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

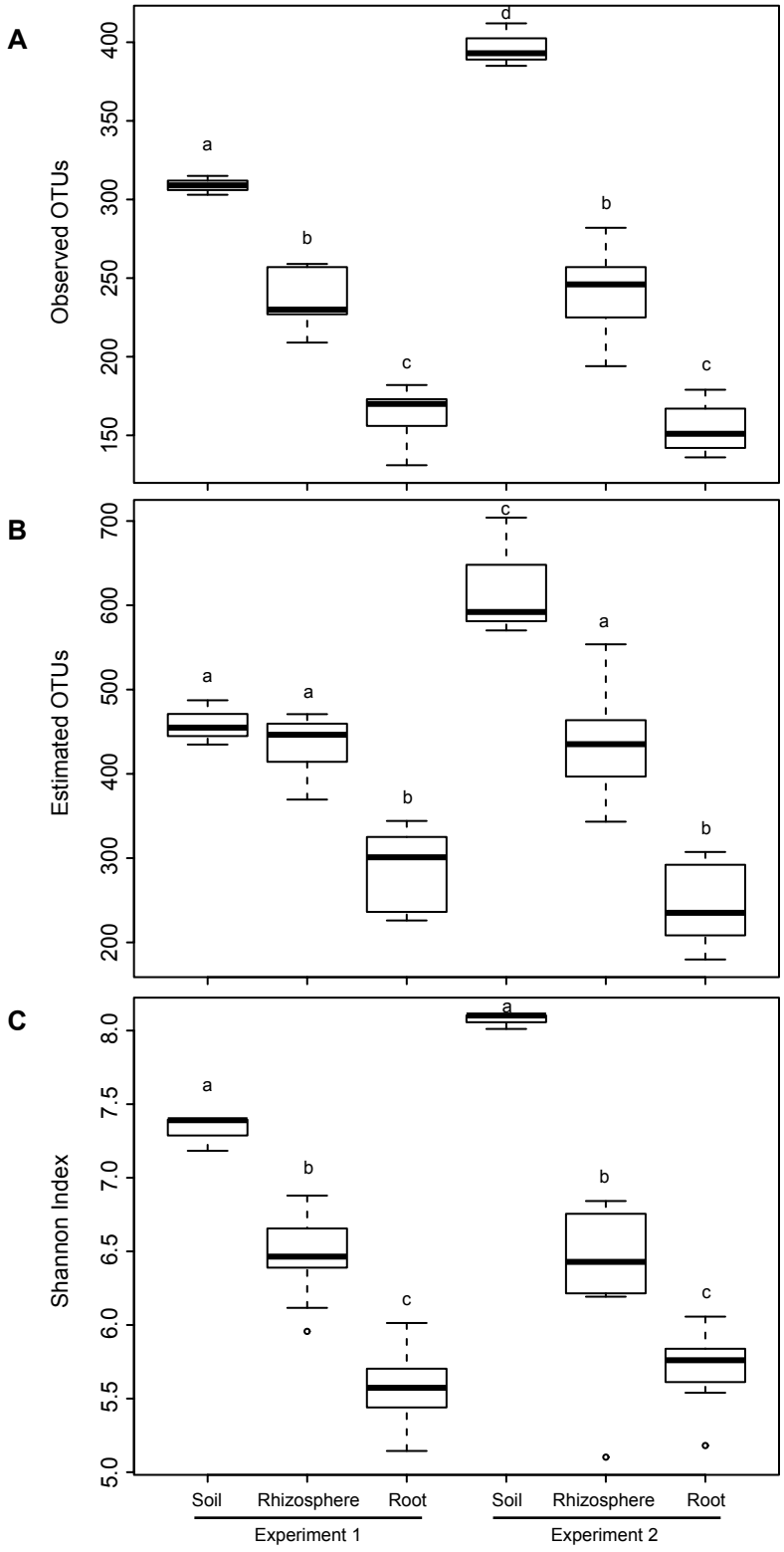
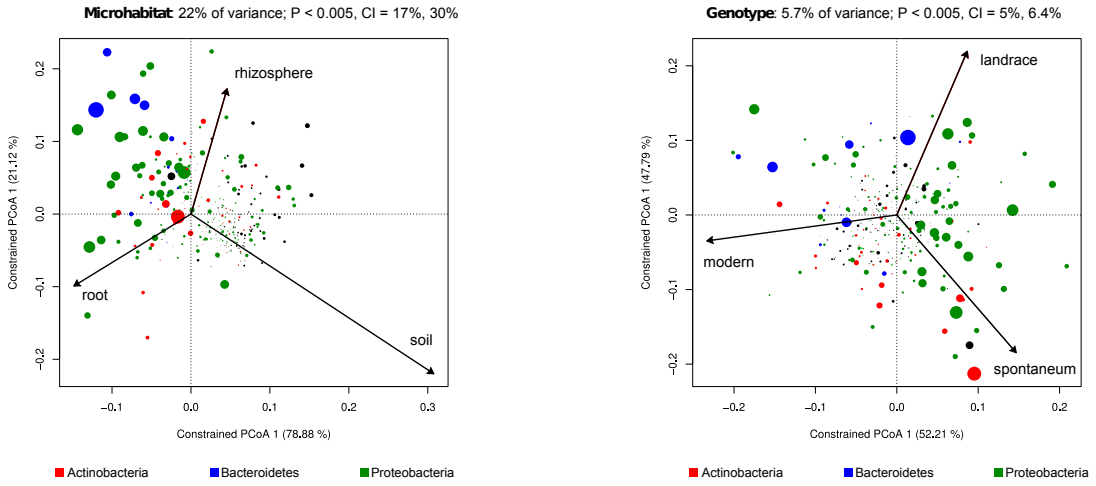
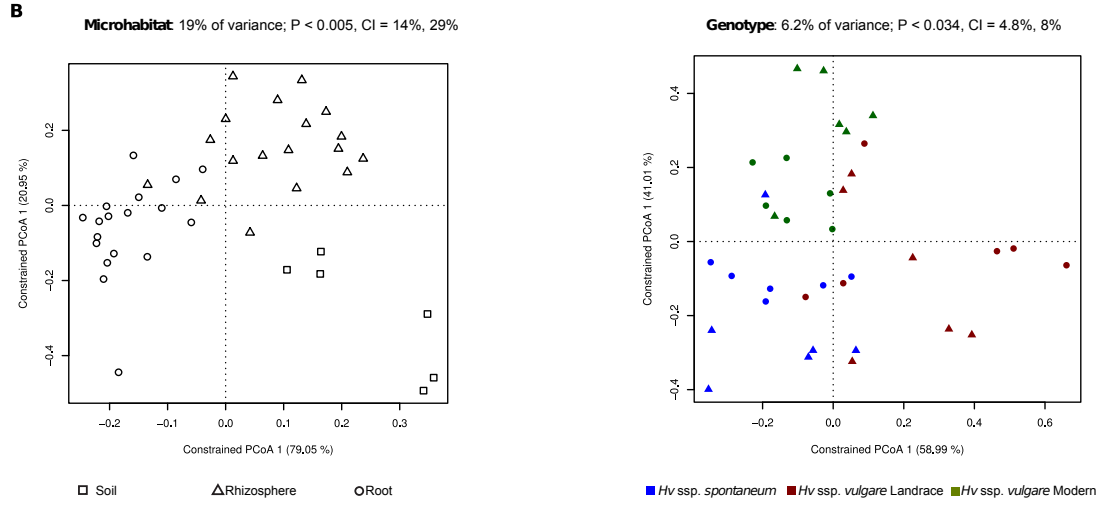


Figure S1. Alpha-diversity calculation of the soil, rhizosphere and root samples. (A) Total number of observed OTUs, (B) Chao1 estimator and (C) Shannon's diversity index. Samples were rarefied to 1,000 reads prior the analysis. Different letters denote statistically significant differences by Tukey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

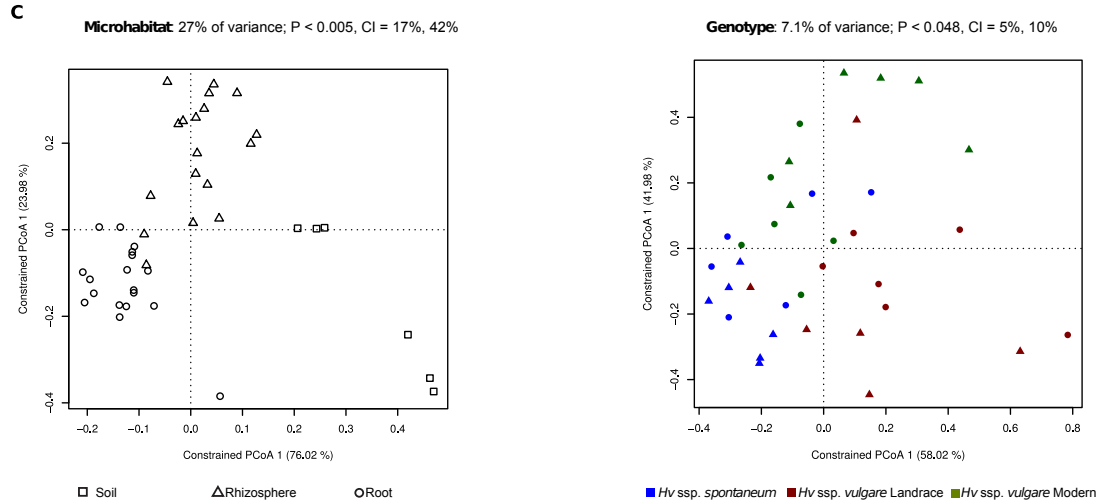
A


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

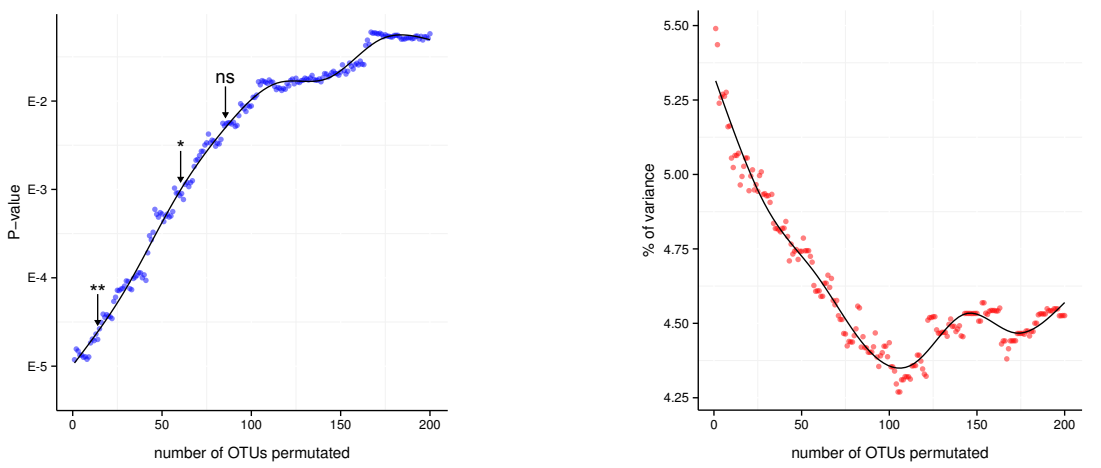

ns
*
**
P−value
E-2
E-3
E-4
E-5
number of OTUs permutated
% of variance
number of OTUs permutated

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.

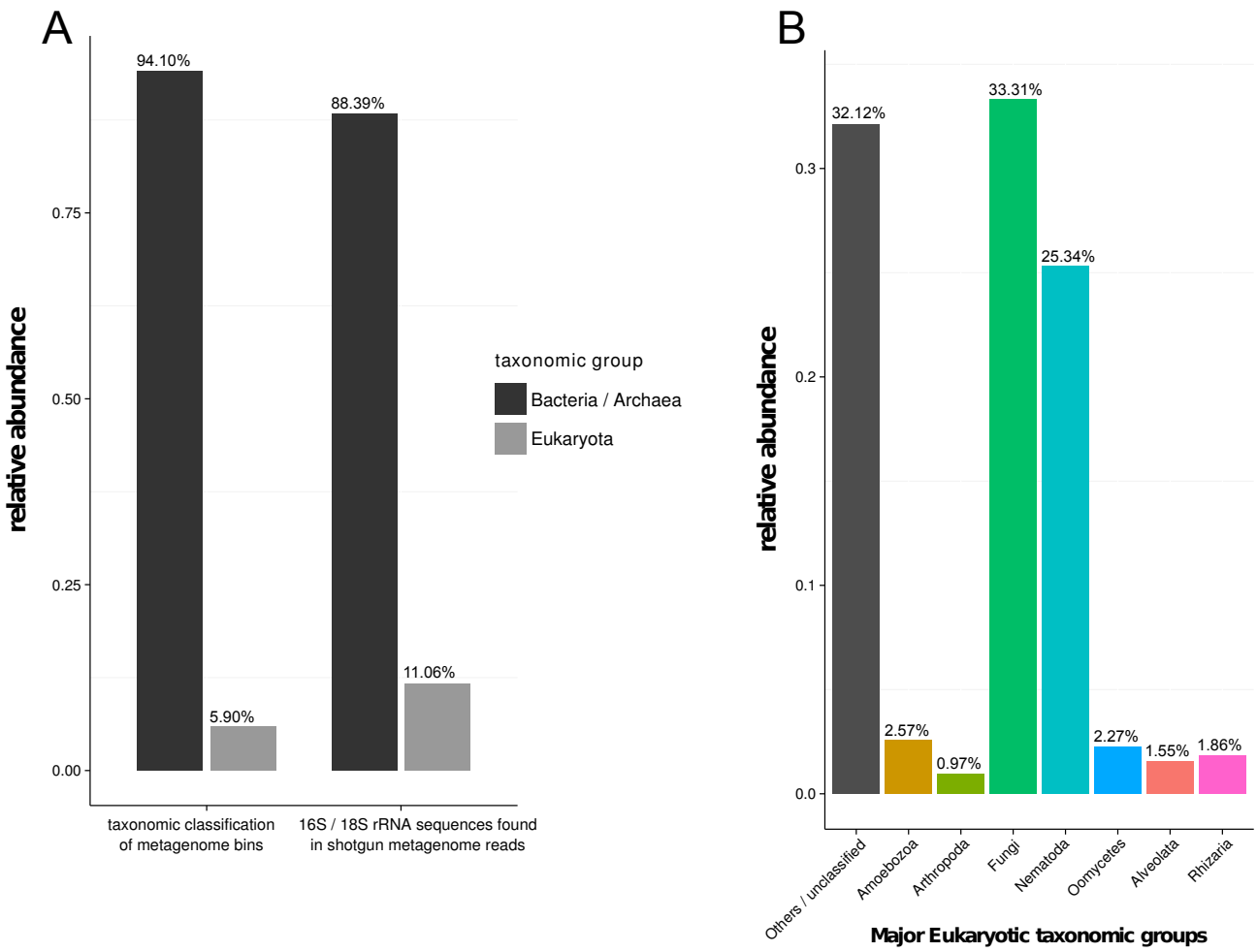


Figure S3. Analysis of Eukaryotic abundances. Comparison of Bacterial to Eukaryotic abundances estimated by taxonomic classification of metagenome bins and comparison of 16S and 18S rRNA gene reads found in the shotgun metagenome reads (A). Relative abundance comparison between major Eukaryotic taxonomic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