothed with a Gaussian kernel function. The green bar indicates the average $\sim D_n/D_s$ over all the samples, the blue bar the average $\sim D_n/D_s$ for all TIGFRAMS annotated with the term "patho" and/or "secretion." The boxplot shows the distribution of the $\sim D_n/D_s$ across all samples for the top 50 ranked TIGRFAM families under positive selection, with families sorted by their median $\sim D_n/D_s$ in descending order. TIGRFAMs annotated with "repeat" or with a mean repetitive value of more than 50% were discarded.

(B) Sequence clusters of residues under positive selection in selected protein families. Top: dots indicate $\sim D_n/D_s$ for a given position in the protein sequence, and their color corresponds to the proportion of gaps in the multiple sequence alignment (MSA). Gray-shaded areas indicate significant clusters of residues under positive selection. Gray-shaded horizontal lines indicate repetitive elements. Bottom: Jensen-Shannon divergence as a function of the positions in the MSA.

under selection in the barley rhizosphere microbiome (Table S5). For example, the GGDEF domain, a previously reported putative bacterial elicitor, essential for motility and biofilm formation (Simm et al., 2004), was under positive selection in the rhizosphere of the wild accession ($p = 0.027$). Of the protein families that had a D_n/D_s ratio significantly higher than the mean, 85.3% had such clusters, whereas they were found in only 34.9% of all detected protein families ($p < 2.2$ E−16, one-sided Fisher's exact test). On average, we found 0.66 ± 1.54 (SD) clusters for each protein family, which spanned 4.0% ± 7.9% (SD) of their amino acid sequence among all families. For the protein families already shown to exhibit significant signatures of positive selection, an average of 6.7 ± 9.0 (SD) clusters were detected.

Furthermore, we identified by de novo prediction 16 putative polymorphic type III secreted effector proteins (T3SEs), of which 30% were under positive selection (Experimental Procedures; Table S6). In addition, 31.5% of these candidate effector proteins contained an average of 5.2 ± 9.8 (SD) clusters of residues under positive selection. This shows that, in the barley rhizosphere microbiota, highly polymorphic bacterial protein families, some of which are known to function in the suppression of plant immune responses, have similar footprints of positive selection as the evolutionary conserved MAMPs (McCann et al., 2012).

Positive Selection Acting on Phages and CRISPR Systems

Interestingly, in our D_n/D_s analysis we found that endoribonuclease gene *cas2* was under strong positive selection. This gene is associated with the clustered, regularly interspaced short palindromic repeat (CRISPR) system, a defense mechanism composed of an array of repeats with dyad symmetry separated by spacer sequences, which, together with a set of CRISPR-associated (CAS) genes, provides protection against phages in Bacteria and Archaea (Westra et al., 2014). In particular, Cas2 participates in the acquisition of new spacers (Barrangou et al., 2007), indicating that the ability to develop resistance to new phages might be an important trait for the bacterial community of the barley rhizosphere (Figure 6B). The enrichment of functional categories related to interactions with bacterial phages in *RR_OTUs* (Table 1) further supports this notion. In addition, we found that the coding sequences of bacteriophage tail and head morphogenesis genes were under positive selection. The

phage tail serves as a channel for the delivery of the phage DNA from the phage head into the cytoplasm of the bacteria. Thus, interactions between bacteria and their phages might have contributed to the positive selection on both the CRISPR-*cas* adaptive immune system of bacteria and on a subset of the bacteriophage proteins observed in the barley rhizosphere.

DISCUSSION

Here, we characterized the rhizosphere and the root microbiota of soil-grown wild, traditional, and modern accessions of barley using a pyrosequencing survey of the 16S rRNA gene. This revealed that the enrichment of members of the families Comamonadaceae, Flavobacteriaceae, and Rhizobiaceae and the virtual exclusion of members of the phyla Firmicutes and Chloroflexi differentiate rhizosphere and root assemblages from the surrounding soil biota. This microbiota diversification begins in the rhizosphere, where a marked initial community shift occurs, and continues in the root tissues by additional differentiation, leading to the establishment of a community inside roots, which is more distinct from the surrounding soil biota.

A comparison to the root and rhizosphere microbial assemblages retrieved from the distantly related dicotyledonous plants *Arabidopsis thaliana* and *A. thaliana* relatives (Bulgarelli et al., 2012; Lundberg et al., 2012; Schlaeppi et al., 2014) revealed both striking differences as well as common features. First, we detected in each of the three tested barley genotypes a marked "rhizosphere effect," i.e., a structural and phylogenetic diversification of this microhabitat from the surrounding soil biota (Figure 3), which we failed to detect in previous studies of *A. thaliana* and *A. thaliana* relatives (Bulgarelli et al., 2012; Schlaeppi et al., 2014). Second, taxonomic classification using the representative sequences of the OTUs enriched in the root microbiota of monocotyledonous barley and dicotyledonous *A. thaliana*, grown in the same soil type, revealed a similar enrichment pattern, although some clear differences were identified (Figure 4). On the basis of our study, the enrichment of members of the families Pseudomonadaceae, Streptomycetaceae, and Thermomonosporacea in root samples of *Arabidopsis* is not seen in barley. Consistently, recent cultivation-independent surveys of the rhizosphere of field-grown maize (Peiffer et al., 2013) and wheat (Turner et al., 2013), two grasses like barley, also revealed almost no enrichment of the aforementioned two actinobacterial taxa. By contrast, enrichment of members of the Microbacteriaceae family appears to be a distinct feature of the barley root microbiota. This suggests the existence of host lineage-specific molecular cues contributing to the differentiation of the root-associated microbiota from the surrounding soil type-dependent bacterial start inoculum. However, the overall conserved microbiota composition in the roots of the monocot barley and the dicot *Arabidopsis*, which diverged ~200 Ma, could be indicative of an ancient plant trait that preceded the emergence of flowering plants. Alternatively, but not mutually exclusive, the conserved microbiota composition might indicate that microbe-microbe interactions serve as a dominant structuring force of the root microbiota in flowering plants.

Our results revealed also a host-genotype-dependent stratification of both the barley root and rhizosphere microbiota (Figure 2B). The host influence on the microbiota profiles is limited, since ~5.7% of the variance can be explained by the factor host genotype and is entirely quantitative. Notably, the host genotype effect is manifested by variations in the abundance of many OTUs from diverse phyla, rather than by single OTUs. Re-analysis of root microbiota abundance data from three *A. thaliana* ecotypes (Schlaeppi et al., 2014), generated with the same 16S rRNA gene primers and using the same computational approach, failed to detect a significant ecotype-dependent effect. By contrast, our results from barley are congruent with a recent investigation of the rhizosphere microbiota of 27 field-grown modern maize inbreds (Peiffer et al., 2013). This study reported a similar proportion of variation attributed to the host genotype (5.0%–7.7% using unweighted or weighted UniFrac distances, respectively) and also a lack of individual bacterial taxa predictive for a given host genotype. Bouffaud and co-workers reported a stratification of the maize rhizosphere microbiota reflecting the major genetic groups emerged during maize diversification, rather than their genetic distance (Bouffaud et al., 2012). These results concur with our findings of accession-dependent microbiota differentiation (Figure 2B) owing to the fact that the tested wild, landrace, and modern accessions represent three distinct phases of the domestication and diversification history of barley (Meyer et al., 2012).

The availability of barley rhizosphere microbiome sequences prompted us to compare the taxonomic classification generated by shotgun DNA sequencing without PCR amplification with the 16S rRNA gene amplicon profiles. This allowed us to determine the presence of microorganisms whose presence cannot be estimated using the 16S rRNA gene primers we have adopted, such as Protists, Fungi, and Archaea. Furthermore, the use of assembly as an intermediate step to improve taxonomic classification of reads and abundance estimates is likely to introduce biases which are not fully understood. In order to assess this effect we retrieved marker genes from the unassembled metagenome reads to be analyzed and used as a control. Correlation tests between the abundance estimates for bacterial taxa obtained with the two methods (0.86 Pearson correlation coefficient; $p < 1.75E{-}12$) indicated that known 16S primer biases, differential ribosomal operon copy number, as well as assembly biases have a minor, but notable, impact on the analysis of beta-diversity, further underlining the importance of using complementary methods for the study of microbial diversity.

Strikingly, we found that Bacteria dominate the annotated barley rhizosphere, whereas the relative abundance of Eukaryotes accounted for only a small fraction. A recent study employing metatranscriptomics to estimate microbial abundances reported a 5-fold higher abundance of Eukaryotes in the oat and pea rhizosphere (16.6% and 20.7%, respectively) compared to that of wheat (3.3%) (Turner et al., 2013). However, since both metatranscriptome and metagenome abundance estimates are based on taxonomic classification using a reference-based method, database-related biases likely play a role in this apparent skew in the community in favor of bacterial taxa. Analysis of 18S rRNA sequences found in the shotgun reads revealed an increased relative abundance of Eukaryotes compared to the results obtained for the metagenome data (11.06% and 5.9%, respectively). However, given the large variation in rRNA operon copy number in eukaryotic genomes, abundance estimates based on 18S read counts are likely to be inflated. We conclude that further studies,

combining alternative markers such as the 18S rRNA gene or internal transcribed spacers (ITSs), targeting broader microbial communities (e.g., Fungi and Oomycetes), are needed to better estimate the phylogenetic composition of the microbiota thriving at the root-soil interface.

Combining our findings from the 16S rRNA gene survey, i.e., that some bacterial taxa are significantly enriched in root and rhizosphere samples with respect to soil (*RR_OTUs*), together with the functional analyses of the rhizosphere metagenome, we were able to map functions to root- and soil-associated taxa. Functional categories significantly enriched in root and rhizosphere (Table 1) corresponded to important traits for the survival and adaptation in these microhabitats, as well as traits related to microbe-microbe interactions and microbe-phage interactions. Importantly, several functions appeared to be relevant for interactions with the host (pathogenic as well as mutualistic), such as the T3SS, regulation of virulence, siderophore production, sugar transport, secretion, invasion, and intracellular resistance, further supporting the hypothesis that the presence of the host plant triggers a functional diversification in the rhizosphere. This is congruent with the observations that plants, through the release of photosynthesis-derived organic compounds into soil (Dakora and Phillips, 2002), can modify the physical, chemical, and biological properties of the rhizosphere to enhance the acquisition of important resources such as water and minerals (McCully, 1999). The growth of barley, like other graminaceous monocotyledons, relies on the secretion and subsequent reuptake of iron-chelating phytosiderophores for the acquisition of scarcely mobile iron ions from soil (Jeong and Guerinot, 2009). Therefore, the observed enrichment of bacterium-derived siderophores in the barley-associated microbial communities indicates that the combined action of microbiota- and host-derived siderophores maximizes the mobilization and bioavailability of the soil-borne iron micronutrient in the rhizosphere.

Out of the 12 categories found to be significantly enriched in the root-associated metagenome bins, only the T3SS was also detected as enriched when we analyzed sequenced isolates. This suggests that the T3SS is a relevant feature of root-associated bacterial taxa in general, whereas the remaining enriched functions detected only by analysis of the metagenome data (Table 1) could correspond to environment-specific features.

Analyzing the coding sequences found in the metagenome data, we observed strong positive selection in proteins that are known to directly interact with the plant host, such as the bacterial T3SS and other outer surface proteins, which might be related to plant-pathogen interactions and secretion (Figure 6). These signs of positive selection are evidence of plant-microbe co-evolution in the rhizosphere and suggest that host-microbe and microbe-microbe interactions exist in these natural community systems that are reminiscent of the arms race co-evolution model established for binary plant-pathogen interactions. Thus, our findings predict that the innate immune system of plants contributes to the selection of bacterial community structure as early as at the root-soil interface. Interestingly, it has been recently noted that balanced polymorphism of resistance genes in *A. thaliana* is maintained in the population through complex community-wide interactions encompassing many pathogen species (Karasov et al., 2014). The substantial number of protein families and the overall scale of positive selection which we identified indicate that metagenomic data are a sensitive tool for studying microevolution within natural environments. However, caution must be exercised when interpreting signatures of positive selection in this context, where the interplay between numerous species, including pathogens, mutualists, and commensals, creates a much more complex system than described by current models of co-evolution.

Previous comparative genomic studies of bacterial CAS genes surprisingly indicated no signs of positive selection, which was attributed to the additional roles of these genes in transcriptional regulation (Takeuchi et al., 2012). A high SNP density, indicative of positive selection, was also found for the CAS proteins csy1 and cse2 in metagenome samples of human gut microbiomes (Schloissnig et al., 2013). The strong signs of positive selection that cas2, one of the three essential proteins of the CRISPR system, exhibited in the barley rhizosphere, along with the positive selection identified for a subset of phage proteins, indicates that natural community systems might allow a more sensitive detection of such effects compared to comparative studies of a relatively small number of isolates. The role of the *cas2* gene in the acquisition of resistance to new phages might be of particular importance in a metabolically active and proliferating bacterial community, such as the rhizosphere microbiota (Ofek et al., 2014), which represents an ideal substrate for bacteriophage infections. Alternatively, the *cas2* gene product could be an elicitor of MAMP-triggered immunity in the host, which preferentially targets indispensable, evolutionary conserved, and broadly distributed microbial epitopes, such as flagellin or EF-Tu (McCann et al., 2012). Thus, the positive selection on CAS genes might simultaneously reflect the pressure exerted by bacteriophages and the host on members of the root-associated microbiota.

The observed overlap of bacterial traits under diversifying selection in the rhizosphere and those found to be significantly enriched in *RR_OTUs* provides direct and independent evidence for the contribution of host-microbe interactions in the selection of the root-associated bacterial microbiota from the surrounding soil biota (e.g., T3SS, virulence regulation and pathogenicity, siderophore production, sugar uptake). Our findings imply that the host innate immune system as well as the supply and demand of functions of root metabolism are relevant host factors for bacterial recruitment. In addition, both the analysis of the metagenome data (e.g., enrichment of T6SS) and the existence of a largely conserved phylogenetic pattern in the root-enriched bacterial taxa in barley and *A. thaliana* (Figure 4) imply that microbe-microbe interactions are also a driving force in the taxonomic differentiation of the root-associated bacterial assemblages. Thus, collectively, our results point toward a model in which the integrated action of microbe-microbe and host-microbe interactions drives root microbiota establishment through specific physiological processes from the surrounding soil biota.

EXPERIMENTAL PROCEDURES

Experimental Design

Surface-sterilized seeds of barley genotypes Morex, Rum, and HID369 were sown onto pots filled with experimental soil collected at the Max Planck Institute of Molecular Plant Physiology, Potsdam, in September 2010 and September 2011. For each accession we organized three biological replicates and repeated the entire experiment using two different samplings of soil

substrate (Table S1). At early stem elongation we excavated the plants from the soil and detached the root systems from the stems. We employed a combination of washing and ultrasound treatments to simultaneously separate the rhizosphere fraction from the roots and enrich for root endophytes. In parallel, bulk soil controls, i.e., pots filled with the same soil and exposed to the same environmental conditions as the plant-containing pots, were processed.

16S Data Analysis

16S rRNA gene sequences were subjected to demultiplexing, quality filtering, dereplication, abundance sorting, OTU clustering, and chimera identification using UPARSE pipeline (Edgar, 2013). Briefly, after removal of barcode and primer sequences, reads were truncated to a length of 290 bp, and only reads with a quality score Q > 15 and no ambiguous bases were retained for the analysis. Chimeras were identified using the "gold" reference database (http://drive5.com/uchime/gold.fa), and OTUs were defined at 97% sequence identity. OTU-representative sequences were taxonomically classified using the RDP classifier (Wang et al., 2007) trained on the Greengenes reference database. The resulting OTU table was used to determine taxonomic relative abundances and subsequent statistical analyses of alpha- and beta-diversity (see Supplemental Experimental Procedures).

Metagenome Data Analysis

Paired-end Illumina reads were subjected to trimming, filtering, and quality control using a combination of custom scripts and the CLC Workbench v5.5.1 and assembled using SOAPdenovo (Heger and Holm, 2000). A small fraction of the partially assembled metagenome samples (on average 3.02% of the reads) was mapped to the annotated barley genomic sequences, and the corresponding contigs or singleton reads were removed (Table S2; Supplemental Experimental Procedures). We used taxator-tk (Dröge et al., 2014) to taxonomically classify the partially assembled metagenome sequences (including unassembled singleton reads) using the NCBI database as a reference. Coding sequences were predicted using MetaGeneMark (Zhu et al., 2010) and annotated using matches to HMM (HMMER v3.0) profiles to the TIGRFAM (Haft et al., 2013) and PFAM (Punta et al., 2012) databases as well as a *k*-mer-based matching using the SEED (Edwards et al., 2012) API and server scripts. To test for a significant enrichment of functional categories in the root-associated bins relative to the remaining bins, we assumed a correspondence at the family level between metagenome bins and root- and rhizosphere-enriched OTUs (RR_OTUs) of these families found in the amplicon survey. To search for signatures of positive selection we first employed HMMER to obtain multiple sequence alignments (MSAs) of orthologous sequences found in the metagenome samples. From each MSA, we calculated neighbor-joining trees and used them to infer D_s and D_n changes. Clusters of residues with significant signs of positive selection were calculated using a sliding window approach. A detailed description of the methods and tools used for the analysis of the metagenome is available in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

The sequences generated in the barley pyrosequencing survey and the raw and assembled metagenomics reads reported in this study are deposited in the European Nucleotide Archive (ENA) under the accession number PRJEB5860. Individual metagenomes are also retrievable on the MG-RAST server under the IDs 4529836.3, 4530504.3, 4524858.3, 4524596.3, 4524591.3, and 4524575.3. The scripts used to analyze the data and generate the figures of this study are available at http://www.mpipz.mpg.de/R_scripts.

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, six tables, one database, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.chom.2015.01.011.

AUTHOR CONTRIBUTIONS

D.B. and P.S.-L. conceived of and designed the experiments. D.B. performed the experiments. D.B. and R.G.-O. analyzed the pyrosequencing data. R.G.-O., P.C.M., J.D., A.W., Y.P., and A.C.M. conceived of and performed the metagenomics analysis. D.B., R.G.-O., P.C.M., A.C.M., and P.S.-L. wrote the paper.

ACKNOWLEDGMENTS

We thank Isa Will and Maren Winnacker for their excellent technical assistance during the preparation of metagenomic DNA samples, Dr. Bruno Huettel and Diana Kuehn (Max Planck Genome Centre Cologne) for the preparation and sequencing of the 454 and Illumina libraries, and Dr. Kurt Stueber for the bioinformatic support. We thank Nina Dombrowski, Dr. Stijn Spaepen, Dr. Girish Srinivas, Dr. Stéphane Hacquard, Dr. Marc Erhardt, and Dr. Daniel Falush for their valuable comments on the manuscript. This work was supported by funds to P.S.-L. from the Max Planck Society, a European Research Council advanced grant (ROOTMICROBIOTA), and the "Cluster of Excellence on Plant Sciences" program funded by the Deutsche Forschungsgemeinschaft. D.B. was supported by a Royal Society of Edinburgh/Scottish Government Personal Research Fellowship co-funded by Marie Curie Actions.

Received: July 1, 2014
Revised: September 25, 2014
Accepted: January 6, 2015
Published: February 26, 2015

REFERENCES

Abbo, S., Pinhasi van-Oss, R., Gopher, A., Saranga, Y., Ofner, I., and Peleg, Z. (2014). Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. Trends Plant Sci. *19*, 351–360.

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS ONE *4*, e6372.

Anderson, M.J., and Willis, T.J. (2003). Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology. Ecology *84*, 511–525.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Boller, T., and Felix, G. (2009). A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. Annu. Rev. Plant Biol. *60*, 379–406.

Bouffaud, M.L., Kyselková, M., Gouesnard, B., Grundmann, G., Muller, D., and Moënne-Loccoz, Y. (2012). Is diversification history of maize influencing selection of soil bacteria by roots? Mol. Ecol. *21*, 195–206.

Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., et al. (2012). Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. Nature *488*, 91–95.

Bulgarelli, D., Schlaeppi, K., Spaepen, S., Ver Loren van Themaat, E., and Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants. Annu. Rev. Plant Biol. *64*, 807–838.

Case, R.J., Boucher, Y., Dahllöf, I., Holmström, C., Doolittle, W.F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl. Environ. Microbiol. *73*, 278–288.

Chelius, M.K., and Triplett, E.W. (2001). The diversity of archaea and bacteria in association with the roots of Zea mays L. Microb. Ecol. *41*, 252–263.

Comadran, J., Kilian, B., Russell, J., Ramsay, L., Stein, N., Ganal, M., Shaw, P., Bayer, M., Thomas, W., Marshall, D., et al. (2012). Natural variation in a homolog of *Antirrhinum* CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. Nat. Genet. *44*, 1388–1392.

Cornelis, G.R., and Van Gijsegem, F. (2000). Assembly and function of type III secretory systems. Annu. Rev. Microbiol. *54*, 735–774.

Dakora, F.D., and Phillips, D.A. (2002). Root exudates as mediators of mineral acquisition in low nutrient environments. Plant Soil *245*, 35–47.

Dröge, J., Gregor, I., and McHardy, A.C. (2014). Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. Bioinformatics. Published online November 10, 2014. http://dx.doi.org/10.1093/bioinformatics/btu745.

Durocher, D., and Jackson, S.P. (2002). The FHA domain. FEBS Lett. *513*, 58–66.

Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat. Methods *10*, 996–998.

Edwards, R.A., Olson, R., Disz, T., Pusch, G.D., Vonstein, V., Stevens, R., and Overbeek, R. (2012). Real time metagenomics: using k-mers to annotate metagenomes. Bioinformatics *28*, 3316–3317.

Guttman, D.S., Gropp, S.J., Morgan, R.L., and Wang, P.W. (2006). Diversifying selection drives the evolution of the type III secretion system pilus of *Pseudomonas syringae*. Mol. Biol. Evol. *23*, 2342–2354.

Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., and Beck, E. (2013). TIGRFAMs and genome properties in 2013. Nucleic Acids Res. *41* (Database issue), D387–D395.

Heger, A., and Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. Proteins *41*, 224–237.

Hong, S., Bunge, J., Leslin, C., Jeon, S., and Epstein, S.S. (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. ISME J. *3*, 1365–1373.

Jeong, J., and Guerinot, M.L. (2009). Homing in on iron homeostasis in plants. Trends Plant Sci. *14*, 280–285.

Jones, D.L., Nguyen, C., and Finlay, R.D. (2009). Carbon flow in the rhizosphere: carbon trading at the soil-root interface. Plant Soil *321*, 5–33.

Karasov, T.L., Kniskern, J.M., Gao, L., DeYoung, B.J., Ding, J., Dubiella, U., Lastra, R.O., Nallu, S., Roux, F., Innes, R.W., et al. (2014). The long-term maintenance of a resistance polymorphism through diffuse interactions. Nature *512*, 436–440.

Kislev, M.E., Nadel, D., and Carmi, I. (1992). Grain and fruit diet 19.000 years old at Ohalo II, Sea of Galilee, Israel. Rev. Palaeobot. Palynol. *73*, 161–166.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glöckner, F.O. (2013). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res. *41*, e1.

Lozupone, C., Lladser, M.E., Knights, D., Stombaugh, J., and Knight, R. (2011). UniFrac: an effective distance metric for microbial community comparison. ISME J. *5*, 169–172.

Lugtenberg, B., and Kamilova, F. (2009). Plant-growth-promoting rhizobacteria. Annu. Rev. Microbiol. *63*, 541–556.

Lundberg, D.S., Lebeis, S.L., Paredes, S.H., Yourstone, S., Gehring, J., Malfatti, S., Tremblay, J., Engelbrektson, A., Kunin, V., del Rio, T.G., et al. (2012). Defining the core *Arabidopsis thaliana* root microbiome. Nature *488*, 86–90.

McCann, H.C., Nahal, H., Thakur, S., and Guttman, D.S. (2012). Identification of innate immunity elicitors using molecular signatures of natural selection. Proc. Natl. Acad. Sci. USA *109*, 4215–4220.

McCully, M.E. (1999). ROOTS IN SOIL: unearthing the complexities of roots and their rhizospheres. Annu. Rev. Plant Physiol. Plant Mol. Biol. *50*, 695–718.

Meyer, R.S., DuVal, A.E., and Jensen, H.R. (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol. *196*, 29–48.

Newton, A.C., Flavell, A.J., George, T.S., Leat, P., Mullholland, B., Ramsay, L., Revoredo-Giha, C., Russell, J., Steffenson, B.J., Swanston, J.S., et al. (2011). Crops that feed the world 4. Barley: a resilient crop? Strengths and weaknesses in the context of food security. Food Security *3*, 141–178.

Ofek, M., Voronov-Goldman, M., Hadar, Y., and Minz, D. (2014). Host signature effect on plant root-associated microbiomes revealed through analyses of resident vs. active communities. Environ. Microbiol. *16*, 2157–2167.

Peiffer, J.A., Spor, A., Koren, O., Jin, Z., Tringe, S.G., Dangl, J.L., Buckler, E.S., and Ley, R.E. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. Proc. Natl. Acad. Sci. USA *110*, 6548–6553.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). The Pfam protein families database. Nucleic Acids Res. *40* (Database issue), D290–D301.

Russell, A.B., Peterson, S.B., and Mougous, J.D. (2014). Type VI secretion system effectors: poisons with a purpose. Nat. Rev. Microbiol. *12*, 137–148.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., et al. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. *37* (Database issue), D5–D15.

Schlaeppi, K., Dombrowski, N., Oter, R.G., Ver Loren van Themaat, E., and Schulze-Lefert, P. (2014). Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. Proc. Natl. Acad. Sci. USA *111*, 585–592.

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., Waller, A., Mende, D.R., Kultima, J.R., Martin, J., et al. (2013). Genomic variation landscape of the human gut microbiome. Nature *493*, 45–50.

Shames, S.R., and Finlay, B.B. (2012). Bacterial effector interplay: a new way to view effector function. Trends Microbiol. *20*, 214–219.

Simm, R., Morr, M., Kader, A., Nimtz, M., and Römling, U. (2004). GGDEF and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. Mol. Microbiol. *53*, 1123–1134.

Takeuchi, N., Wolf, Y.I., Makarova, K.S., and Koonin, E.V. (2012). Nature and intensity of selection pressure on CRISPR-associated genes. J. Bacteriol. *194*, 1216–1225.

Turner, T.R., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., Osbourn, A., Grant, A., and Poole, P.S. (2013). Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. ISME J. *7*, 2248–2258.

Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. *73*, 5261–5267.

Weber, E., and Koebnik, R. (2006). Positive selection of the Hrp pilin HrpE of the plant pathogen *Xanthomonas*. J. Bacteriol. *188*, 1405–1410.

Westra, E.R., Buckling, A., and Fineran, P.C. (2014). CRISPR-Cas systems: beyond adaptive immunity. Nat. Rev. Microbiol. *12*, 317–326.

Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. *38*, e132.

Cell Host & Microbe, Volume *17*

Supplemental Information

Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley

Davide Bulgarelli, Ruben Garrido-Oter, Philipp C. Münch, Aaron Weiman, Johannes Dröge, Yao Pan, Alice C. McHardy, and Paul Schulze-Lefert

Supplemental Information

Supplemental Data

Database S1 corresponds to an excel file containing the following data:

Worksheet Design_Barley: Design of the barley experiment;

Worksheet Barley_Phyla and Barley_Families relative abundances matrices for the taxonomic assignment at Phylum and Family level, respectively;

Worksheet PCR: Touch-down PCR programme used in this study;

Worksheets ws2 and ws3: OTU tables (including taxonomy information) 16S rRNA gene barley survey absolute counts and RA > 0.5% log2 transformed, respectively;

Worksheets ws4, ws5 and ws6: Count matrices for the observed OTUs, Chao and Shannon indices, respectively;

Worksheets ws7 ws8 and ws9: Bray-Curtis, Unweighted and Weighted Unifrac distance matrices of OTU count tables, respectively;

Worksheets ws10: Count matrices for the *Root*, *Rhizo* and *RR_OTUs* sub-communities retrieved from the wild accession;

Worksheets ws11: Count matrices for the *Root*, *Rhizo* and *RR_OTUs* sub-communities retrieved from the landrace accession;

Worksheets ws12: Count matrices for the *Root*, *Rhizo* and *RR_OTUs* sub-communities retrieved from the modern accession;

Worksheet ws13: Taxonomic abundances for each sample determined from shotgun metagenome data using taxator-tk;

Worksheet ws14: Taxonomic abundances determined from 16S rhizosphere samples using the NCBI reference database;

Worksheet ws15: Taxonomic abundances determined from 16S rRNA sequences found with Meta RNA in the shotgun metagenome reads;

Worksheet ws16: Eukaryotic abundances determined from 18S rRNA sequences found with Meta RNA in the shotgun metagenome reads.

Supplemental Tables

	Golm#2	Golm#3
Sampling date	Sep 10	Sep 11
Organic C(%)	1	2,5
Texture(%)		
Clay	4,2	1
Silt	4,2	1
Sand	91,6	98
Classification[1]	Sand	Sand
pH	7,12	6,88
Mineral content(mg/Kg)[2]		
Phosphorous	12,87	15,14
Potassium	27,75	28,18
Magnesium	5,44	5,44
Calcium	84,84	41,57
Nitrate	13,7	13,7

1 Soil texture classification according to FAO
2 Determined with H_2O extraction

Table S1. Physical and chemical characterisation of the experimental soil substrates used in this study. Relates to Figure 3.

general assembly statistics

sample	total number of reads	number of reads assembled	perc. of reads assembled	total number of contigs (inc. singleton reads)	partially assembled sample size (contigs and singleton reads) (in bp)	assembled sample size (only assembled contigs) (in bp)	number of contigs longer than 500 bp	assembled sample size (only contigs longer than 500 bp) (in bp)	longest contig (in bp)	N50*	N90*
A	103,797,442	66,279,430	63.85	53,279,288	4,919,889,439	1,381,997,074	26,780	18,548,447	6,857	651	520
B	57,126,852	37,536,598	65.71	29,989,928	2,728,559,798	675,158,135	18,979	17,247,207	10,609	910	539
C	58,328,864	38,593,720	66.17	31,303,983	2,810,008,297	689,700,519	18,593	14,155,374	8,916	729	530
D	89,089,142	61,767,764	69.33	46,674,741	4,093,136,563	1,124,208,336	13,768	8,826,307	6,380	609	516
E	55,564,696	42,793,240	77.02	31,874,397	2,547,674,014	558,176,685	3,419	2,151,299	8,687	589	512
F	87,005,576	65,939,656	75.79	47,529,870	3,913,772,227	1,012,605,301	14,400	9,401,280	11,687	620	517
total / avg.	450,912,572	312,910,408	69.65	240,652,207	21,013,040,338	5,441,846,050	95,939	70,329,914	8,856	685	522

barley filtering

sample	number of filtered contigs (inc. singleton reads)	number of filtered reads	longest filtered contig (in bp)	total size of filtered contigs (in bp)	percentage of partially assembled sample size filtered	number of reads filtered	perc. of total reads filtered	perc. of assembled reads filtered
A	352,403	1,356,735	2,634	53,799,852	1.09	1,590,014	1.53	2.04
B	325,554	1,207,028	3,356	50,282,653	1.84	1,437,008	2.52	3.22
C	344,874	1,305,705	3,122	53,474,095	1.90	1,536,723	2.63	3.38
D	547,023	2,199,147	2,934	84,293,478	2.06	2,522,633	2.83	3.56
E	554,381	2,387,301	3,558	87,154,810	3.42	2,667,531	4.80	5.58
F	664,223	2,977,645	2,903	104,422,724	2.67	3,289,624	3.78	4.52

sample description and env. variables

sample	age (weeks after sowing)	description	developmental stage	temperature	humidity (%)	photoperiod	climate environment	disease status
A	8	modern accession	early stem elongation	20°C day/18°C night	70	16h day/8h night	greenhouse	not detectable
B	8	wild accession	early stem elongation	20°C day/18°C night	70	16h day/8h night	greenhouse	not detectable
C	8	landrace accession	early stem elongation	20°C day/18°C night	70	16h day/8h night	greenhouse	not detectable
D	8	modern accession	early stem elongation	20°C day/18°C night	70	16h day/8h night	greenhouse	not detectable
E	8	wild accession	early stem elongation	20°C day/18°C night	70	16h day/8h night	greenhouse	not detectable
F	8	landrace accession	early stem elongation	20°C day/18°C night	70	16h day/8h night	greenhouse	not detectable

sample	soil batch	host genotype	host common name	host scientific name	comments
A	Golm #2	Morex	barley	Hordeum vulgare ssp. vulgare	Morex is a cultivated malting variety from USA
B	Golm #2	HID369	wild barley	Hordeum vulgare ssp. spontaneum	HID is a wild accession from Israel
C	Golm #2	Rum	barley	Hordeum vulgare ssp. vulgare	Rum is a landrace (i.e. cultivated by subsistence farmers) from Jordan
D	Golm #3	Morex	barley	Hordeum vulgare ssp. vulgare	Morex is a cultivated malting variety from USA
E	Golm #3	HID369	wild barley	Hordeum vulgare ssp. spontaneum	HID is a wild accession from Israel
F	Golm #3	Rum	barley	Hordeum vulgare ssp. vulgare	Rum is a landrace (i.e. cultivated by subsistence farmers) from Jordan

* N50 and N90 statistics were calculated on contigs larger than 500 bp

Table S2. Description of shotgun metagenome samples, assembly statistics and filtering of barley contaminant sequences. Relates to Figure 5.

Protein Family	median D_N/D_S [‡]	Biological Function [§]	sample p-value [†] M1	W1	L1	M2	W2	L2	genotype p-value
Wild genotype									
TIGR02780	1,21	P-type conjugative transfer protein TrbJ		***					2,05E-03
TIGR00049	1,41	iron-sulfur cluster assembly accessory protein		*					2,07E-03
TIGR00916	1,93	protein-export membrane protein, SecD/SecF family		***	***				9,63E-03
TIGR01543	1,50	phage prohead protease, HK97 family		***					9,63E-03
TIGR03426	1,41	rod shape-determining protein MreD		***					1,32E-02
TIGR02118	1,72	TIGR02118: conserved hypothetical protein	***	***	*	*	***	+	1,79E-02
TIGR02464	1,69	conserved hypothetical protein		***					4,22E-02
TIGR01552	1,59	prevent-host-death family protein		***	***	*	**	*	4,46E-02
TIGR00613	1,52	DNA repair protein RecO		*					4,46E-02
TIGR00616	1,91	recombinase, phage RecT family					**		4,69E-02
Modern genotype									
TIGR01128	0,71	DNA polymerase III, delta subunit	**						1,28E-12
TIGR00225	1,37	C-terminal processing peptidase	**						1,28E-05
TIGR03358	1,84	type VI secretion protein, VC_A0107 family	***					***	1,15E-02
TIGR01216	1,47	ATP synthase F1, epsilon subunit	+			***			1,39E-02
Landrace genotype									
TIGR03355	1,35	type VI secretion protein, EvpB/VC_A0108 family						***	4,32E-08
TIGR03197	1,37	tRNA U-34 biosynthesis protein MnmC			**				7,61E-04
TIGR03930	2,43	WXG100 family type VII secretion target	***	***	***	***		***	7,86E-03
TIGR04085	2,47	radical SAM additional 4Fe4S-binding SPASM domain	***	***	***	***	***	***	1,18E-02
TIGR01382	1,52	intracellular protease, PfpI family			*				1,95E-02
TIGR03544 [¶]	1,78	DivIVA domain		*	***	***	*	***	4,30E-02
TIGR00026	1,69	deazaflavin-dependent oxidoreductase	**		***	**		***	4,96E-02

[†] based on one-sided Fisher's exact test with 5% FDR

[‡] Non-synonymous / synonymous substitution rate

[§] based on TIGR annotation

[¶] high abundant, based on ANOVA analysis

+=p <0.1

* =p <0.05

** =p <0.01

** =p <0.001

Table S3. Protein families with evidence for significantly enhanced positive selection in one of the three genotypes (wild, modern, landrace). Of the 115 protein families with enhanced signs of positive selection found in one of the genotype comparisons, the 21 protein families that also showed significant signs of positive selection in at least one sample are shown. Families with repetitiveness values of more than 50% were not considered. Relates to Figure 6.

			sample p-value†					
Protein Family	median D_N/D_S‡	Biological Function§	M1	W1	L1	M2	W2	L2
TIGR03357	2.36	VI_zyme: type VI secretion system lysozyme-like protein	***	+	***	*		***
TIGR02511	2.16	type_III_tyeA: type III secretion effector delivery regulator, TyeA family						*
TIGR04183	1.98	Por_Secre_tail: Por secretion system C-terminal sorting domain	***	***	***	***	***	***
TIGR03347	1.97	VI_chp_1: type VI secretion protein, VC_A0111 family	*	+	**			
TIGR03344	1.94	VI_effect_Hcp1: type VI secretion system effector, Hcp1 family	+			***		*
TIGR03358	1.84	VI_chp_5: type VI secretion protein, VC_A0107 family	***					***
TIGR01707	1.80	gspI: type II secretion system protein I	**			*		
TIGR03354	1.78	VI_FHA: type VI secretion system FHA domain protein		*				
TIGR03349	1.73	IV_VI_DotU: type IV/VI secretion system protein, DotU family		*				
TIGR03363	1.62	VI_chp_8: type VI secretion-associated protein, ImpA family		*				
TIGR03355	1.35	VI_chp_2: type VI secretion protein, EvpB/VC_A0108 family						***

† based on Fisher's test with a 5% FDR

‡ Non-synonymous / synonymous substitution rate

§ based on TIGR annotation

+ = p <0.1
* = p <0.05
** = p <0.01
** = p <0.001

Table S4. Protein families of bacterial secretion systems found to be under positive selection (with significant D_N/D_S statistic) in one or more samples and a maximal repetitiveness value less than 70%. Relates to Figure 6.

Putative elicitors from McCann et al.	semantic overlap
GGDEF domain/EAL domain protein‡	GGDEF: diguanylate cyclase (GGDEF) domain
Radical SAM domain protein	rSAM_more_4Fe4S: radical SAM additional 4Fe4S-binding SPASM domain
ExoDNase I SbcB	exoDNase_III: exodeoxyribonuclease III
YjeF-related protein	yjeF_nterm: YjeF family N-terminal domain
Capsular polysaccharide biosynthesis protein‡	eps_fam: capsular exopolysaccharide family
Thiazole synthase ThiG	ThiI_C_thiazole: thiazole biosynthesis domain
Cytochrome C oxidase assembly protein CtaG	ccoO: cytochrome c oxidase, cbb3-type, subunit II
DNA repair protein RecN	recO: DNA repair protein RecO
Acetyl-CoA acetyltransferase PhbA-2	AcCoA-C-Actrans: acetyl-CoA C-acetyltransferase

† based on TIGRFAMS with significant high D_N/D_S values on genotype and sample comparison

‡ Significant P value ($P < 0.05$) of the LRT between models M7 and M8 for nonpathogen genomes.

Table S5. Semantic overlap of proteins families under positive selection with a significant D_N/D_S statistic in one or more sample compared to the whole dataset and the putative elicitors reported by McCann and co-workers (Case et al., 2007). Relates to Figure 6.

			p-value[a]						
Protein Family	median D_N/D_S [b]	Biological Function[c]	M1	W1	L1	M2	W2	L2	Gene Ontolotgy
TIGR01614	2,61	pectinesterase inhibitor domain			*	+	***		
TIGR02266	2,57	Myxococcus xanthus paralogous domain	***	***			**		
TIGR03761	2,11	integrating conjugative element protein, PFL_4669 family		***					
TIGR00377	1,94	anti-anti-sigma factor	***			***		**	GO:0006355
TIGR01843	1,77	type I secretion membrane fusion protein, HlyD family							GO:0008565
TIGR00687	1,58	pyridoxal kinase							GO:0008478
TIGR00792	1,52	sugar (Glycoside-Pentoside-Hexuronide) transporter							GO:0006812
TIGR00556	1,52	phosphopantetheine-protein transferase domain	*	*					
TIGR00014	1,28	arsenate reductase (glutaredoxin)							GO:0008794
TIGR00223	1,27	aspartate 1-decarboxylase							GO:0004068
TIGR00678	1,11	DNA polymerase III, delta' subunit							GO:0003887
TIGR02224	0,89	tyrosine recombinase XerC							GO:0006310
TIGR02541	0,77	flagellar rod assembly protein/muramidase FlgJ							GO:0001539
TIGR03691	0,73	proteasome, alpha subunit							GO:0004298
TIGR02211	0,51	lipoprotein releasing system, ATP-binding protein							GO:0005524
TIGR00263	0,08	tryptophan synthase, beta subunit							GO:0000162

[a] based on one-sided Fisher's exact test with 5% FDR
[b] Non-synonymous / synonymous substitution rate
[c] based on TIGR annotation

+ = p <0.1
* = p<0.05
**=p<0.01
***=p<0.001

Table S6. Predicted putative Type III effectors. Candidate effectors were found using the EffectiveT3 tool on the consensus sequence of the multiple sequence alignment. The standard classification setting (trained with all effector sequences as described in (Jehl et al., 2011) and standard parameters (minimum score cut-off = 0.9999) were used. Relates to Figure 6.

Supplemental Figures

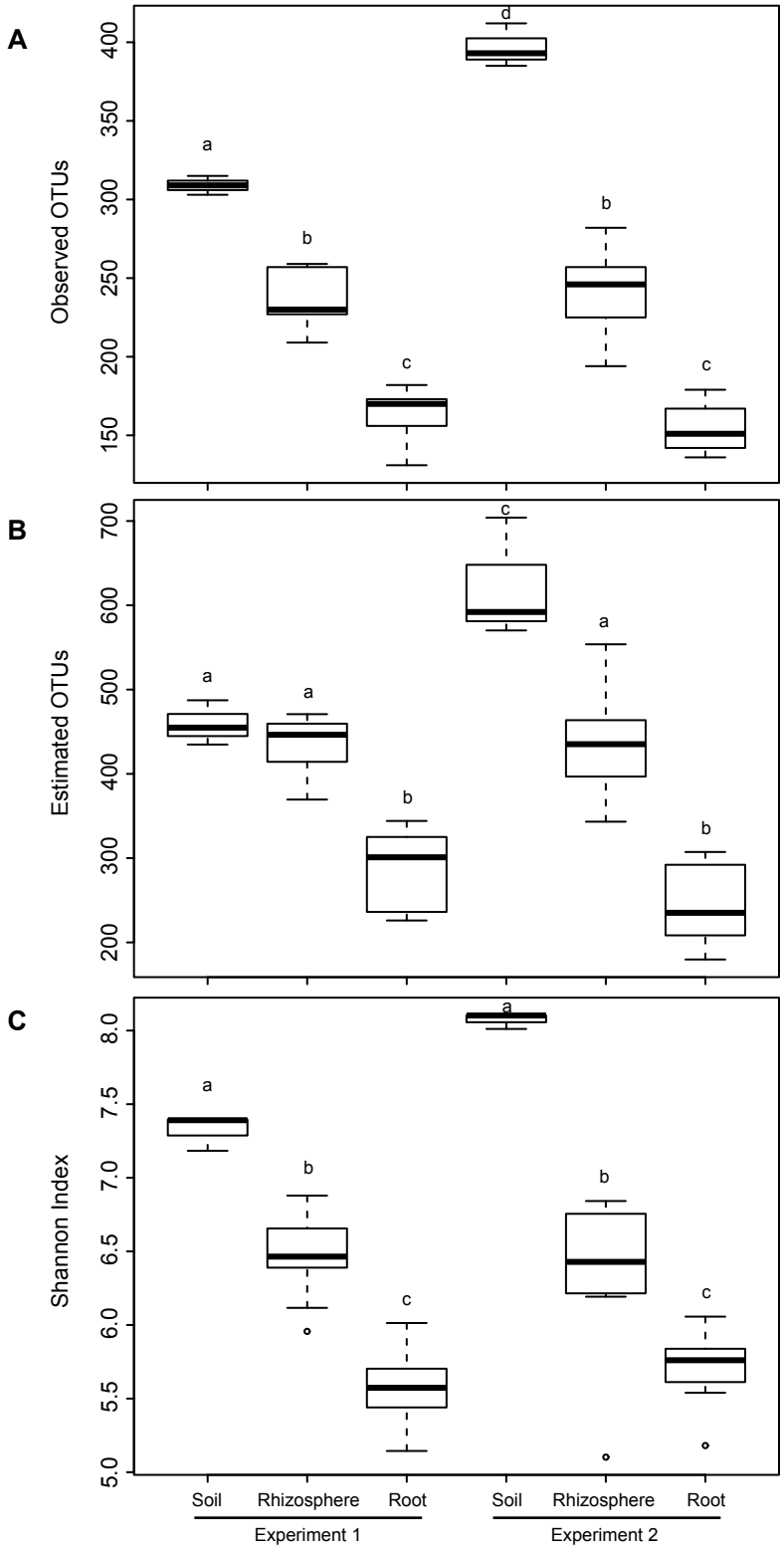
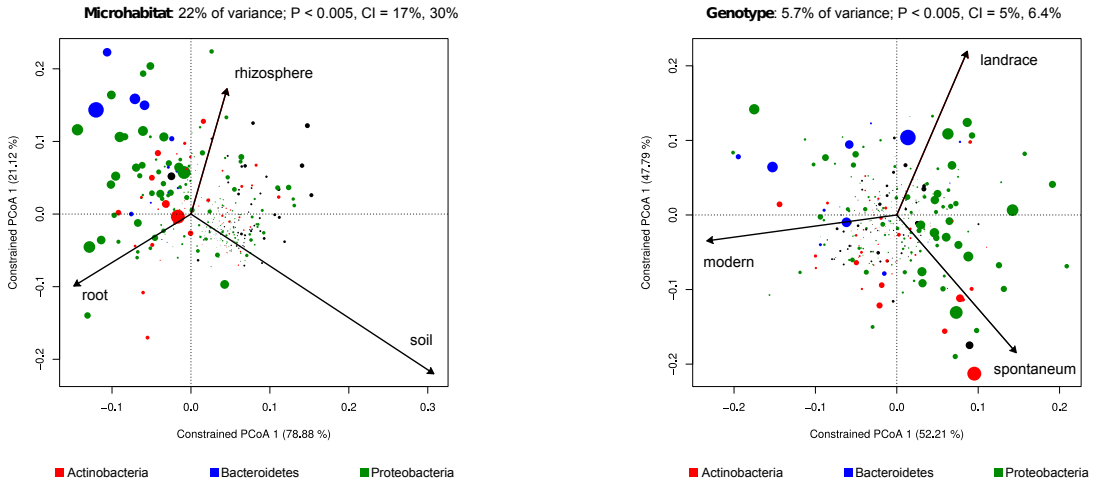
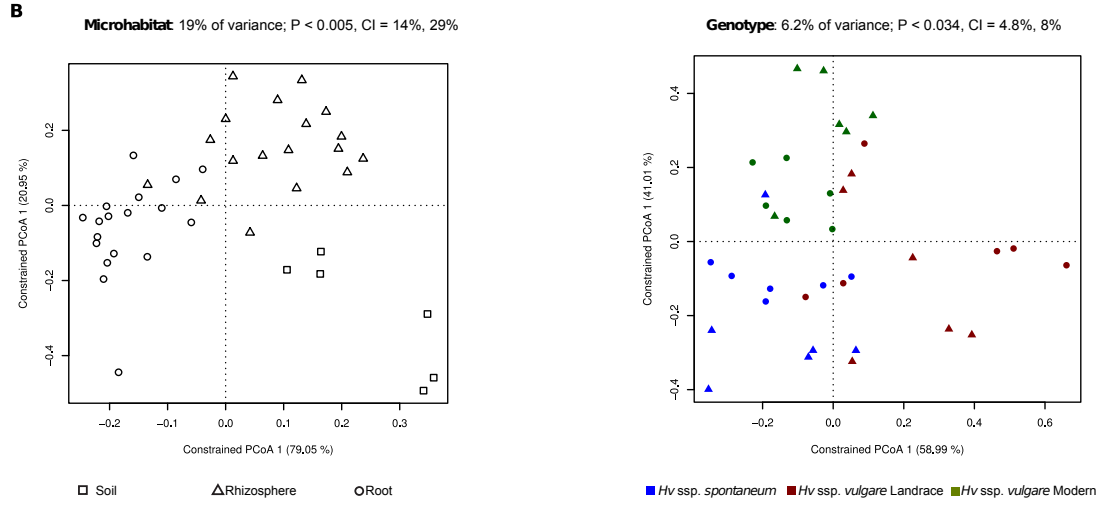


Figure S1. Alpha-diversity calculation of the soil, rhizosphere and root samples. (A) Total number of observed OTUs, (B) Chao1 estimator and (C) Shannon's diversity index. Samples were rarefied to 1,000 reads prior the analysis. Different letters denote statistically significant differences by Tukey test at $p<0.05$. Relates to Figure 2.

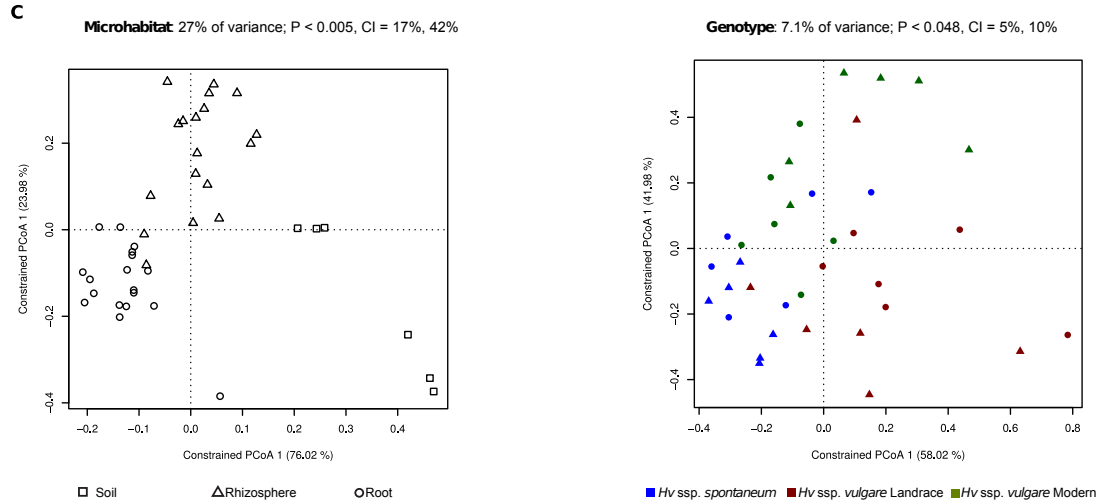
A


Microhabitat: 22% of variance; P < 0.005, CI = 17%, 30%
rhizosphere
root
soil
Constrained PCoA 1 (21.12 %)
Constrained PCoA 1 (78.88 %)
Actinobacteria
Bacteroidetes
Proteobacteria
Genotype: 5.7% of variance; P < 0.005, CI = 5%, 6.4%
landrace
modern
spontaneum
Constrained PCoA 1 (47.79 %)
Constrained PCoA 1 (52.21 %)
Actinobacteria
Bacteroidetes
Proteobacteria

B


Microhabitat: 19% of variance; P < 0.005, CI = 14%, 29%
Constrained PCoA 1 (20.95 %)
Constrained PCoA 1 (79.05 %)
Soil
Rhizosphere
Root
Genotype: 6.2% of variance; P < 0.034, CI = 4.8%, 8%
Constrained PCoA 1 (41.01 %)
Constrained PCoA 1 (58.99 %)
Hv ssp. *spontaneum*
Hv ssp. *vulgare* Landrace
Hv ssp. *vulgare* Modern

C


Microhabitat: 27% of variance; P < 0.005, CI = 17%, 42%
Constrained PCoA 1 (23.98 %)
Constrained PCoA 1 (76.02 %)
Soil
Rhizosphere
Root
Genotype: 7.1% of variance; P < 0.048, CI = 5%, 10%
Constrained PCoA 1 (41.98 %)
Constrained PCoA 1 (58.02 %)
Hv ssp. *spontaneum*
Hv ssp. *vulgare* Landrace
Hv ssp. *vulgare* Modern

D

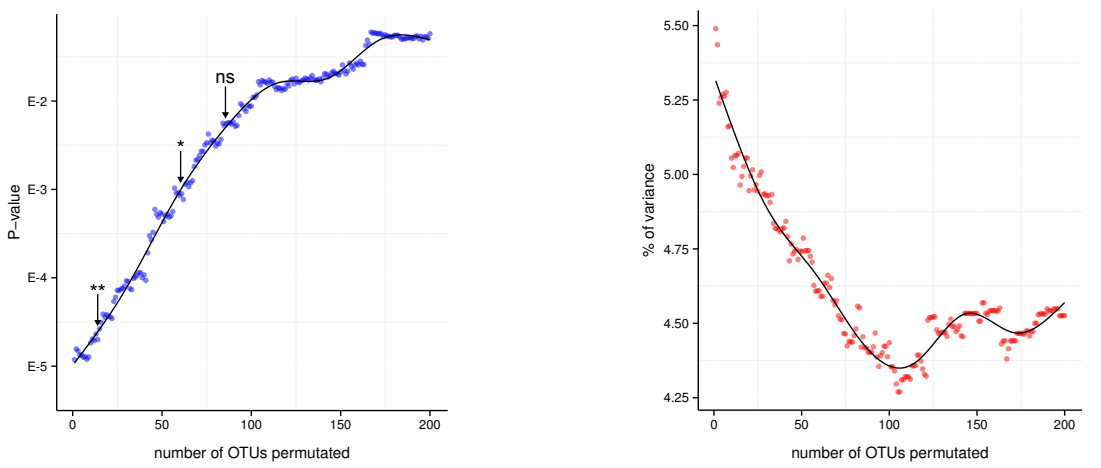

ns
*
**
P−value
E-2
E-3
E-4
E-5
number of OTUs permutated
% of variance
number of OTUs permutated

Figure S2. (A) OTU scores of PCoA analysis. The arrows point to the centroid of the constrained factor: microhabitat (left) and genotype (right). Circle size depicts to the relative abundances of each OTU (log scale) and colors illustrate different phyla. The percentage of variation explained by each axis refers to the fraction of the total variance of the data explained by the constrained factor. **(B) Constrained Principal coordinate analysis (PCoA) analysis based on weighted UniFrac distances,** constrained by microhabitat (24 % of the overall variance; P < 5.00E-2, 5,000 permutations) and by accession (5.8 % of the overall variance; P < 5.00E-2, 5,000 permutations). **(C) Constrained Principal coordinate analysis (PCoA) analysis based on unweighted UniFrac distances,** constrained by microhabitat (9% of the overall variance; P < 5.00E-2, 5,000 permutations) and by accession (5.5 % of the overall variance; not significant). **(C) Permutation analyses of the constrained ordination.** Analysis of the impact of randomly permutating the most relevant OTUs (ranked by their contribution to the ordination space) on the significance of the genotype effect (left) and on the fraction of overall variance of the data explained by the projection (right). Relates to Figure 2.

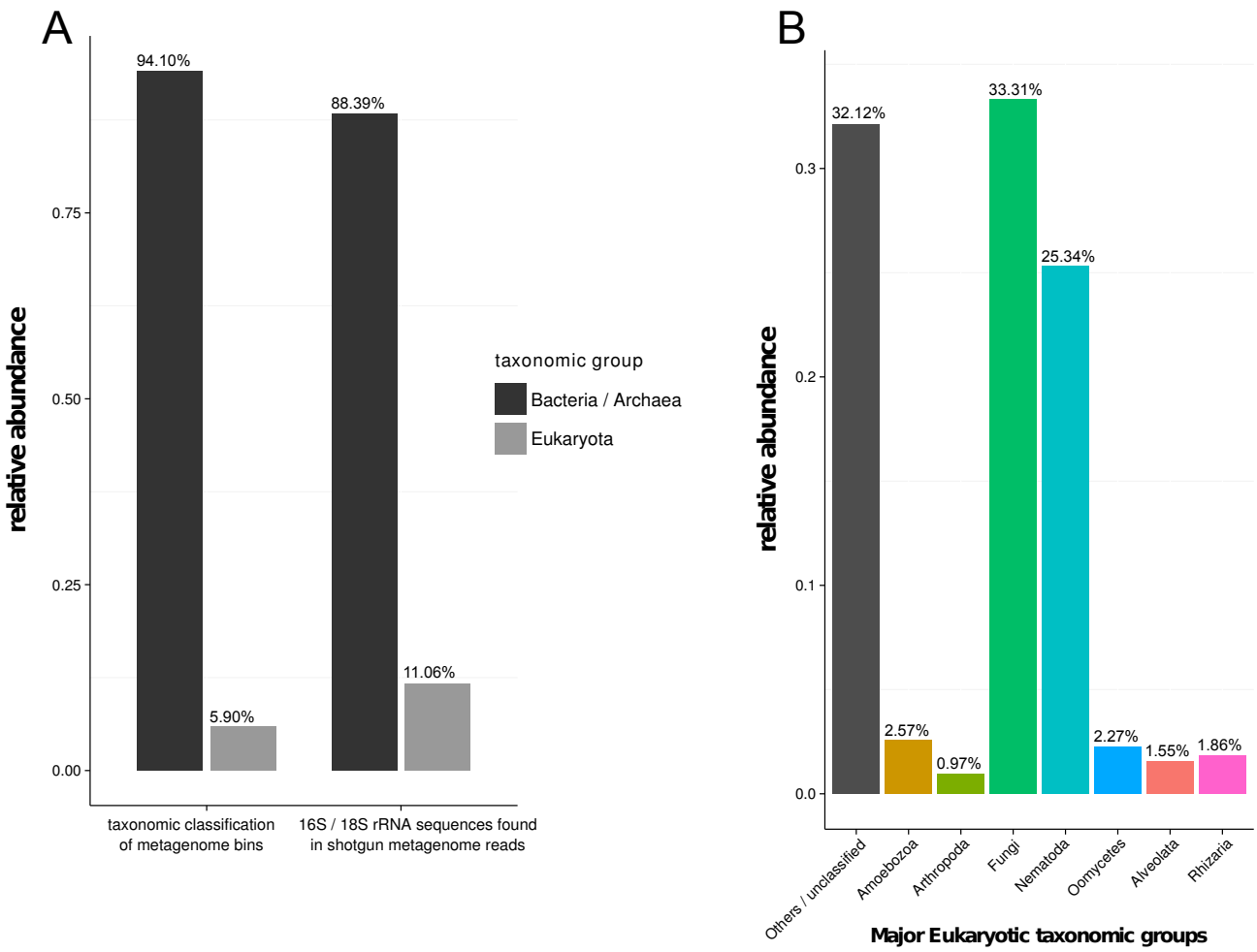


Figure S3. Analysis of Eukaryotic abundances. Comparison of Bacterial to Eukaryotic abundances estimated by taxonomic classification of metagenome bins and comparison of 16S and 18S rRNA gene reads found in the shotgun metagenome reads (A). Relative abundance comparison between major Eukaryotic taxonomic groups found in the barley rhizosphere metagenome by analysis and classification of 18S rRNA sequences (B). Relates to Figure 5.

Supplemental Experimental Procedures

Experimental design

The soil substrates used in this investigation were collected at the Max Planck Institute of Molecular Plant Physiology (52.416 N/ 12.968 E, Potsdam –Golm, Germany) in September 2010 and September 2011 stored and prepared for use as previously described (Bulgarelli et al., 2012) . The geochemical characterization, as obtained from the 'Labor für Boden- und Umweltanalytik' (Eric Schweizer AG, Thun, Switzerland) is provided in Table S1.

Seeds of the *H. vulgare* ssp. *spontaneum* accession "HID369" were kindly provided by Prof. Marteen Koorneef, Department of Plant Breeding and Genetics, Max Planck Institute for Plant Breeding Research, Cologne, Germany. Seeds of the *H. vulgare* ssp. *vulgare* cultivar "Rum" and cultivar "Morex" were kindly provided by Prof. Maria von Korff, Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany. "HID 369" is a barley wild accession collected in the Mount Meron region, Israel (Korneef M, personal communication), "Rum" is a landrace traditionally cultivated in Jordan (Samarah et al., 2009) while "Morex" is a cultivated malt variety developed in the United States. These materials were chosen since "Morex" is one of the reference genotype for barley genetic and genomic investigations (Mayer et al., 2012) and for the availability of recombinant inbred lines and double haploids populations existing among these genotypes (Korneef M., von Korff M., personal communications).

Before planting, seeds were surface-sterilized with a combination of bleach and ethanol treatments as previously described (Bulgarelli et al., 2010). Surface sterilized

seeds were sown onto 9 cm diameter plastic pots filled with experimental soil, which were placed at 4°C in the dark for stratification during 5 days prior to relocation to the cultivation greenhouse. We grew the plants under long day conditions 16 hours light (day) and 8 hours dark (night), 20°C during the day and 18°C during the night at a relative humidity of 70 %. After germination, barley seedlings were thinned to one plant per pot and transferred for three weeks in a climatic chamber under long day conditions (16 hours light and 8 hours dark) at 4°C to synchronise the development of wild and cultivated accessions. After this vernalisation treatment, pots were transferred in the cultivation greenhouse and the plants were maintained for additional four weeks, under the aforementioned growth conditions, until all plant genotypes reached the early stem elongation developmental stage. Unplanted pots were subjected to the same conditions as the planted pots to prepare the control soil samples at harvest. For each barley accession and unplanted soil, three biological replicates, defined as individual pot, were processed. The entire experiment has been performed twice using two distinct samplings of the experimental soil.

Preparation of the metagenomic DNA from soil, rhizosphere and root samples.

Roots were separated from the adhering soil particles and the defined root segment of 6 cm length starting 0.5 cm below the root base was harvested. Only seminal roots were included in the analysis and, when present, nodal roots have been excised from the root system before downstream processing. Roots were collected in 50 ml falcons containing 10 ml PBS-S buffer (130 mM NaCl, 7 mM Na2HPO4, 3 mM NaH2PO4, 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 subjected to a second washing treatment (20 minutes at 180 rpm in 3 ml PBS-S buffer).

Doubled-washed roots were then 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 microbes thriving in close association with root tissues (Fig. S2). Roots were removed from PBS-S, rinsed in a fresh volume of 10 ml PBS-S buffer and grinded with mortar and pestle in liquid nitrogen. Pulverised roots were collected in 15 ml falcon tubes and stored at -80°C until further processing. In parallel, a subset of soil-grown root samples subjected to double washing only or double washing and sonication was used to perform a scanning electronic microscopy investigation of the rhizoplane as previously described (Bulgarelli et al., 2012). The soil suspensions collected in the falcon tubes after the first and second washing treatments were combined, centrifuged at 4,000*g* for 20 minutes and the pellet, referred to as the rhizosphere, was frozen in liquid nitrogen and stored at -80°C until further processing. Soil samples were collected from unplanted pots in a soil depth of -0.5 to -6.5 from the surface corresponding to 6 cm root length, frozen in liquid nitrogen and stored at -80°C until further processing. Total DNA was extracted with the FastDNA® SPIN Kit for Soil (MP Biomedicals, Solon, USA) following the manufacturer's instructions. Samples (pulverised roots, rhizospheric and unplanted soils) 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. DNA samples were eluted in 80 µl DES water and DNA concentrations were determined using the NanoDrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, USA).

16S rRNA gene amplicon library preparation and pyrosequencing

Amplicon libraries were generated using the PCR primers 799F (5'-AACMGGATTAGATACCCKG-3')(Chelius and Triplett, 2001) and 1193R (5'-ACGTCATCCCCACCTTCC-3') (Bodenhausen et al., 2013) spanning ~400 bp of the hypervariable regions V5-V7 of the prokaryotic 16S rRNA gene. For multiplexed pyrosequencing we utilized the 799F primer fused at the 5‘ end with a sample specific (Database S1), error-tolerant 6-mer barcode (N´s) followed by a SfiI 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 10ng/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 MgCl2, 200 µM of dNTPs and 400 nM of each fusion primer. The PCR reactions were assembled in a laminar flow and amplified using the touch-down protocol described in Database S1. To minimize stochastic PCR effects samples were amplified with 3 independently pipetted mastermixes in triplicate reactions per mastermix. Triplicate reactions of each sample were pooled per mastermix and a 10 µl aliquot inspected on a 1.5% agarose gel for the lack of detectable 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 to separate the ~450bp 16S rRNA gene amplicon from the ~800bp 18S rRNA gene amplicon typically generated by the PCR primers 799F and 1193R (Bulgarelli et al., 2012) The smaller PCR product was excised from the agarose gel and purified 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 NanoDrop 1000 Spectrophotometer (Thermo Scientific, Wilmington, USA). Purified PCR products were pooled in equimolar amounts and concentrated using the QIAamp DNA Micro Kit (Qiagen, Hilden, Germany). Ligation of the amplicon pools to 454 adapters, emulsion PCR and pyrosequencing were performed at the Max Planck Genome Centre in Cologne (http://mpgc.mpipz.mpg.de/home/) as previously described (Schlaeppi et al., 2014).

Alpha and betadiversity analysis on the 16S rRNA gene dataset

We used the command *alpha_diversity.py* in QIIME to determine the Shannon (OTU evenness) and the Chao1 (OTU richness) indices as well as the total number of observation on the OTU table rarefied at 1,000 counts per sample. Statistical analysis (ANOVA, TukeyHSD) were performed in R. The on average index values were corrected for microhabitat and experiment.

Betadiversity calculations were computed using non-rarefied OTU counts. Bray-Curtis, weighted and unweighted UniFrac dissimilarity matrices were constructed in QIIME using the command *beta_diversity.py* on ‰ OTU relative abundances $\log_2$ transformed. Only OTUs with a relative abundance above 5 ‰ in at least one sample were included in the analysis. Permutational multivariate analyses of variance were

performed in R using the function *adonis.* Constrained principal coordinates analyses were performed in R using the log_2 transformed OTU table. We used the function capscale, constraining by the environmental variables microhabitat and host genotype. Statistical significance of the ordinations as well as confidence intervals for the variance was determined by an ANOVA-like permutation test (functions permutest and anova.cca) with 5,000 permutations. All the R functions used for the beta diversity calculations were retrieved from the R package vegan v2.0.8 (Dixon, 2003) .

To assess the influence of individual OTUs on the observed genotype effect, we first ranked the OTUs based on their relative contributions to the ordination space. We then randomly permutated the abundances of each OTU and repeated the analysis for each bootstrap sample (100 repetitions). To assess the contribution of individual OTUs to the significance of the effect we averaged the p-values and the percentage of variance explained across repetitions.

Statistical analysis on taxa and OTU counts

To identify taxa (Fig. 1) and OTUs (Fig. 3) enriched in rhizosphere and root microhabitats compared to unplanted soil we employed linear statistics on RA values (log2, > 5 ‰ threshold) using a custom script developed from the R package Limma. Differentially abundant taxa and OTUs between two microhabitats were calculated using moderated t-tests. The resulting p-values were adjusted for multiple hypotheses testing using the Benjamini-Hochberg correction. Ternary plots were constructed as previously described (Bulgarelli et al., 2012) .

Taxonomic comparison of the barley and Arabidopsis root enriched microbiota

We retrieved the sequences of soil, rhizosphere and root samples of Arabidopsis thaliana grown in the same soil type used in the Barely survey from our former database (Bulgarelli et al., 2012). Sequences reads were subjected to the same UPARSE pipeline adopted for the barley sequences. Note that upon UPARSE processing and in silico depletion of reads assigned to Chloroflexi, only the 30 samples containing at least 1,000 high quality sequencing reads have been included in the downstream analysis. Arabidopsis root enriched OTUs were determined using linear statistics on RA values (log2, > 5 ‰ threshold). To compare the taxonomic distribution of OTUs enriched in barley and Arabidopsis roots and their relative abundances within their respective communities we first used MOTHUR (Schloss et al., 2009) to assign NCBI taxonomy IDs to each of OTU representative sequences. We then generated a tree based on the taxonomic relationships of the different OTUs within the NCBI database. The tree-plot was generated by mapping the log-transformed relative abundance of each root-enriched OTUs onto the reduced NCBI tree using the MOTHUR taxonomic assignments.

Shotgun sequencing of the rhizosphere DNA preparations

Total DNA prepared from rhizosphere samples were combined in a sample- and experiment wise-manner and were sheared to an average size of 200 bp (COVARIS, Woburn, USA). Sequencing libraries were prepared from fragmented DNA according to the suppliers' recommendations (TruSeq DNA sample preparation v2 guide, Illumina, San Diego, USA). Libraries were quantified by fluorometry, immobilized and processed onto a flow cell with a cBot (Illumina, San Diego, USA) followed by sequencing-by-synthesis with TruSeq v3 chemistry on a HiSeq2500 (100bp Paired-end, Illumina, San Diego, USA).

Metagenome quality filtering and assembly

Raw paired-end Illumina reads (99 bp average read length) were processed using custom scripts and the CLC Genomics Workbench v5.5.1 for adapter removal, ambiguity, length (reads <60 bp) and quality trimming (15 Phred score; 0.03 Solexa scale). High-quality reads of each sample were assembled using the SOAP-denovo assembler (Heger and Holm, 2000) with the metagenomics model and default parameters (command: *SOAPdenovo-31mer all -s soapdenovo.config -K 23 -R -p 40*).

Barley sequence filtering

Genomic sequences for the Barke, Morex and Bowman barley cultivars were downloaded from ftp://ftpmips.helmholtz-muenchen.de/plants/barley/public_data/barley_data_archive_21Nov12.tgz. Assembled contigs and unassembled singleton reads from the six rhizosphere metagenome samples were mapped to barley annotated genomic sequences with the BWA-MEM program (Li, 2013) with default parameters settings, allowing to find partial matches. Sequences with a mapping score larger than 13 (allowing for a 0.05 % incorrect alignment) were considered to represent potential barley sequences and removed from the dataset. When analysing the eukaryotic diversity present in the barley rhizosphere, all remaining sequences classified as belonging to the Poales order by *taxator-tk* (see below, 3.09% of reads), were assumed to be residual barley contaminants and also subsequently removed. For analysis of bacterial and archaeal diversity and functional enrichment, all contigs and singleton reads not classified as belonging to either domain were not included.

Taxonomic assignment

The partially assembled metagenome sequences (including reads which were not part of larger contigs) was taxonomically classified with *taxator-tk* (Droge et al., 2014). The nonredundant reference sequence collection used for the taxonomic assignment with *taxator-tk* was generated from the following resources: ncbi-refseq-microbial_56, ncbi-draftgenomes-bacteria_sequences from 22/11/2012, ncbi-genomes-bacteria from 22/11/2012 and ncbi-hmp from 16/10/2012. The software uses several passes of sequence similarity searches for estimation of a robust set of the closest evolutionary neighbors for individual regions of a sequence and assignment of a taxonomic identifier for the overall sequence using their lowest common ancestor in the NCBI taxonomy. Taxonomic annotations with *taxator-tk* have low error rates and allow an extensive taxonomic profiling of a metagenome simultaneously for Bacteria, Archaea and Eukaryotes without PCR primer biases or biases introduced by marker gene copy number variations. Relative abundances were calculated by mapping the reads back to the assembled contigs and determining the number of reads assigned to each taxon.

Gene prediction and functional annotation

Complete and partial coding sequences (CDSs) for protein encoding genes were predicted with MetaGeneMark (Zhu et al., 2010). Contigs with less than 100bps were discarded as prediction accuracy is low for very short sequences (Trimble et al., 2012) Because of the size of the data set (more than 150000 CDSs in total) a fast and reliable probabilistic method, namely profile HMMs, was used for sequence homology detection: CDSs were annotated based on matches to the profile Hidden

Markov Models (HMMs) to the TIGRFAM 13.0 (Haft et al., 2013) and the Pfam 27.0 databases (Punta et al., 2012) using the hmmsearch command of HMMER 3.0 (Eddy, 2009). Matches to Pfam and TIGRFAM HMMs with an E-value of at least 0.01, a bit score of at least 25 and additionally exceeding the gathering threshold (-ga option of the hmmsearch command) were further considered. Each Pfam and TIGRFAM family has such a manually defined gathering threshold for the bit score that was set by the curator of the protein family. Per sample on average 25811 CDSs were assigned to 915 families. To annotate the CDSs with SEED subsystems we applied a *k*-mer based matching (Edwards et al., 2012; Overbeek et al., 2005). The CDSs were first mapped to FIGFAMS by searching *k*-mers of length 9 and requiring at least two matching *k*-mers at most 600 bp apart. The FIGFAMs were mapped to subsystems via their functional roles.

Metagenomic SSU rRNA gene profiling

We joined the paired-end reads that overlapped by at least 5 bp. If there was less or no overlap we inserted 20 ambiguous nucleotides (N's), which approximately corresponded to the mean gap size between pairs of reads.

The joined paired-end reads were searched for hits of 16S and 18S rRNA gene family with Meta RNA (version HMMER 3.0) using default settings (Huang et al., 2009). We used MOTHUR (Schloss et al., 2009) to assign NCBI taxonomy identifiers to the 16S rRNA sequences.

Identification of differential functional enrichment

To test for enrichment of functions in bacterial taxa associated to *RR_OTUs* we retrieved family-level taxonomic bins, which included sequences assigned to lower

ranks within a family. We then determined which of these bins corresponded to the same families as the rhizosphere and root-enriched *RR_OTUs* and calculated abundances of functional categories for each bin (SEED subsystems level 2). Finally, we employed a non-parametric statistical test (Mann-Whitney) to test for a significant enrichment of functional categories in the rhizosphere bins relative to the remaining bins, controlling for false discovery rate (FDR) using the Benjamini and Hochberg procedure.

To compare with functional enrichment in sequenced isolates, we took the same taxonomic groups used for comparisons within the metagenomes and retrieved all their sequenced isolates from the NCBI database of microbial genomes (accessed on 15/01/2014). In total, we downloaded and annotated 1,233 bacterial genomes belonging to both, the soil as well as the root-associated taxa, and conducted the exact same statistical analysis as previously described.

Estimation of non-synonymous and synonymous substitution rates

CDSs were annotated based on their protein family membership using HMMER with the TIGRFAM 13.0 database (see Gene prediction and functional annotation). With the hmmalign command of HMMER 3.0 a multiple sequence alignment (MSA) of the sample protein sequences was created using each TIGRFAM protein family. Based on the MSA and the CDS nucleotide sequences, a codon alignment was constructed for each protein family with pal2nal 14 (Suyama et al., 2006) using default parameters. We then used clearcut, a relaxed neighbour joining algorithm (Evans et al., 2006; Saitou and Nei, 1987) to reconstruct a phylogenetic tree for each protein family from the obtained MSA. Neighbour joining instead of the slower maximum

likelihood methods was used because of the size of the data set. For this, additive pairwise distances were calculated with a slightly modified version of clearcut, were gaps were not counted as mismatches. This was done because the gaps in the alignments were mostly of technical origin, caused by the alignment of short contigs to longer reference sequences. Using an in-house tool (phylorecon), amino acid sequences and coding sequences were reconstructed for the internal nodes of each protein family tree using maximum parsimony as a criterion and the numbers of synonymous (D_s) and non-synonymous (D_n) changes were inferred with correction for multiple substitutions as in (Tusche et al., 2012). We excluded low-confidence positions in the alignment with a large number of gaps (more than 50%) for the calculation of the mean D_n/D_s value for each protein family. A one-sided Fisher's test was performed to identify protein families with a significant enrichment of D_n versus D_s changes in comparison to the entire sample. The false discovery rate (FDR) was controlled using the Benjamini and Hochberg procedure and alpha set to 5%. The ~dN/dS was then approximated for every set of sequences found in the metagenome samples belonging to the same protein family as $(D_n/N) / (D_s/S)$, or $(D_n/D_s) / (S/N)$ (Pond and Muse, 2005), where the ratio of synonymous (S) to nonsynonymous (N) sites was set to 0.5. Complex repeat architectures and gapped approximate repeats were detected with RADAR (Heger and Holm, 2000) on the MSA consensus sequences. The quotient of the number of sites annotated with a repeat element and the length of the consensus sequence served as measure of repetitiveness. This quotient was used for filtering of protein families that had a large number of repeat elements.

Detection of clusters with a significant D_n/D_s *statistic*

To determine sequence regions of the proteins families with significant signs of positive selection, D_N/D_S was calculated based on the number of nonsynonymous to synonymous substitutions by sliding a window of 10 amino acids in length over the MSA. A one-sided Fisher's test of exact probability was performed for each window and the FDR controlled with the Benjamini and Hochberg procedure with alpha set to 5 %. Each window with a corrected p-value lower than 0.05 and a mean gap ratio in this window lower than 60% was reported as significant and neighboring windows merged into clusters, corresponding to sites on the consensus sequence with significantly higher D_n/D_s statistic compared to the rest of the sample.

Prediction of transmembrane helices, effector proteins and secreted proteins

Transmembrane helices in the consensus MSA sequence were predicted with TMHMM v. 2.0) (Sonnhammer et al., 1998) using default parameters. Jensen-Shannon divergence was calculated for every position of the sequence with conservation_code (Capra and Singh, 2007). Bacterial secreted proteins were predicted with EffectiveT3 (Jehl et al., 2011) with the standard set classification module and selective cut-off (0.9999) setting.

Supplemental References

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

Bulgarelli, D., Biselli, C., Collins, N.C., Consonni, G., Stanca, A.M., Schulze-Lefert, P., and Vale, G. (2010). The CC-NB-LRR-type Rdg2a resistance gene confers immunity to the seed-borne barley leaf stripe pathogen in the absence of hypersensitive cell death. PLoS One *5*.

Bulgarelli, D., Rott, M., Schlaeppi, K., Ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., Rauf, P., Huettel, B., Reinhardt, R., Schmelzer, E., *et al.* (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. Nature *488*, 91-95.

Capra, J.A., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. Bioinformatics *23*, 1875-1882.
Case, R.J., Boucher, Y., Dahllof, I., Holmstrom, C., Doolittle, W.F., and Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol *73*, 278-288.
Chelius, M.K., and Triplett, E.W. (2001). The Diversity of Archaea and Bacteria in Association with the Roots of Zea mays L. Microbial ecology 41, 252-263.
Dixon, P. (2003). VEGAN, a package of R functions for community ecology. J Veg Sci *14*, 927-930.
Droge, J., Gregor, I., and McHardy, A.C. (2014). Taxator-tk: Precise Taxonomic Assignment of Metagenomes by Fast Approximation of Evolutionary Neighborhoods. Bioinformatics.
Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. Genome informatics International Conference on Genome Informatics *23*, 205-211.
Edgar, R.C. (2013). UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods *10*, 996-998.
Edwards, R.A., Olson, R., Disz, T., Pusch, G.D., Vonstein, V., Stevens, R., and Overbeek, R. (2012). Real Time Metagenomics: Using k-mers to annotate metagenomes. Bioinformatics *28*, 3316-3317.
Evans, J., Sheneman, L., and Foster, J. (2006). Relaxed neighbor joining: a fast distance-based phylogenetic tree construction method. Journal of molecular evolution *62*, 785-792.
Haft, D.H., Selengut, J.D., Richter, R.A., Harkins, D., Basu, M.K., and Beck, E. (2013). TIGRFAMs and Genome Properties in 2013. Nucleic acids research *41*, D387-395.
Heger, A., and Holm, L. (2000). Rapid automatic detection and alignment of repeats in protein sequences. Proteins *41*, 224-237.
Huang, Y., Gilna, P., and Li, W. (2009). Identification of ribosomal RNA genes in metagenomic fragments. Bioinformatics 25, 1338-1340.
Jehl, M.A., Arnold, R., and Rattei, T. (2011). Effective--a database of predicted secreted bacterial proteins. Nucleic acids research *39*, D591-595.
Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ARXIV *eprint arXiv:1303.3997*.
Mayer, K.F.X., Waugh, R., Langridge, P., Close, T.J., Wise, R.P., Graner, A., Matsumoto, T., Sato, K., Schulman, A., Muehlbauer, G.J.*, et al.* (2012). A physical, genetic and functional sequence assembly of the barley genome. Nature *491*, 711-+.
Overbeek, R., Begley, T., Butler, R.M., Choudhuri, J.V., Chuang, H.Y., Cohoon, M., de Crecy-Lagard, V., Diaz, N., Disz, T., Edwards, R.*, et al.* (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic acids research *33*, 5691-5702.
Pond, S.L.K., and Muse, S.V. (2005). HyPhy: Hypothesis testing using phylogenies. Statistical Methods in Molecular Evolution, 125-181.
Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012). The Pfam protein families database. Nucleic acids research *40*, D290-301.
Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution *4*, 406-425.

Samarah, N.H., Alqudah, A.M., Amayreh, J.A., and McAndrews, G.M. (2009). The Effect of Late-terminal Drought Stress on Yield Components of Four Barley Cultivars. J Agron Crop Sci *195*, 427-441.
Schlaeppi, K., Dombrowski, N., Oter, R.G., Ver Loren van Themaat, E., and Schulze-Lefert, P. (2014). Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. Proc Natl Acad Sci U S A *111*, 585-592.
Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Proceedings / International Conference on Intelligent Systems for Molecular Biology ; ISMB International Conference on Intelligent Systems for Molecular Biology *6*, 175-182.
Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic acids research *34*, W609-W612.
Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75, 7537-7541.
Trimble, W.L., Keegan, K.P., D'Souza, M., Wilke, A., Wilkening, J., Gilbert, J., and Meyer, F. (2012). Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC bioinformatics *13*, 183.
Tusche, C., Steinbruck, L., and McHardy, A.C. (2012). Detecting patches of protein sites of influenza A viruses under positive selection. Molecular biology and evolution *29*, 2063-2071.
Zhu, W., Lomsadze, A., and Borodovsky, M. (2010). Ab initio gene identification in metagenomic sequences. Nucleic acids research *38*, e132.